Regression analysis of crime rate in North Carolina

Stone Jiang

Introduction

The purpose of this report is to understand the key determinants of crime in order to generate effective, actionable policy recommendations for political candidates running for election in the state of North Carolina. To achieve this goal, we will examine a cross-sectional dataset of crime rate in various counties of North Carolina in the year 1987. We provide three carefully interpreted linear regression models to explore the predictive power of the independent variables in this dataset with respect to our dependent variable crime rate, focusing on how potential relationships may be utilized for better control and detection of crime. This work will focus on addressing the overarching research question: What are the major deterrents and motivators of crime, and how does the existence of factors from both categories influence overall crime rate?

Specifically, we will look to three subquestions to help narrow down our focus:

1. How does fear of arrest and convictions deter crime across North Carolina?
2. How does wage for all types of employees influence overall crime rate?
3. What variables are strong predictors of crime, and at the same time are robust across the entire state, irrespective of location and population density?

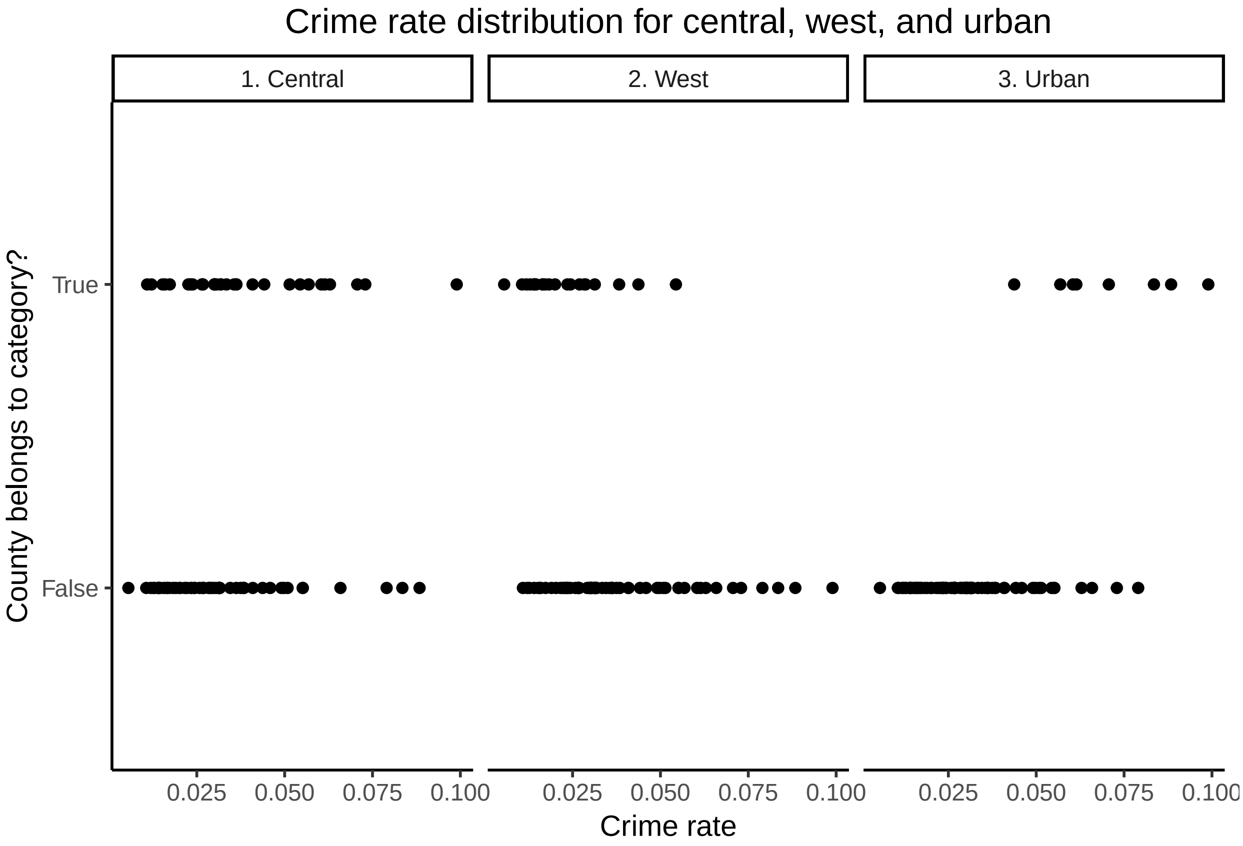
We will provide rational for variable and model selection using both background information and automated methods. The development, evaluation, and interpretation of our models were divided in three stages, which can be found in sections *Model 1*, *Model 2* and *Model 3*. Finally, the policy recommendations derived from this process are presented in the *Policy Recommendations and Concluding Remarks* section.

Initial Data Cleaning

For this project, the dependent variable of interest is crime rate. Before choosing the best independent variables for our models, the data was examined and cleaned as follows. First, we omitted the last rows of the csv as they were empty and did not contain data. Second, we eliminated the one duplicated entries, as identified by unique county ID (193). Third, we verified that the datatype of all of our variables are consistent with expectation, and that no missing values were detected. Here, we converted the prbconv variable from a factor to a numeric variable. R initially read this variable as a factor due to omitted rows at the end of our having non-numerical values. This leaves us with data points for 90 counties.

After this initial data cleaning, we identified columns we believed could be potential casual predictors of crime rate. N.B. county and year were disregarded in our models because they are identifiers.

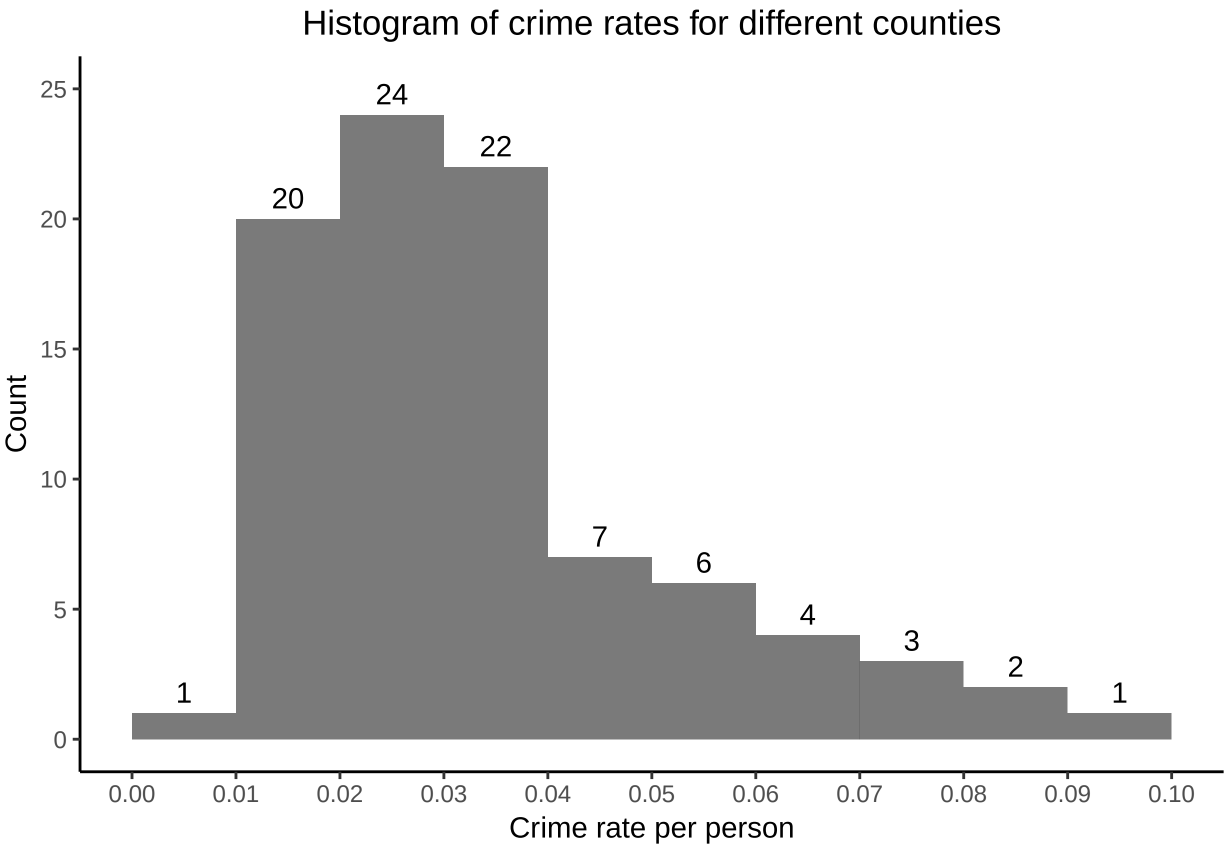
The variables west, central and urban are categorical, and came encoded as 0 and 1. This can be used directly in the regression as identifiers. For example, a regression model containing the variable "west" would have a non-zero coefficient for counties in the west, and a 0 coefficient for counties on the east. We first looked to see whether the distribution of crime rate is different depending on the location (west vs central) and whether the county was urban.



For Central versus not Central North Carolina, the crime rate distribution is relatively even. Counties in Western North Carolina appear to have less crime on average than those labeled as not Western. Counties labeled as Urban have more crime on average.

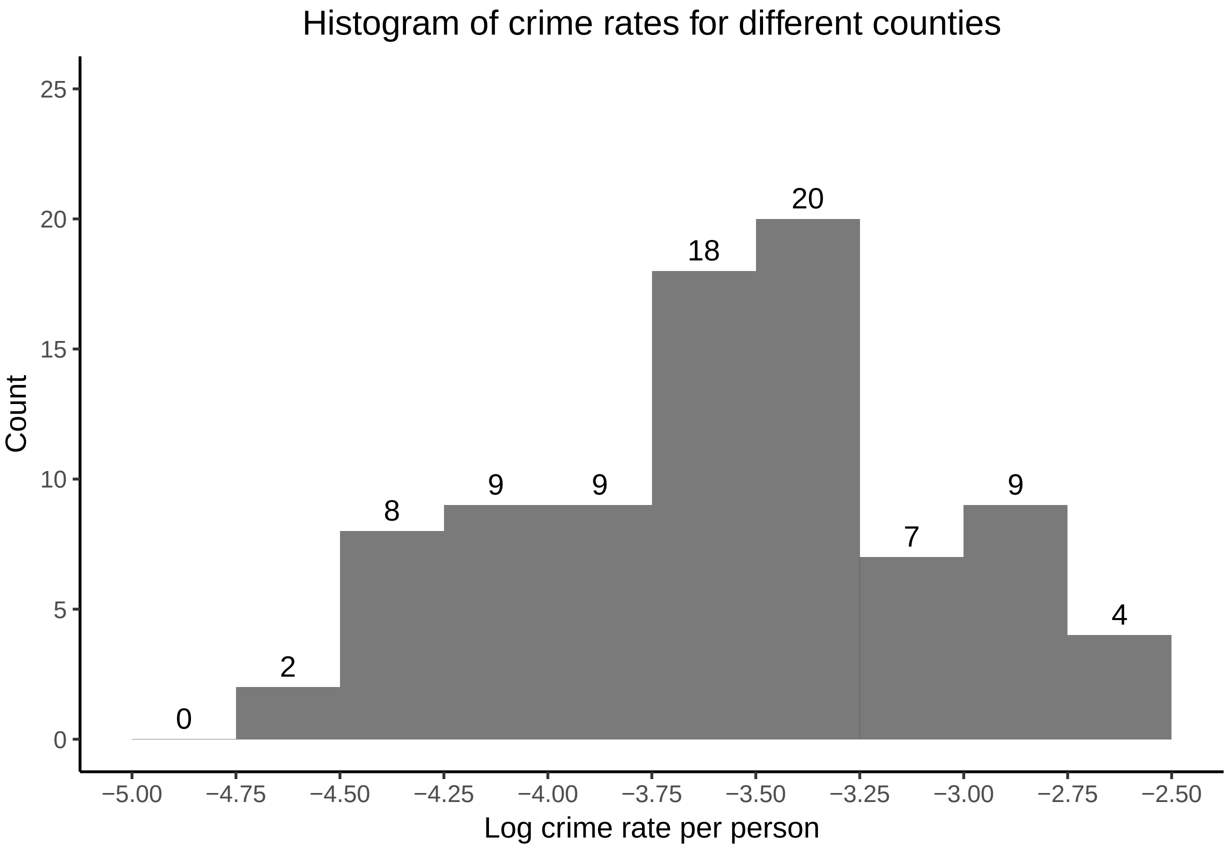
We next examined the dependent variable crime rate as defined by crime rate per capita.

```
##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966
```



First, we see that the mean crime rate across counties is higher than the median, at about 3.4 crimes per 100 people. The crime rate is greater than 0 for all counties and highly skewed toward larger values. For this project, we aimed to interpret our model coefficients as how changes in explanatory variables affect changes in crime rate. Since the baseline crime is different for different counties, it is beneficial to transform crime rate into the log of crime rate. This allows us to interpret changes in crime rate as a percentage, which makes comparisons across counties more meaningful For example, a 0.01 change in crime rate for the lowest county (0.005) is a much larger percent change than for the largest county (0.099), but a 1% change is comparable regardless of the baseline crime rate.

The following figure shows a histogram of the logarithms of crime rate per county in North Carolina.



The distribution of the logarithms of crime rate follow a closer to normal distribution with no outliers. This is also a desirable for the creation of predictive models.

Model 1

Key Variables

For our initial model, we would like to focus on factors that intuition says should influence crime. We believe that there are four variables which represent deterrents to crime: probability variables of arrest, conviction, prison sentence, and the severity of punishment in average sentence days. Collectively, we will call these variables the “fear factors.”

For example, we believe the higher the chance someone believes they will be arrested, convicted, or sent to prison, the less likely they will commit a crime. Also, the more severely someone believes they will be punished for the crime (as measured by prison day sentences), the less likely they will commit a crime. Out of these four, we believe probability of arrest and probability of conviction will have the greatest effects. The reason is that a single arrest or conviction can permanently damage someone’s record. For most people who have never committed crimes before, just the idea of possibly getting in trouble with the police could be enough to deter them. In addition, there are many crimes that result in fines, community service, and other forms of punishment that does not involve prison.

The wage variables can either deter or motivate individuals to commit a crime. We believe that the more satisfied someone is with their income, the less likely they will commit a crime because they are more likely to attain their desires without the need of illegal routes. Along the same lines, unemployment is likely to lead to increased crime rates. Too high of a wage, especially in blue collar jobs that are non-customer facing and physical in nature, means some employees can be “priced out”. As wage goes up, individuals paid that wage are expected to do more, lowering the amount of workforce necessary, leading to greater unemployment.

For our base model, we will look at only what we consider traditional blue collar jobs that are non-customer facing and physical in nature: construction and manufacturing. **We also take the log of all wage variables**: this is standard practice as we want to measure the effect of a percent changes in salary, and not absolute changes, a similar argument to taking the log of crime rate.

Before performing EDA on the variables listed above, we note why we have chosen to exclude the other variables in our base model. Omitting these variables can potentially introduce bias into the model, but for the first model, we wish to only use the key determinants of crime.

Additional variables

We believe that density should be a positive predictor of crime. Previous studies have shown that, even though the relationship between population density and crime rate is complex, a highly dense population areas present higher crime rates up to about a density limit of 500 people per squared mile, which North Carolina falls under [1]. We will discuss density further in models 2 and 3.

Tax could reflect how people vote [2]. Tax is also linked specifically to income-producing crimes [3]. According to the literature, for these specific kind of crimes, taxation has an important deterrent effect as it increases the risk criminals assume from these activities. Since we are not differentiating between kinds of crimes, this is left aside for now.

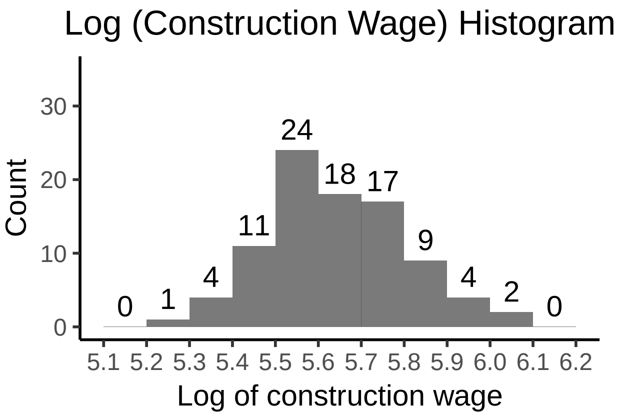
Percent young male may also be a potential strong predictor of crime in counties where young male have particularly bad influence. For example, criminologists believe that women are always less likely to commit crimes than men [4].

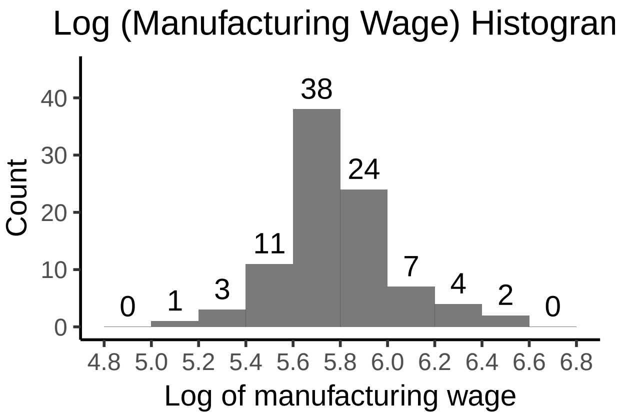
We will not consider the variable “mix,”" because we are interested in crime rate regardless of the nature of the offense.

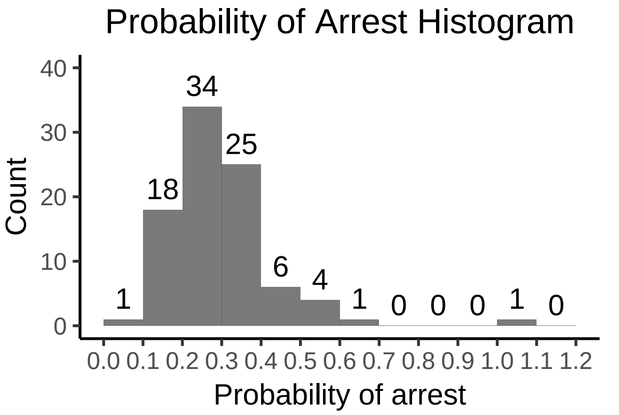
For all other variables, we will consider them in models 2 and 3, and in our final recommendations.

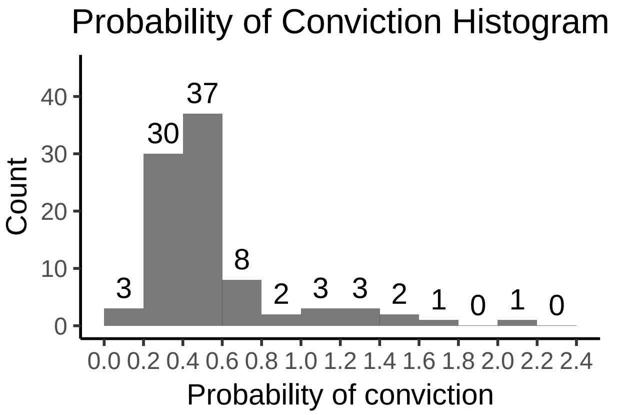
EDA and a note on "probability" variables

For our 4 variables, we first performed EDA to ensure that we had reasonable data. We plotted a grid of histograms and look at the distribution of the explanatory variables.









We see that the log of the two wage variables have no outliers. Both are somewhat upward skewed in that there are more data points above the mode of each, but overall, the distribution looks fairly symmetric.

For the probability of arrest, defined as the number of arrests to offences, the general trend is a skew to the right, with one datum above 1. An explanation for this is that we are looking at a cross-sectional data pooled from a multi-year study. For example, if data collection started in June of one year, and people committed an offence in January of that year but not arrested until after June, that person could appear in this data set as having been arrested but not committing an offence. As a result, this variable is not precisely probability, but rather a metric variable indicating level of arrest per offense.

For the probability of conviction, defined as ratio of conviction to arrests, there are many more data points skewed to the right. This variable is confounded by the fact that one does not necessarily need to be arrested to be convicted of a crime. Another mechanism is a citation, which are issued in place of arrests for smaller crimes. Therefore, it is reasonable for this variable to exceed 1 as well.

Model and Coefficient Interpretation

We now build our regression model. **Note that in the EDA, we had previously log transformed all wage variables, and stored them within the X_wage_transformed data frame.**

```
##
## Call:
## lm(formula = data$logcrmrte ~ prbarr + prbconv + wcon + wmfg,
##     data = X_wage_transformed)
##
## Coefficients:
## (Intercept)       prbarr      prbconv         wcon         wmfg
##     -8.4338      -1.6815      -0.7070       0.4696       0.5408
```

The coefficient for prbarr represents the effect of the variable prbarr on crime rate. Specifically, keeping all other variables constant, we can convert this coefficient to an exact change. For every 0.1 unit increase in prbarr:

$$\begin{aligned}\%\Delta\text{crmrte} &= 100 * [\exp(\beta * \Delta\text{prbarr}) - 1] \\ &= 100 * (e^{-1.6815*0.1} - 1) \\ &\approx -15.5\%\end{aligned}$$

we expect a -15.5% change in crime rate (or 15.5% decrease in crime rate). Since this variable represents the "fear factor" we presented above, this provides support for our hypothesis that the effection is negative. An analogous interpretation can be said for probability of conviction. Keeping all other variables constant, for a 0.1 increase in prbconv:

$$\begin{aligned}\%\Delta\text{crmrte} &= 100 * [\exp(\beta * \Delta\text{prbarr}) - 1] \\ &= 100 * (e^{-0.7070*0.1} - 1) \\ &\approx -6.83\%\end{aligned}$$

we see a 6.83% decrease in crime. With these variables, we measure how a perceived probability of getting in trouble with the legal system deters crime. These seem to be practically signficant as they consistitute large percent decreases in crime.

The coefficients on the wage variables represents how a percent change in average wage in that industry relates to a percent change in crime rate. This can be used directly in the interpretation since these coefficients are small. Namely, keeping all other coefficients constant, a 1% increase in construction wage leads to a 0.47% increase in crime rate, while a 1% increase in manufacturing leads to a 0.54% increase in crime rate. This is consistent with our hypothesis as well. Even though the changes are small (and likely practical insignificant), we may be observing the fact that more people are losing their jobs as wages increase (and labor per individual also increases).

Classical Linear Model Assumptions

At this point, we will evaluate the Classical Linear Model Assumptions, and perform hypothesis testing to see whether each of our coefficients are statistically significant.

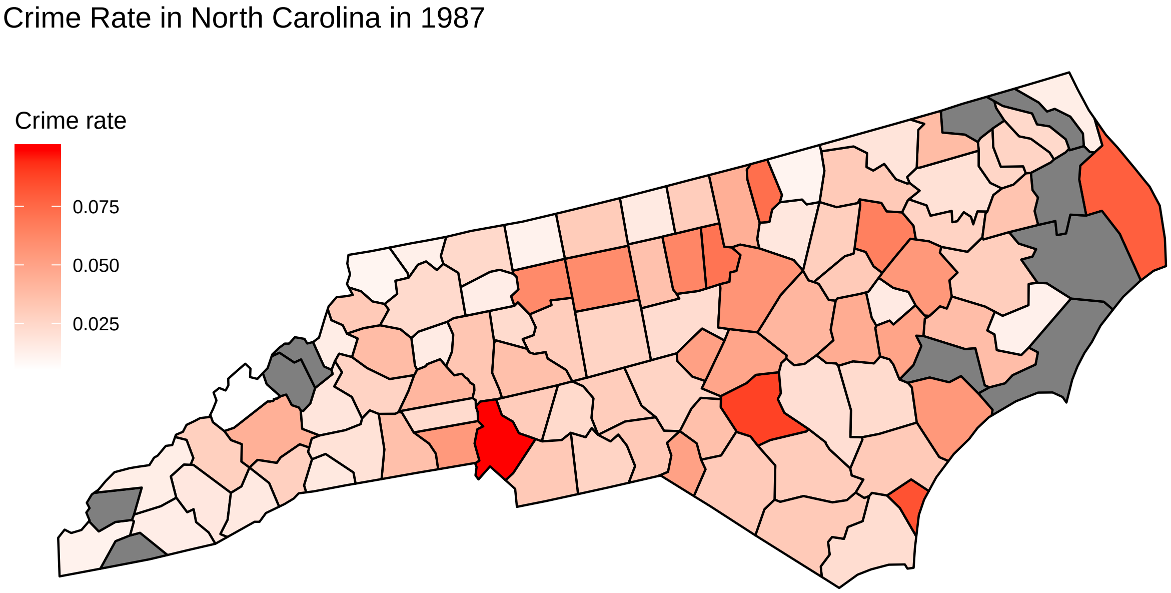
CLM 1: Linear in parameters

Nothing to assess here. We define the model with an error term such that the parameters are linear (and assume this model is the population model and estimate its parameters). The independent variables can be transformed in any way, including taking logs as we have done.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

CLM 2: Random Sampling

This is a rare case where we actually have a majority of the population at hand. We are interested in crime rate in the state of North Carolina, which has 100 counties. We have data for 90 of these counties. We can generate a visualisation to see where which counties were eliminated to see if there was systematic geographic bias. This is done with the plot_usmap package.



We see that the 10 counties without data (in black) are somewhat clustered along the eastern and western/north western boarders of North Carolina. But since we have data points for even clustered geographic regions where data is missing, we should be able to draw fairly reasonable conclusions about crime in the state as a whole.

Within each county, which we can view as our available population, we have no reason to believe that the sampling was not random, or even in some cases a consensus. For example, it is not hard to imagine that the crime rate per capita could be calculated from police records as a consensus. Our police per capita, data from the FBI, is also likely a consensus. Wage variables are likely a sample of employees, at least from available data reported to the IRS. We have no reason to believe that this sample was biased in any way. Overall, given the limited information, we have little reason to drastically doubt an IID sample within our available population of 90 counties.

CLM 3: No perfect multi-collinearity

First, multi-collinearity is guaranteed when we have more features than samples, which is not the case here. Second, multi-collinearity can occur when one variable is a perfect linear combination of another set of variables. In that case, the one of those variables are regressed on the remaining of the group, the R^2 will be 1. R would have warned us if this were the case (by evaluating whether the covariance matrix is singular), so we have fulfilled this requirement. We can further evaluate this using the VIF for each coefficient to evaluate whether some degree of multicollinearity should be of worry. This is done as follows.

```
##   prbarr  prbconv     wcon     wmfg
## 1.090235 1.026432 1.244954 1.164963
```

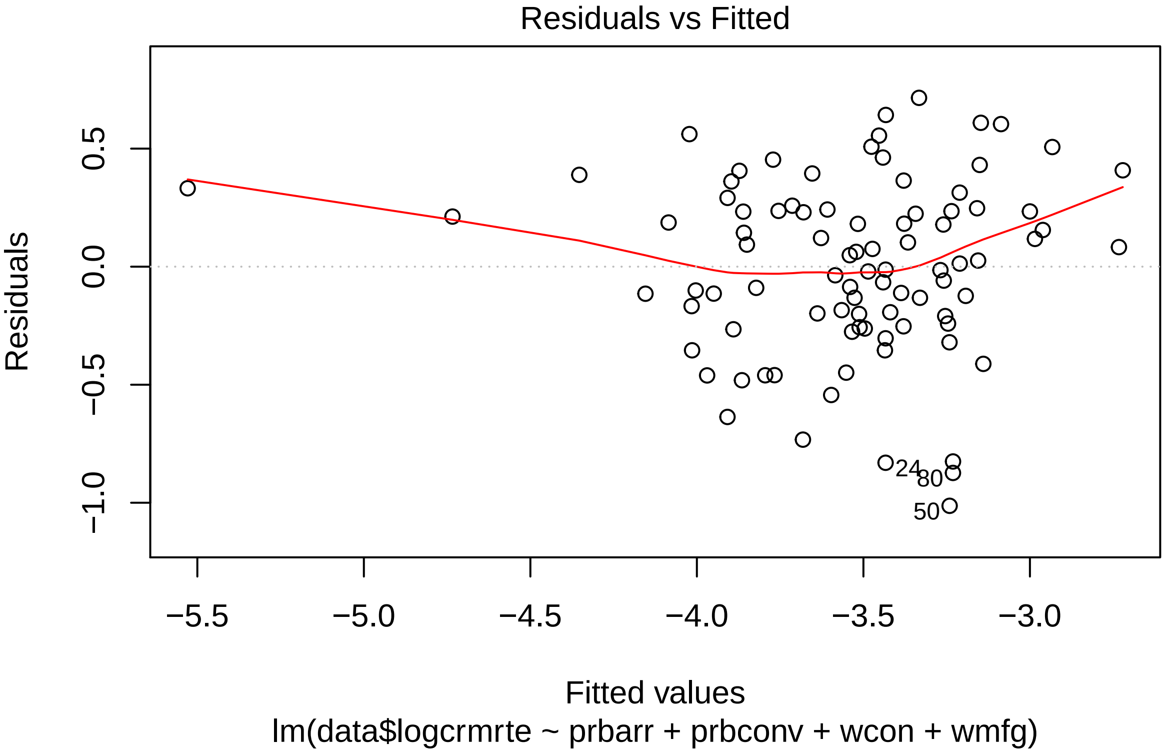
We see that all VIF factors are significant below 4, which means we do not have significant multi-collinearity to worry about.

CLM 4: Zero Conditional Mean

Zero conditional mean states that the expected value of the error term is 0 for all values of the independent variables x_k.

$$E(u|x_1, x_2, ..., x_k) = 0$$

Under zero conditional mean, we expect that the residuals on the residuals versus fitted value plot to have an expected value of 0 across the board. To check this, we plot the residual agains the fitted values for our set.



Based on this plot, we see that unfortunately, the line adopts a U shape. However, the curvature is a result of very few data points on the extreme ends of the fitted values. In the middle where the bulk of our data is, from -4 to just before -3, the line seems flat and centered around 0. However, above 3, the 6 data points are all above 0. The conclusion is that our model most likely does not satisfy CLM 4. We will need to adjust our model by adding more parameters in order to capture more of the variation in crime rate due to omitted variables.

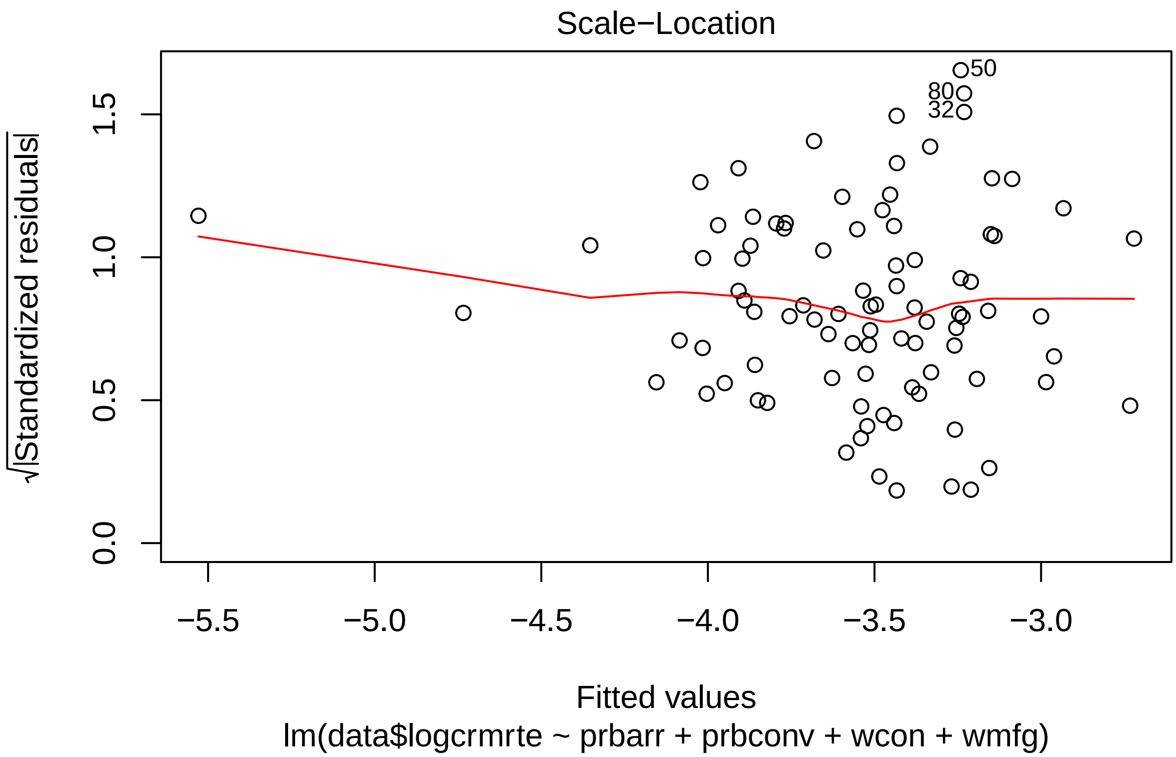
CLM 5: Homoskedasticity

Homoskedasticity assumption is that the variance of the error terms are constant for any combination of x_k values.

$$Var(u|x_1, x_2, ..., x_k) = \sigma^2$$

Examining the fitted values versus residuals plot above, while the spread (larger the spread the greater the estimated variance) appears to be slightly larger around fitted values of around 3.75 (around -1 to 0.5) than around 4 (around -0.5 to 0.5), overall there are no major observable patterns in differences in variance as a function of x.

We can also check the scale-location plot. If homoskedasticity were achieved, we would expect a horizontal line across this plot:



We see that this line is roughly horizontal from -5 to -3. The only major curvature is the single data point around -5.5. However, this is likely due to small sample size for that particular fitted values. Discrepancies such as that observed are much more likely when the sample size is small. This indicates that we most likely have close to homoskedasticity.

One way to test for homoskedasticity is the Breusch-Pagan Test. The null hypothesis of the test states that we have homoskedasticity. We will test at a standard significance level of 0.05.

$$H_0 : \text{Homoskedasticity}$$
$$H_a : \text{Heteroskedasticity}$$

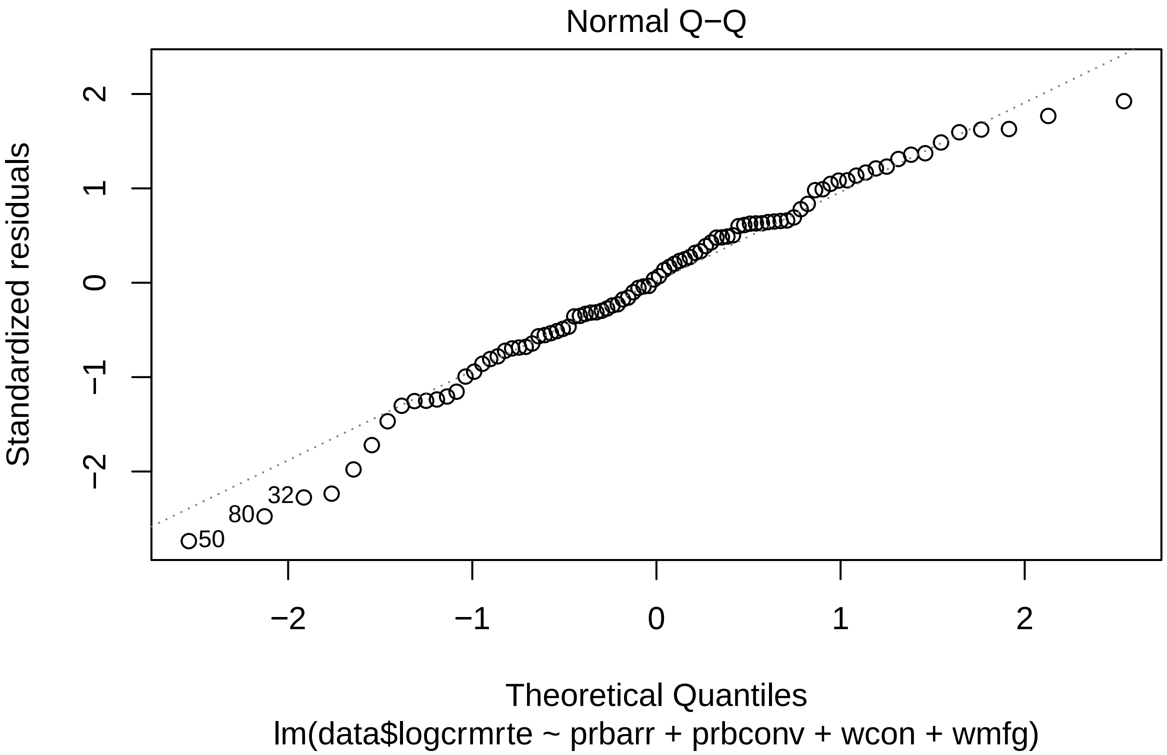
```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 4.0136, df = 4, p-value = 0.4042
```

Since the $p - value >> 0.05$, we fail to reject the null hypothesis that we have homoskedasticity.

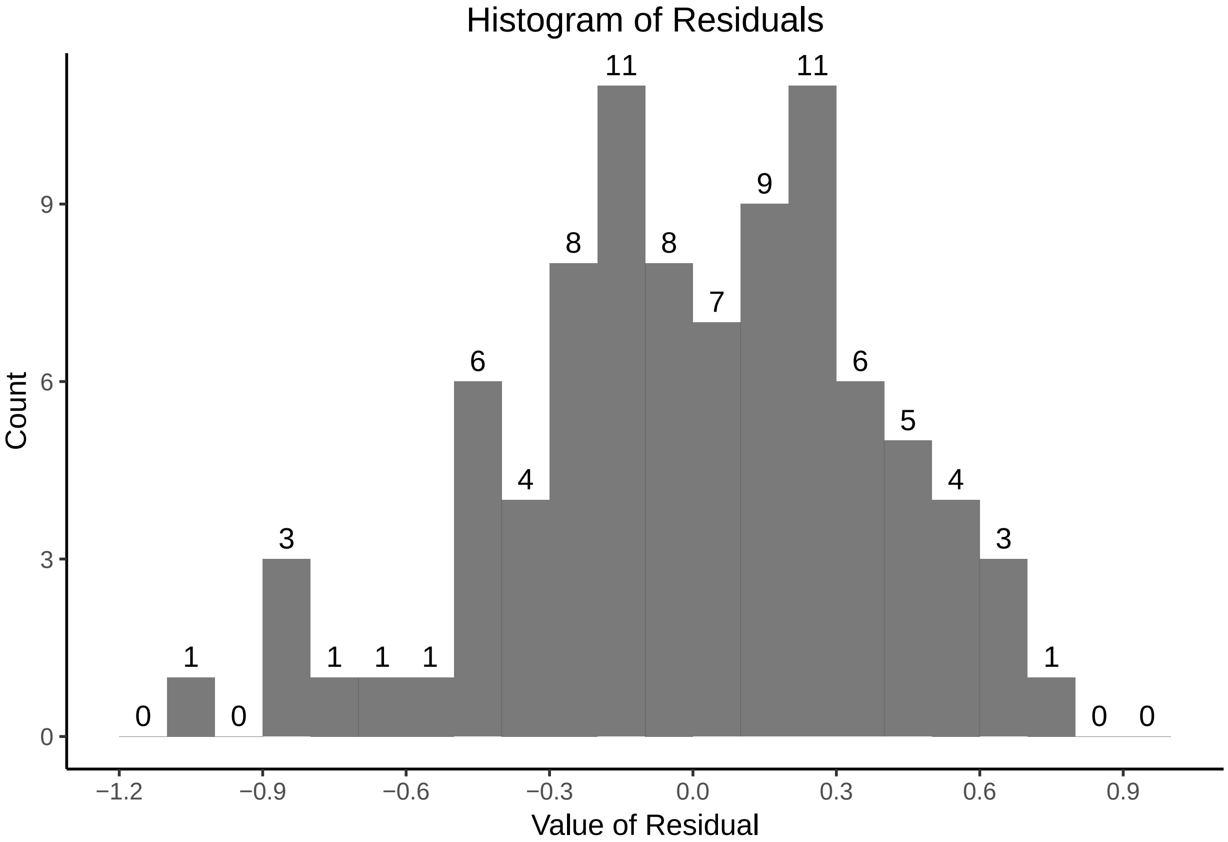
In any case, it is good practice to almost always use heteroskedastic robust errors, especially since we have some doubt from the residuals versus fitted values plot.

CLM 6: Normality

CLM 6 assumes that population error is independent of the explanatory variables x_1 through x_k, and that the error term is normally distributed with mean 0 and constant variance. We can check this with the qqplot of the fitted values versus residuals plot.



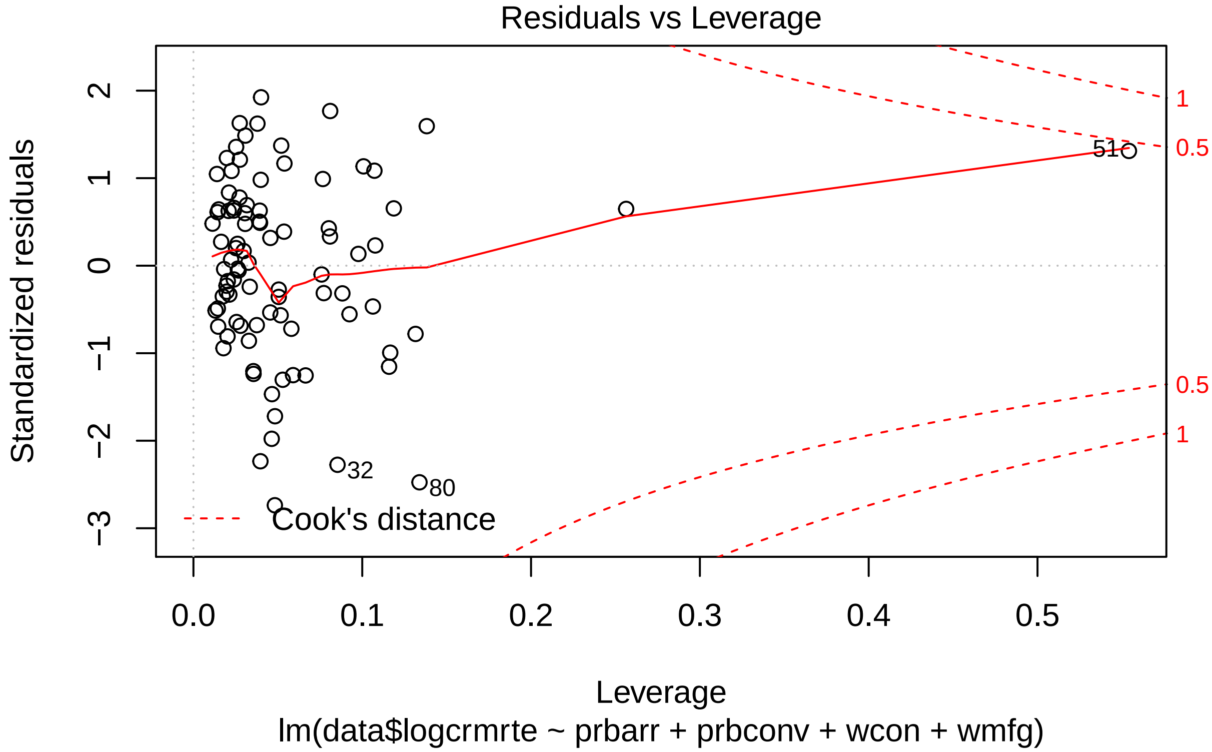
Not even counting the exception of extreme values, the data points wavering back and forth, which could indicate a kurtosis problem. We can further visualise the residuals in a histogram.



Based on the histogram, the data does not appear very normal. In fact, it is somewhat bimodal around -0.4, and 0.2.

In any case, since our sample size 90 is much greater than 30, asymptotics also kicks in, ensuring that the sampling distribution of our coefficients are approximately normal. This will be important in statistical testing.

Finally, we would like to check and see if there are any outliers in our model that might have significant influence:



We see that data point 51 could be problematic. Its Cook's distance is still below 1, which means it does not have high enough influence for us to worry about at the moment.

We will now perform statistical testing to see whether the four coefficients we included are statistically significant. To do this, we derive heteroskedastic errors from the vcovHC function from the sandwich package. This function produces a covariance matrix, and the standard errors are the square root of the diagnal.

```
## (Intercept)      prbarr     prbconv        wcon        wmfg
##   1.9047222   0.4570443   0.1447330   0.3162314   0.2486414
```

We see that prbarr and wcon have the largest standard errors. In order to look at the statistical significance of our statistics, we can perform a t-test using the robust standard errors. Since we have large sample size, the sampling distribution of our statistic is approximately normal, so our statistic is distributed as a t-distribution:

$$\frac{\hat{\beta_j} - \beta_j}{se(\hat{\beta_j})} \sim t_{n-k-1}$$

For all the 4 betas, we will use a 2-sided test as significance level 0.05

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

To perform the test for all of the variables, we use the coeftest package, specifying the degrees of freedom as sample size - 4 (number parameters except beta_0) - 1, and the heteroskedasticity-consistent estimation of the covariance matrix.

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -8.43382    1.90472 -4.4278 2.813e-05 ***
## prbarr      -1.68150    0.45704 -3.6791 0.0004097 ***
## prbconv     -0.70698    0.14473 -4.8847 4.815e-06 ***
## wcon         0.46959    0.31623  1.4849 0.1412566
## wmfg         0.54077    0.24864  2.1749 0.0324169 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the statistical test, we see that both the probability of arrest and probability of conviction are both highly significant variables with p-values << 0.001, while the wage of construction is not statistically significant. The wage of manufacturing is somewhat statistically significant, with a p-value close to 0.05.

Based on this simple, we have evidence to support the following answers to our research questions:

1. Fear of arrest and conviction are both likely highly significant factors in deterring crime.
2. Blue collar wages in industries that are non-customer facing and primarily physical labor in nature are positively correlated with crime. We have yet to determine how other wages related to crime rate.

To make a definitive policy recommendation, we will require additional exploration of other co-variates in our other models.

Model 2

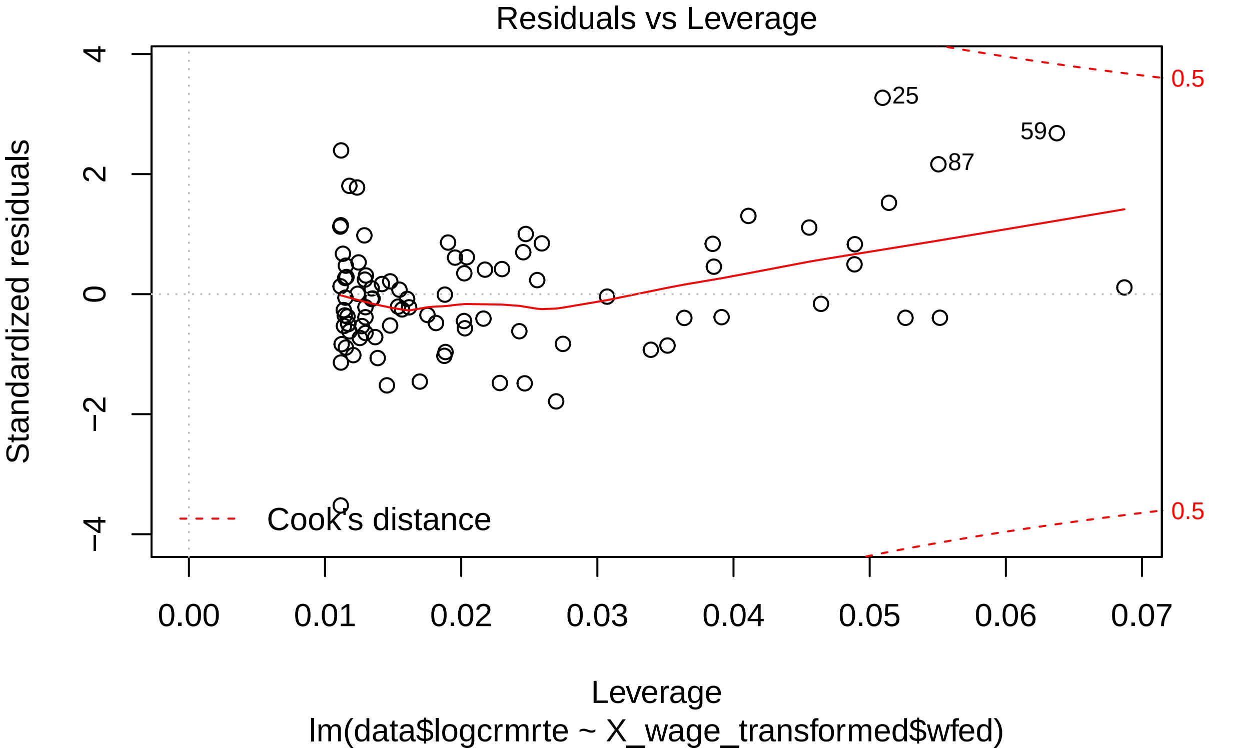
Exploration of independent variables

For model 2, we wanted to add in other covariates meant to increase accuracy of prediction. To do this, we wanted to first get a sense of crmrte correlation with all numeric variables. We also parse our data into X (numeric variables) and y (crmrte)

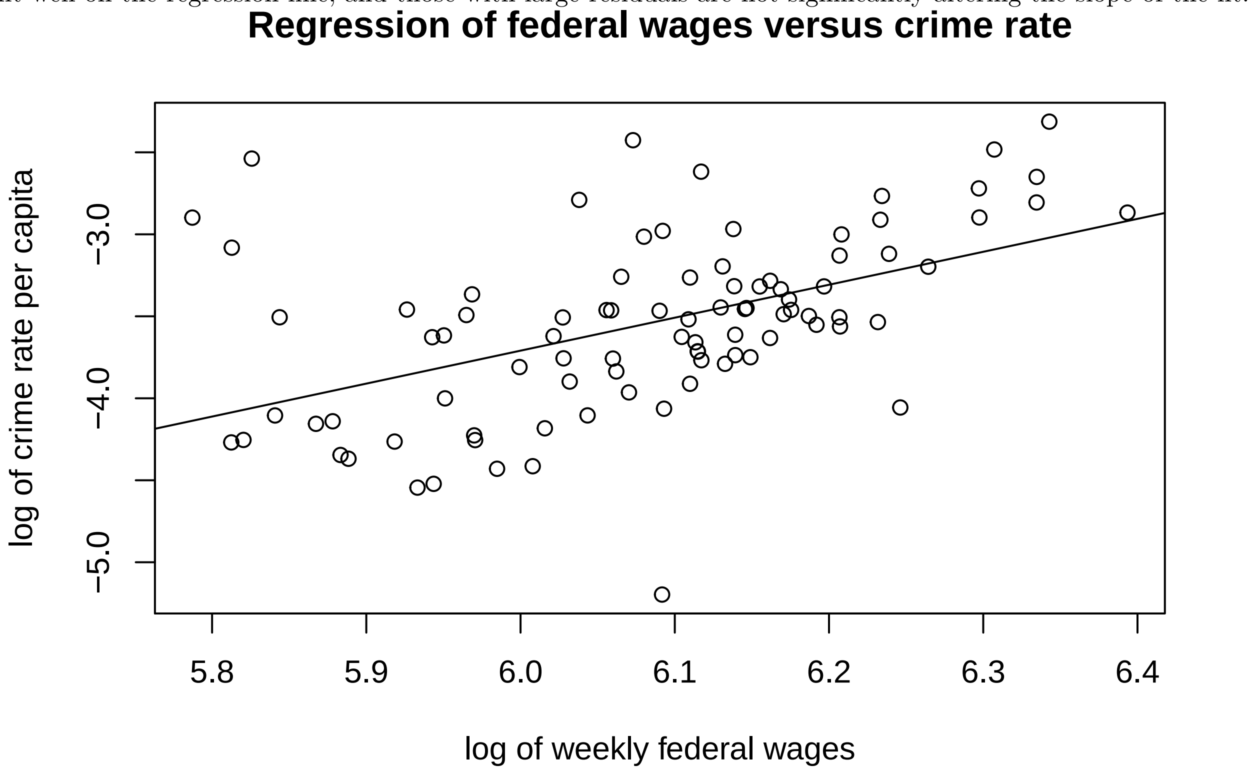
```
##    variable  crmrte_cor
## 1   density  0.63302339
## 2      wfed  0.52330585
## 3      wtrd  0.39379240
## 4      wcon  0.39371486
## 5     taxpc  0.35832339
## 6      wmfg  0.30753731
## 7      wfir  0.29324265
## 8      wloc  0.28856678
## 9   pctymle  0.27815466
## 10 pctmin80  0.23291821
## 11     wtuc  0.20146493
## 12     wsta  0.16970208
## 13     fips  0.02376789
## 14  prbpris  0.02147024
## 15    polpc  0.01040580
## 16   avgsen -0.04936931
## 17     wser -0.11312801
## 18      mix -0.12473445
## 19  prbconv -0.44681361
## 20   prbarr -0.47276691
```

It should be no surprise that density is the best single positive predictor of crime rate. Highly dense population areas present more opportunities for crime, and also have larger wealth gaps. We would like to exclude this variable from our regression model 2, and instead, we will save this variable for model 3 in order to check the robustness of our model 2. This will be used address our third research question: can we develop a model robust across the entire state, irrespective of county and density within each county? Furthermore, urban is a similar categorical variable that is directly related to density, and will serve the same purpose in model 3.

Surprisingly, wfed is the second best single predictor of crime rate, which seems surprising. It is difficult to rationalize this relationship from theory. Instead, we first perform a bivariate regression to see if any influential outliers are present.



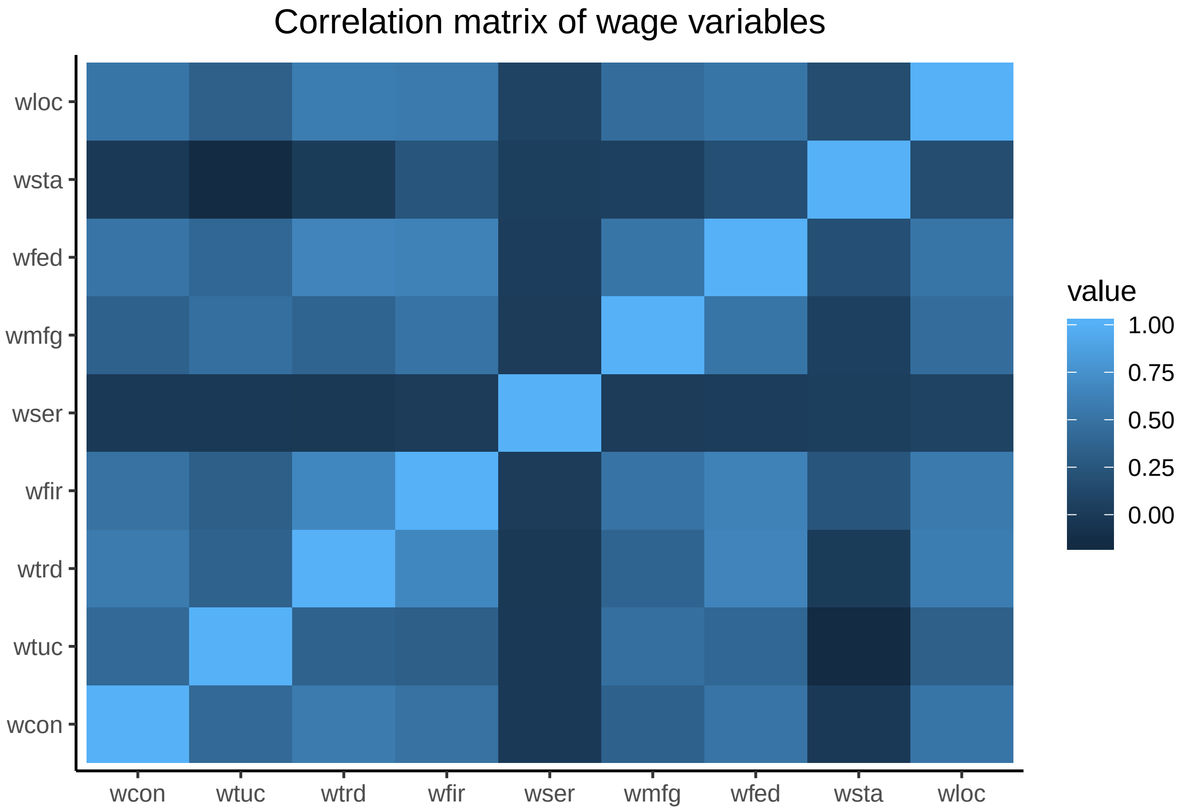
We can see that all data points have low Cook's distance, significantly less than 1. As a result, we do not have any outliers. We can also visually inspect the fit and come to the same conclusion that most data points fit well on the regression line, and those with large residuals are not significantly altering the slope of the fit.



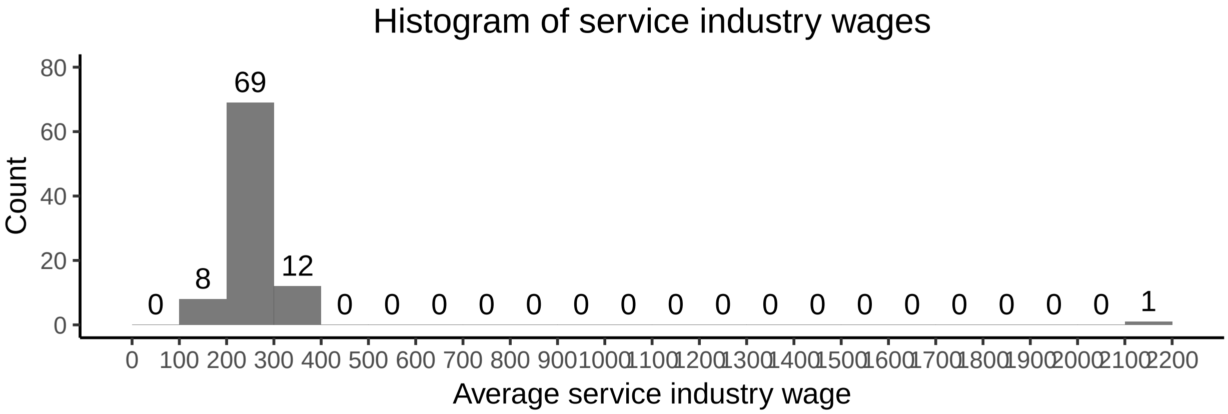
We will now take a closer look at the rest of the wage variables by generating an all-to-all correlation plot of each of the wages.

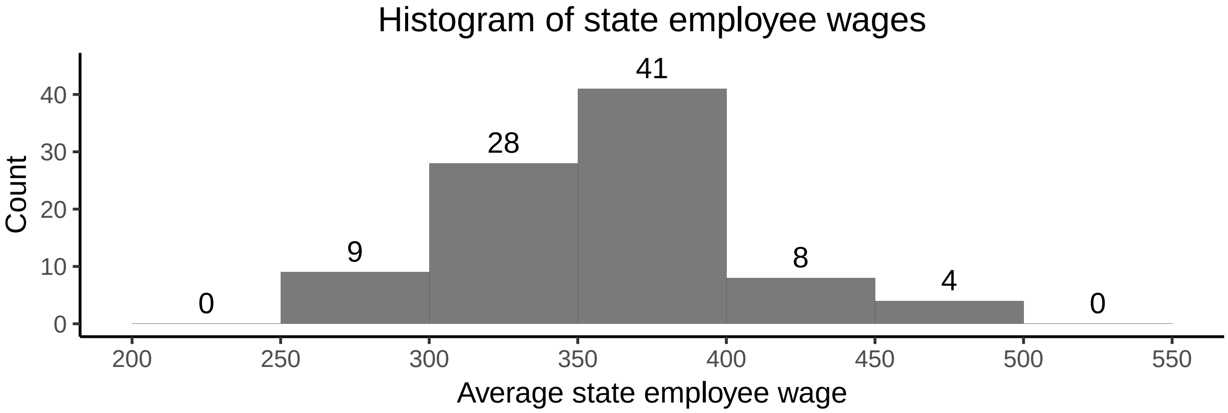
Wages Variables

At this point, it is important to mention that there has been previous research showing that lack of satisfaction with wages is an important predictor of crime rate [5]. Therefore, the analysis presented in this section is a relevant part of the EDA and variable selection for this project.

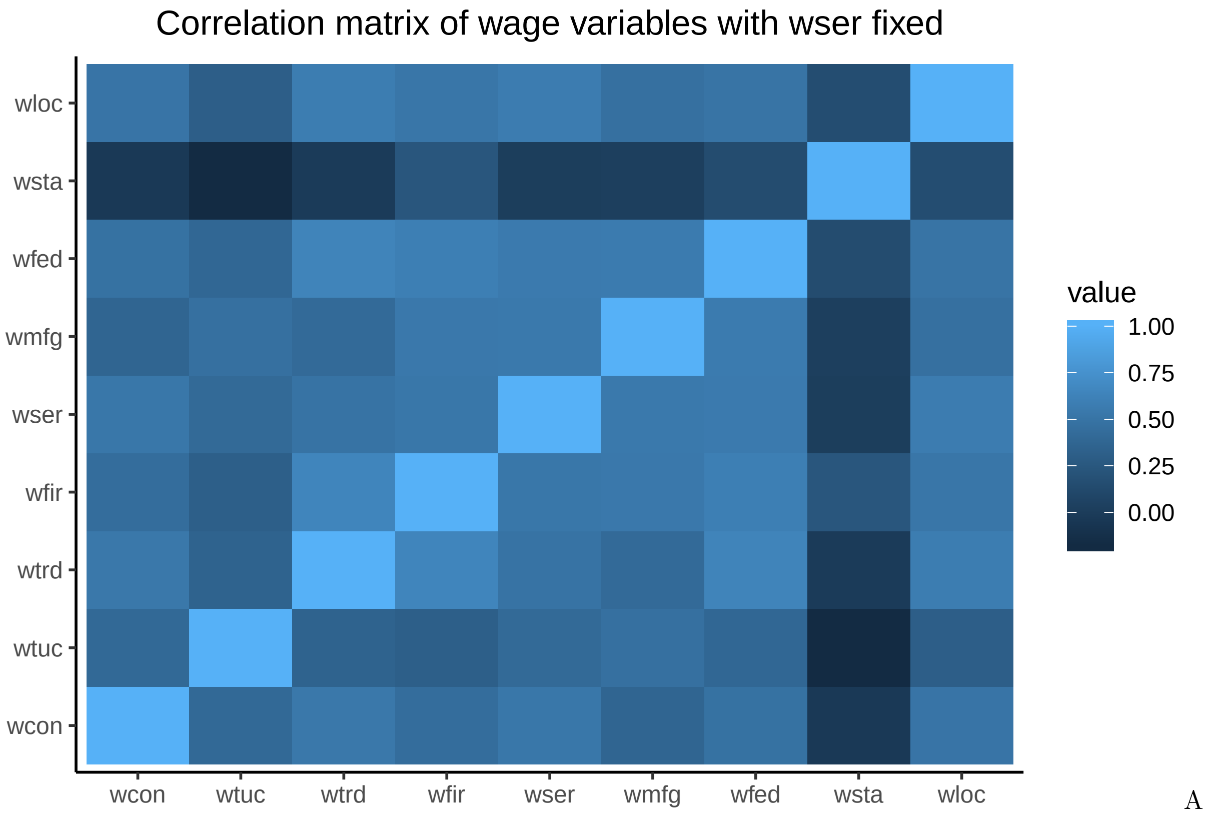


Interestingly, wser and wsta are both relatively dark compared to the rest of the data set. If the data was accurate, then that's a good indication of strong independent predictors within the wage category, so we perform EDA to ensure that these variables are reasonable:





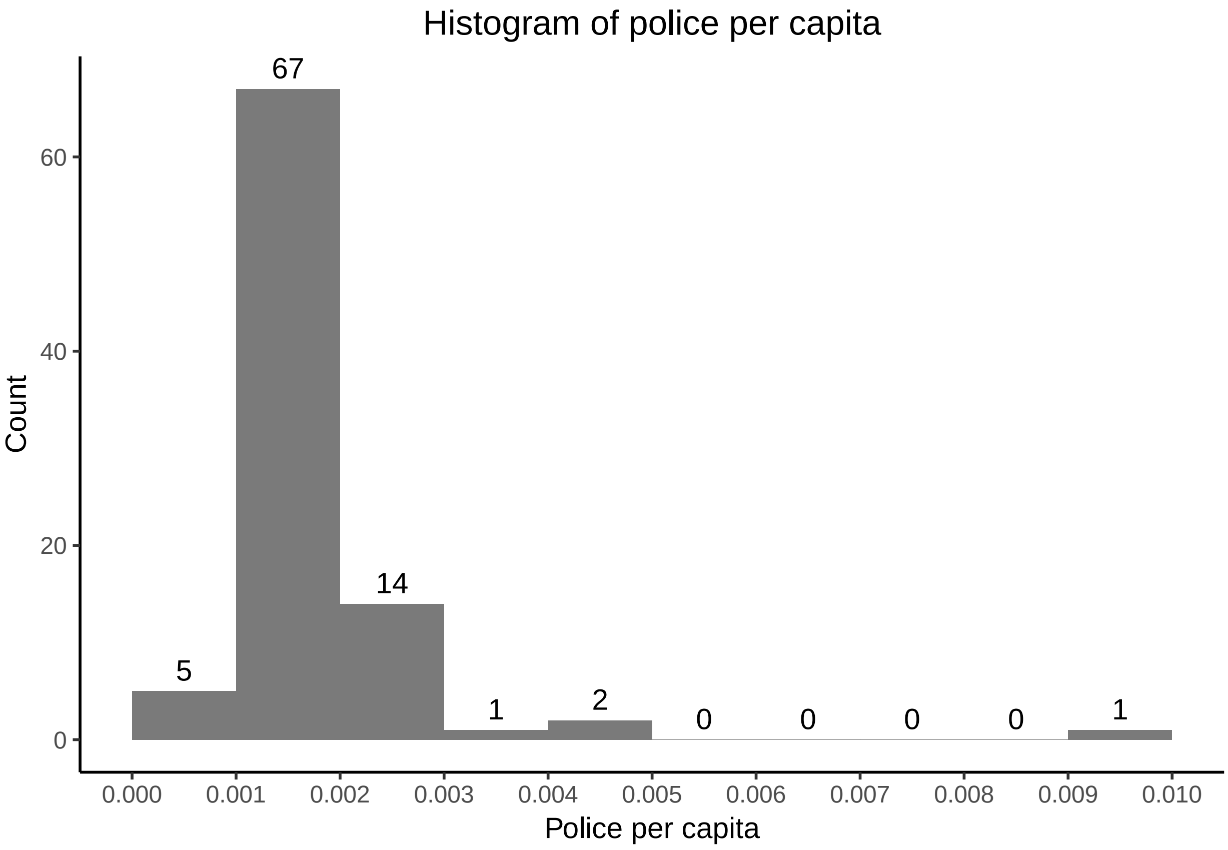
The state wage is somewhat right skewed, likely because regions with more highly ranked officials make more than most average employees. However, there is not any red flags. However, we see that there's a huge outlier in service industry wages, more than 10 fold. Very likely this is an error in which the decimal was shifted by 1. The county is Warren County, which is not known to have atypical service industry wages. Even if the data point is accurate, it may be significantly skewed due to non-random sampling of CEOs rather than general workers. Therefore, we strongly believe this data point is incorrect and does not represent our target population. We choose to fix this point by changing the value to the average wser across the state. After doing this, we then plot the all-to-all correlation matrix again.



A corrected wser now has a weak correlation with all other wage variables. wsta still represents a good independent predictor, and wfed is the single best predictor of crime in the wage category of variables.

Police per Capita

It was previously stated that more police could result in stronger detection of crime. This effect has been previously studied and observed for violent crimes [6]. Therefore we look into the effect of the police per capita variable.



Again, we see a significant outlier. However, we believe this could be real for the following reason.

```
##
## ===============================================
##                crmrte_abs prbarr prbconv polpc
## -----------------------------------------------
## outlier county   0.006    1.091   1.500  0.009
## state average    0.034    0.295   0.551  0.002
## -----------------------------------------------
```

The probability of arrest for this outlier is more than 3x state average, as is the probability of conviction. The police force is 4.5 times the state average. More police in a county means more crime gets detected and responded to in a timely fashion. Interestingly, the rate of crime is about 5-6x less than state average. This is likely a result of the fact that people are less likely to commit crime because they know the police is highly effective, and any crime that is caught will likely lead to more arrests. This is all to say that we actually believe this is a legitimate data point.

Demographic and Geographic Variables

We believe percent minority will be a good independent predictor of crime. Minorities are more likely to be targeted for the police, especially in rural regions [7]. Young male won't be considered a priority because the nature of the crime is not specified. If the data was only for example, petty theft, perhaps young male would be a good predictor.

We saw previously that eastern and western N.C. had apparently different crime rates, and will explore the effects of these variables within the context of our model.

Prison and prison sentence

Due to the significance of the fear variables we included in model 1, we also believe that probability and length of prison sentence are both potentially important and would like to evaluate these variables co-variates in model 2.

Model 2 Variables

At this point, we've outlined a few variables both from model 1 and EDA that we think would be fruitful to include in our model 2. In order to perform variable selection in an automated fashion, we begin with a global AIC optimization using a combination of forward and backward selection, and then apply hypothesis testing to further adjust our model.

```
##
## Call:
## lm(formula = y ~ prbarr + prbconv + polpc + taxpc + pctmin80 +
##     pctymle + wfed + wsta, data = X_stepwise)
##
## Coefficients:
## (Intercept)       prbarr      prbconv        polpc        taxpc
##   -8.971281    -2.170226    -0.785200   161.991529     0.006035
##    pctmin80      pctymle         wfed         wsta
##    0.011242     3.568045     1.374461    -0.503116
```

First, we happily see that the insignificant wcon wage variable (and the somewhat signficant wmfg) in our first model did not show up in this AIC optimized model. The AIC mixed model selected a lot of the same variables from our EDA, with the addition of two variables, percent young male, and tax revenue per capita. The model did not include the probability of prison and length of prison sentence. We will evaluate these with separate specifications.

Hypothesis Testing for further model and variable selection

First, we test the significance of the coefficients generated by the AIC minimized model. Note that we have yet to evaluate the CLM for the AIC.mixed model. However, due to asymptotics, we know that the t-test can be used as long as we use robust standard errors in case we have hetereoskedasticity.

For all the betas, we will use a 2-sided test as significance level 0.05:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

To perform the test for all of the variables, we use the coeftest package, specifying the degrees of freedom as sample size - 8 (number parameters except beta_0) - 1, and the heteroskedasticity-consistent estimation of the covariance matrix.

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -8.9712809   2.3524359 -3.8136 0.0002662 ***
## prbarr       -2.1702256   0.3073045 -7.0621 5.115e-10 ***
## prbconv      -0.7852002   0.0990757 -7.9253 1.054e-11 ***
## polpc       161.9915286  42.8434794  3.7810 0.0002976 ***
## taxpc         0.0060345   0.0039908  1.5121 0.1343975
## pctmin80      0.0112424   0.0014907  7.5416 5.968e-11 ***
## pctymle       3.5680450   2.5768716  1.3846 0.1699643
## wfed          1.3744611   0.3245105  4.2355 5.987e-05 ***
## wsta         -0.5031155   0.2457556 -2.0472 0.0438783 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All coefficients are significant ($p<0.05$) with the exception of taxpc and pctymle. For all variable coefficient except taxpc and pctymle, we reject the null hypothesis and state that the coefficient is in fact not zero.

Same cannot be said about percent young male, and taxpc. However, the test evaluates each individual coefficient on their own. It could be that these two variables are jointly significant in the following fashion: both variables could just be proxies for number of families with children. People with families are more likely to vote in support of better schools and safety for their families, resulting in higher taxes. In this case, these variables could be potentially jointly significant. We check this with an F-test using linearhypothesis from the car package, specified below at alpha 0.05, with two-tailed test:

$$H_0 : \beta_{taxpc} = \beta_{pctymle} = 0$$
$$H_a : \text{H0 is not true}$$

```
## Linear hypothesis test
##
## Hypothesis:
## taxpc = 0
## pctymle = 0
##
## Model 1: restricted model
## Model 2: y ~ prbarr + prbconv + polpc + taxpc + pctmin80 + pctymle + wfed +
##     wsta
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     83
## 2     81  2 1.9224 0.1529
```

We see that the since the p-value is 0.1529, which means we again fail to reject H0. As a result, we have support that these variables do not have joint significance. We will remove these variables from our model as a result.

Next, we wanted to assess whether probability of conviction and length of prison sentence are important predictors. To do so, we specify a new model with these covariates:

We then perform t-test, analogous to that above for mixed.AIC under the null hypothesis that these coefficients are 0:

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -7.0655071   2.6470705 -2.6692  0.009184 **
## prbarr       -2.6217099   0.3509217 -7.4709 8.204e-11 ***
## prbconv      -0.8970018   0.1025211 -8.7494 2.487e-13 ***
## polpc       235.8846839  57.5891803  4.0960 9.902e-05 ***
## pctmin80      0.0121218   0.0015703  7.7194 2.677e-11 ***
## wfed          1.1875914   0.3555734  3.3399  0.001269 **
## wsta         -0.5119346   0.2398877 -2.1341  0.035862 *
## prbpris      -0.2382048   0.4114301 -0.5790  0.564217
## avgsen       -0.0044576   0.0122709 -0.3633  0.717349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To our surprise, the t-test produced very high value for these coefficients, which means we fail to reject H0 that the coefficients are in fact 0. Again, we test joint significance:

```
## Linear hypothesis test
##
## Hypothesis:
## prbpris = 0
## avgsen = 0
##
## Model 1: restricted model
## Model 2: data$logcrmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed +
##     wsta + prbpris + avgsen
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     83
## 2     81  2 0.2116 0.8097
```

The p-value is very large. As a result, we have evidence there is not joint significance and will not include these variables in our final model.

We saw from the EDA that many of the wage variables, with the exception of wsta, were correlated with each other. Even though each of the individuals variables did not make into our AIC model, could those excluded be jointly significant? We can test this by generating a model specfication with all wage variables. The null hypothesis for the test below is that all wage variables have a coefficient of 0:

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   -9.600942   3.590043 -2.6743  0.009163 **
## prbarr        -2.586265   0.388889 -6.6504 3.971e-09 ***
## prbconv       -0.869426   0.106655 -8.1518 5.638e-12 ***
## polpc        221.014206  66.361760  3.3304  0.001340 **
## pctmin80       0.012334   0.001701  7.2509 2.948e-10 ***
## wfed           0.875603   0.434071  2.0172  0.047208 *
## wsta          -0.394959   0.288797 -1.3676  0.175469
## wcon           0.281991   0.249611  1.1297  0.262145
## wtuc           0.072343   0.226892  0.3188  0.750720
## wtrd           0.234918   0.289443  0.8116  0.419544
## wfir          -0.188005   0.355142 -0.5294  0.598085
## wser          -0.208284   0.269147 -0.7739  0.441410
## wmfg           0.102478   0.159551  0.6423  0.522619
## wloc           0.341645   0.671749  0.5086  0.612513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking the significance of each coefficient, none of the newly included wage variables are statistically significant on their own. We now check joint signficance with an F-test, under the null that there is no joint significance:

```
## Linear hypothesis test
##
```

```
## Hypothesis:
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wloc = 0
##
## Model 1: restricted model
## Model 2: data$logcrmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed +
##     wsta + wcon + wtuc + wtrd + wfir + wser + wmfg + wloc
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     83 5.5752
## 2     76 5.1863  7    0.3889 0.8141 0.5784
```

These new wage variables were also not joint significantly due to a large p-value. Therefore, in accordance with the concept of parsimony, we will not include any other wage variables in our model.

At this point, in every test so far, we see that prbarr and prbconv are consistently highly statistically significant. This brings up an interesting question as to whether an interaction exists between these two variables. Consider the two counties below:

```
##
## ======================
##   county prbarr prbconv
## ----------------------
## 1   99   0.154   1.234
## 2   51   0.155   0.260
## ----------------------
```

Both have nearly equivalent prbarr, but county 99 has a much higher prbconv. We believe that the effect of prbarr may be higher in county 99 due to an interaction between the two variables. The reason is that if criminals know they will be rarely convicted, they may not care very much if they get arrested, whereas if they know they will be convicted nearly every time, then the same prbarr will be a stronger deterrent.

We can test this hypothesis by including an interaction term into our regression, and perform hypothesis testing under the null hypothesis that the coefficient of the interaction is 0:

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     -7.1855001   2.6004932 -2.7631  0.007066 **
## prbarr          -2.4371217   0.8577007 -2.8415  0.005664 **
## prbconv         -0.8288363   0.3571120 -2.3209  0.022774 *
## polpc          243.5384580  76.9304428  3.1657  0.002172 **
## pctmin80         0.0120104   0.0016917  7.0995 4.111e-10 ***
## wfed             1.1728582   0.3569706  3.2856  0.001499 **
## wsta            -0.5105240   0.2259974 -2.2590  0.026539 *
## prbarr:prbconv  -0.2620833   1.6823107 -0.1558  0.876583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the p-value is 0.88, and as a result, fail to reject H0 that the interaction term is not 0. As a result, we will not include this interaction term in our final model.

Based on experimenting with these alternative specifications, our AIC model, and statistical testing, we now arrive at our proposed model 2.

To further assess model_2, we want to use the RESET test to check our variable specification, and see if whether polynomial terms should be included in our model to improve its predictive power. At a significance level of 0.05:

$$\begin{aligned} H_0 &:\text{second order polynomial not needed} \\ H_a &:\text{second order polynomial is needed} \end{aligned}$$

```
##
##  RESET test
##
## data:  model_2
## RESET = 0.70955, df1 = 6, df2 = 77, p-value = 0.6429
```

Based on the RESET test, we see that polynomial terms will not help improve our model due to a large p-value.

Classical Linear Model Assumptions

We now access the CLM assumptions of our final model 2.

CLM 1 and 2 are identical to our model 1.

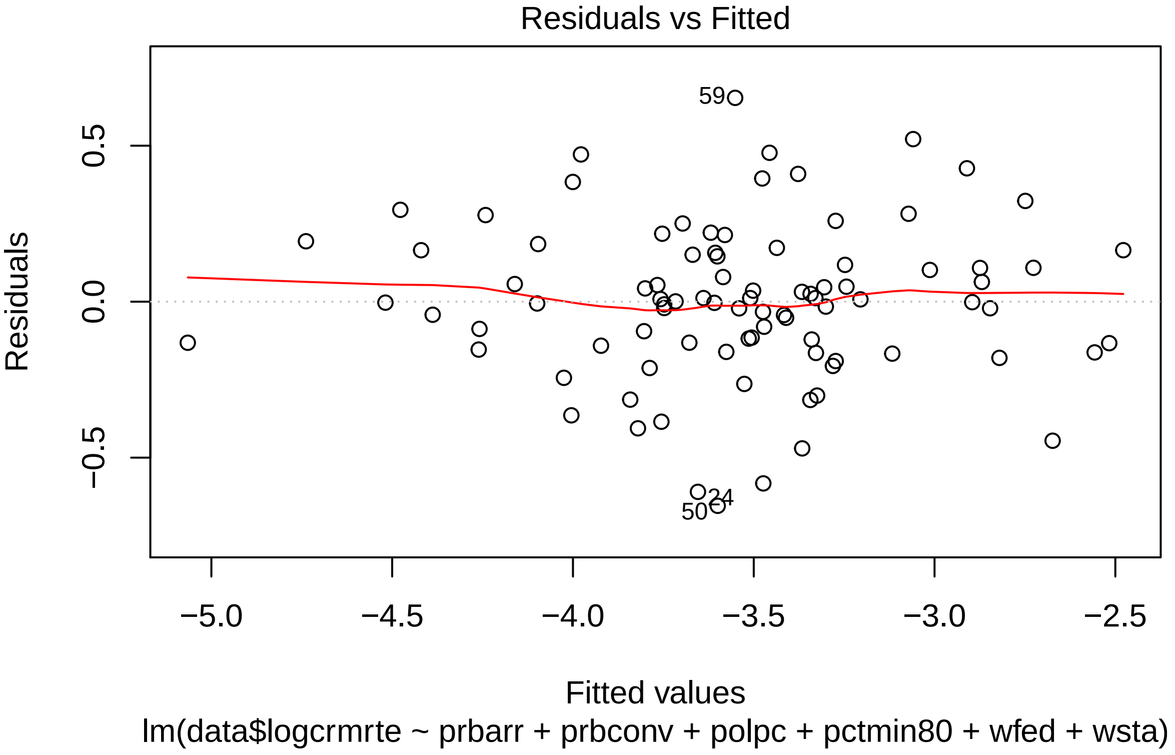
CLM 3: No perfect multi-collinearity

R confirmed no perfect multi-collinearity. The VIFs are also all significantly below 4, which means we do not have multi-collinearity to worry about:

```
##   prbarr  prbconv    polpc pctmin80     wfed     wsta
## 1.481696 1.120632 1.527763 1.103844 1.152948 1.095947
```

CLM 4: Zero Conditional Mean

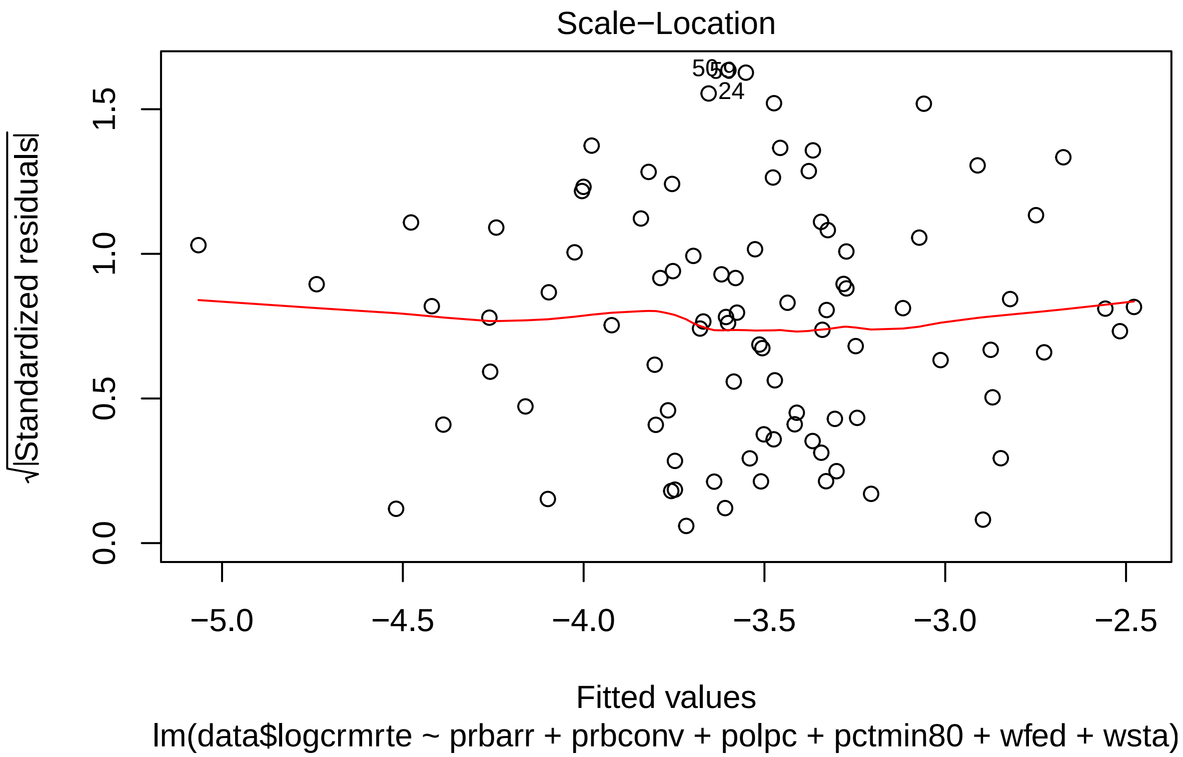
To check zero conditional mean, we plot the residual against the fitted values for our set.



Based on this plot, we see that the line is very much linear even at extreme values. The average value of the residual is approximately 0 for all fitted vales. Zero Conditional Mean is met for our model of 6 variables.

CLM 5: Homoskedasticity

Examining the fitted values versus residuals plot, the spread is slightly larger around fitted values of around 3.5 (around -0.9 to 0.6) than around 4 (around -0.4 to 0.5). We can also check the scale-location plot:



We see that this line is roughly horizontal from -5 to -3. The only major curvature is the single data point around -4. This indicates that we most likely have close to homoskedasticity.

One way to test for homoskedasticity is the Breusch-Pagan Test. The null hypothesis of the test states that we have homoskedasticity. We will test at a standard significance level of 0.05.

$$H_0 : \text{Homoskedasticity}$$
$$H_a : \text{Heteroskedasticity}$$

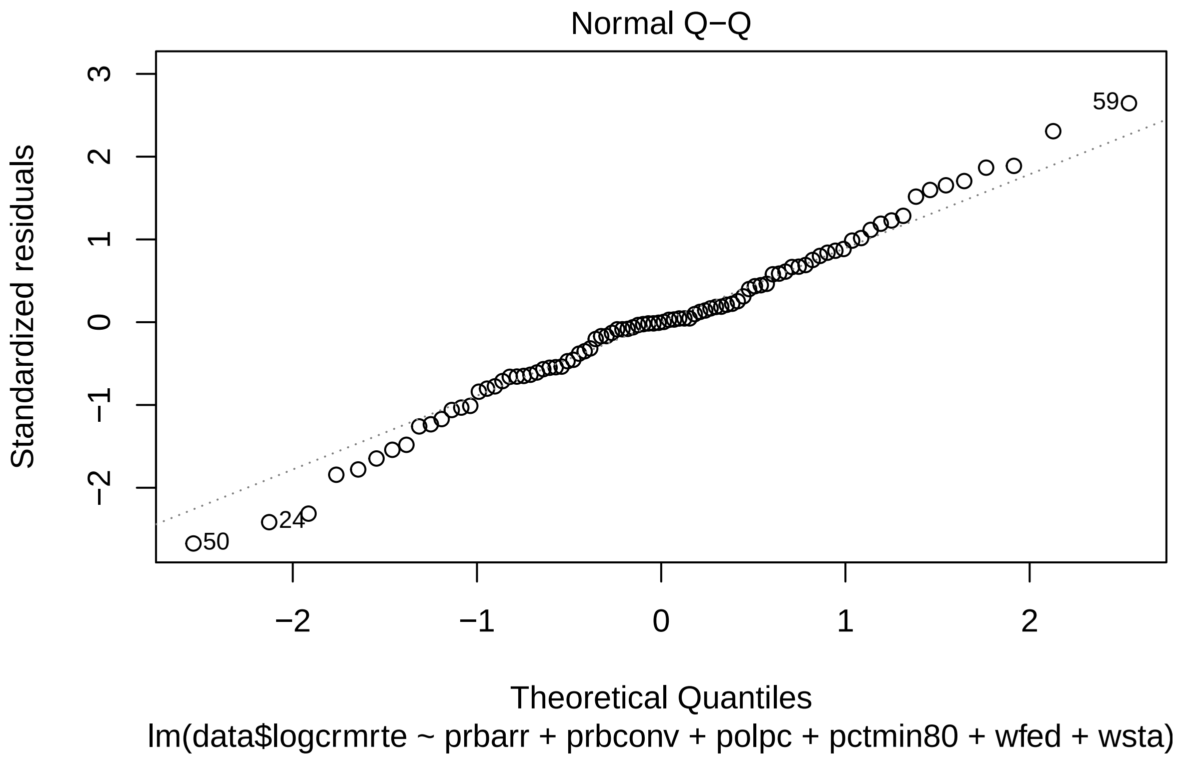
```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 22.302, df = 6, p-value = 0.001067
```

Since the $p - value < 0.05$, we reject the null hypothesis that we have homoskedasticity. Our sample size is relatively small, so the test suggests we have a major deviation from homoskedasticity.

Since our visualization and hypothesis test gave inconsistent results, we will use heteroskedastic robust errors for reporting and statistical testing.

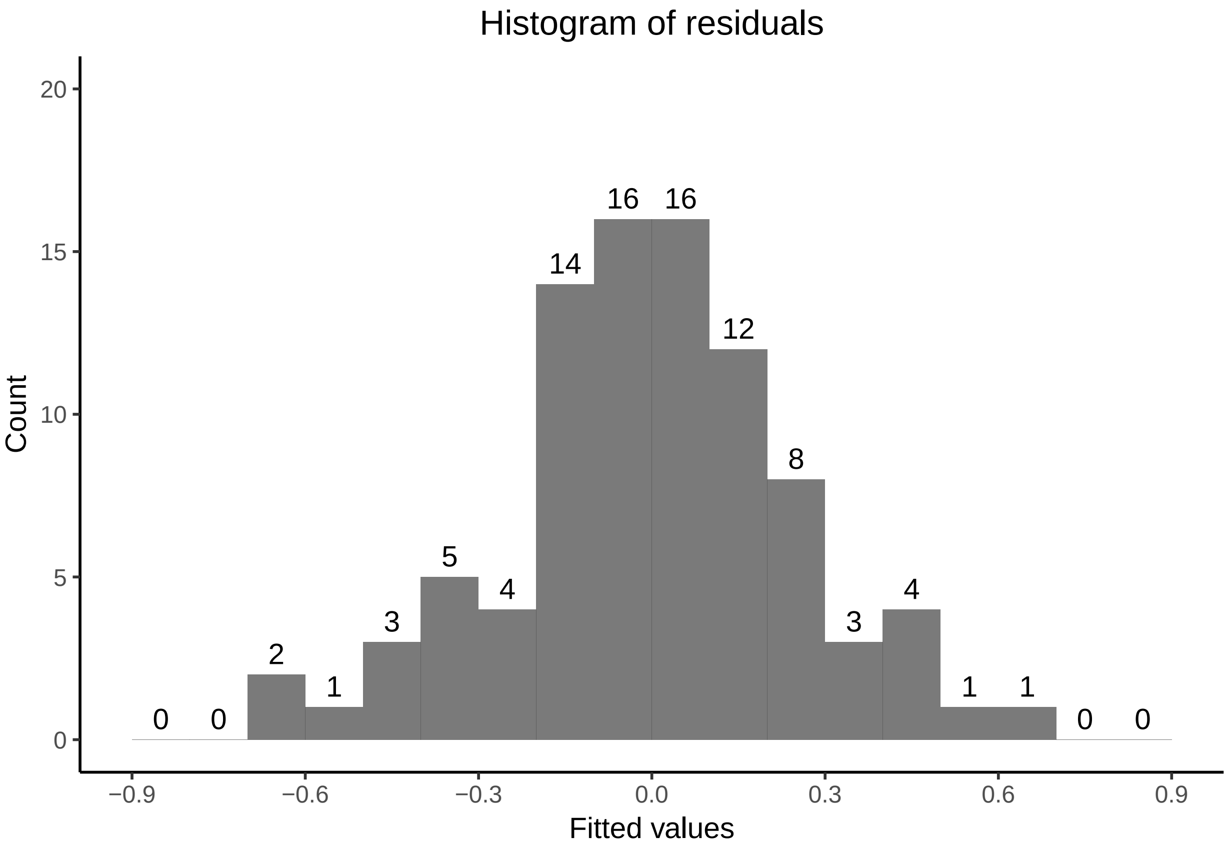
CLM 6: Normality

CLM 6 assumes that population error is independent of the explanatory variables x_1 through x_k, and that the error term is normally distributed with mean 0 and constant variance. We can check this with the qqplot of the fitted values versus residuals plot.



Not counting the extreme values, the data points wavering back and forth, which could indicate a kurtosis problem. Also, most differ from where we would like them to be on the line, so this indicates we most likely do not have normality of errors.

We can visualise the residuals in a histogram.



Based on the histogram, the data does not appear very normal. The tail seems to taper off too fast from the peak in the middle. In any case, since our sample size 90 is much greater than 30, asymptotics also kicks in, ensuring that the sampling distribution of our coefficients are approximately normal.

Arrest versus conviction

In order to fully address our research question, we will test one more item about our model:

1. Is the effect of probability of arrest versus conviction the same on crime rate? If not, which fear variable is more important for deterring crime?

To do this, we will use the linearHypothesis test from the car package.

For both cases, we will test at a significance level of 0.05:

$$H_0 : \text{the two coefficients are the same}$$
$$H_a : \text{the two coefficients are different}$$

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr - prbconv = 0
##
## Model 1: restricted model
## Model 2: data$logcrmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed +
##     wsta
##
## Note: Coefficient covariance matrix supplied.
##
```

```
##   Res.Df Df      F    Pr(>F)
## 1     84
## 2     83  1 32.557 1.737e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the difference is highly statistically significant, so reject H0 state that we have evidence suggesting the effects are different.

```
##  (Intercept)       prbarr      prbconv        polpc     pctmin80
##  -7.00398772  -2.63251938  -0.90117838 230.04857582   0.01208904
##         wfed         wsta
##   1.16445048  -0.51967238
```

Based on the value of the coefficients, since the coefficient of probability of arrest is larger, this variable is a stronger deterrent than probability of conviction.

Interpretation

Keeping all other variables constant, a 1% increase in wfed results in an approximately 1.16% increase crime rate. State wage seems to have a negative effect: a 1% increase state wage, keeping all other variables constant, results in a -0.52% change in crime rate. For all other variables, we take the exact transformation below as we did in model 1:

$$\%\Delta \text{crmrte} = 100 * [\exp(\beta * \Delta \text{variable}) - 1]$$

We do this in one step in R, looking at a 0.1 increase for the arrest and conviction variables, a 0.1 increase in pctmin80, and a 0.001 increase in polpc. We chose these values based on the size of the coefficients and provide the proper interpretation below:

```
##      prbarr     prbconv    pctmin80       polpc
## -23.1451754  -8.6176504   0.1209635  25.8661149
```

In this model, keeping all other variables constant, increasing prbarr by 0.1 units decreases crime by 23.1%, and increasing prbconv by 0.1 units decreases crime by 8.62%. These all seem like significant decreases in crime rate with a relatively small increase in the "fear factors." In addition, a 0.1 increase in pctmin80, or 1 additional minority per 1000 non-minority, increase crime by only 0.12%. This does not seem like a practically significant amount for a relatively small increase in minority population. Finally, a 0.001 increase in polpc, or an additional police officer for every 1000 people, will increase measured crime rate by 25.9%. This seems like a huge increase.

We now have information for fully address our first two subquestions. We save recommendations for our final section of the report.

1. Fear of arrest and convinctions are consistently the best deterrents of crime. Compared to each other, fear of arrest is statistically more signficant and also by a practical margin. By increasing arrest probability by 0.1, our model predicts a 23.1% decrease in crime, most likely noting that greater arrests leads to bigger sense of police presence, and a greater fear that crime committed will lead to an arrest.

Police per capita is strongly correlated with crime rate. The more police presence, the more likely crime is detected and responded to. The more crime that is present in an area, the more police are necessary to keep order. Furthermore, we saw that police has one key outlier: county 51 has very high police per capita, high arrest and conviction rates, but low crime rate. This is likely an outlier where the capacity for police to respond to crime has met or exceed the rate of crime.

2. Overall, most of the wage variables do not significantly influence crime rate, either on their own, or jointly within our models. The exception are federal wages and state wages. Federal wages in a county are positively related to crime, while state wages are negatively related with crime.

One interpretation could be that monetary incentives at the state level might be in place for individuals who are able to effectively combat crime (one example of this is the State Patrol). A higher the wage in one area for state employees could mean that more crime-combating individuals are present in that county, leading to lower crime rates, and higher overall pay as a result of those employees.

A positive correlation between wfed and crime rate could be a result of an omitted variable: general wealth level of the county. Higher ranking federal officials are much more likely to congregate in wealthy areas compared to poorer areas, and wealthier areas may be subject to more white-collar crime. We don't think that increasing federal salary would increase crime, but that the relationship is a result of a third variable.

Model 3

The purpose of our third model is to address our final research question, which is whether our parsimonious model 2 is robust to other covariates as measured by predictive power, and inference on the coefficients. We are especially interested in whether including density, the very strong single best predictor of crime rate, would greatly change our model conclusions. For model 3, we will include all co-variables.

First we evaluate whether this generates a reasonable AIC using the classical penalty term of k=2, and adjusted r squared, a parsimony-adjusted correlation coefficient.

```
##
## ==============================
##                Model 3 Model 2
## ------------------------------
## Adj. R Squared  0.816   0.777
## AIC            16.574  21.075
## ------------------------------
```

We see that including all covariates improves both the AIC and adjusted R^2, but not by much. However, the interpretation of Model 3 is much less clean than model 2 due to the presence of many more variables. Adding back variables such as density does not significantly alter our adj. R^2.

We then look at how much coefficients themselves actually change.

```
##
## =================================================================
##             model_2 coefficients model_3_coefficients percent_change
## -----------------------------------------------------------------
## (Intercept)       -7.004               -8.274           18.136
## prbarr            -2.633               -1.915          -27.259
## prbconv           -0.901               -0.687          -23.758
## polpc            230.049              156.394          -32.017
## pctmin80           0.012                0.009          -24.202
## wfed               1.164                1.027          -11.819
## wsta              -0.520               -0.395          -24.047
## -----------------------------------------------------------------
```

Our coefficients have changed between ~20%-30%. Depending on the individual coefficient, this could represent a relatively large change. The largest percent change was in police. Rather than a 25.9% increase in crime for every 1 additional police officer for 1000 people, we now have a 16.9% increase using the same calculations shown previously.

One very important fact is that none of the signs on the coefficients have changed, meaning the direction of influence was correctly predicted in model 2. By omitting the variables present in the dataset in model 2, we did not get different conclusions in terms of the direction of influence, although the magnitude of the effect was different than in model 3.

We now evaluate the CLM assumptions.

Classical Linear Model Assumptions

CLM 1 and 2 are identical to our model 1.

CLM 3: No perfect multi-collinearity

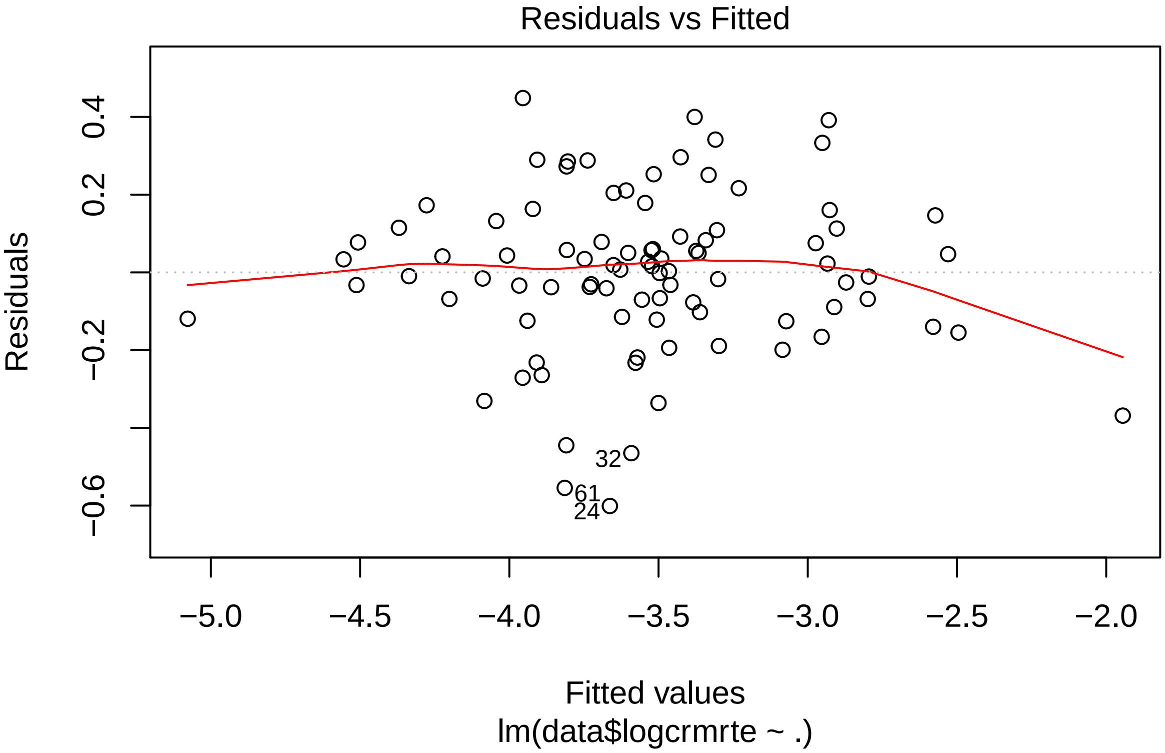
R would have warned us if this were the case that we had perfect multi-collinearity, so in this case we have fulfilled this requirement.

```
##   prbarr  prbconv  prbpris   avgsen    polpc  density    taxpc     west
## 2.364452 1.761170 1.218374 1.825254 2.990818 5.443675 1.972683 3.271060
##  central    urban pctmin80      mix  pctymle     wcon     wtuc     wtrd
## 2.080564 3.957133 2.912945 1.922204 1.503969 2.211167 1.768986 3.065321
##     wfir     wser     wmfg     wfed     wsta     wloc
## 2.596418 2.384276 2.277978 3.076669 1.640593 2.447422
```

We see that all VIF factors except for density and urban are significantly below 4. Since we will not use model 3 for recommendations, but only as a validation that model 2 is robust, this high VIF factor is ok for our purposes.

CLM 4: Zero Conditional Mean

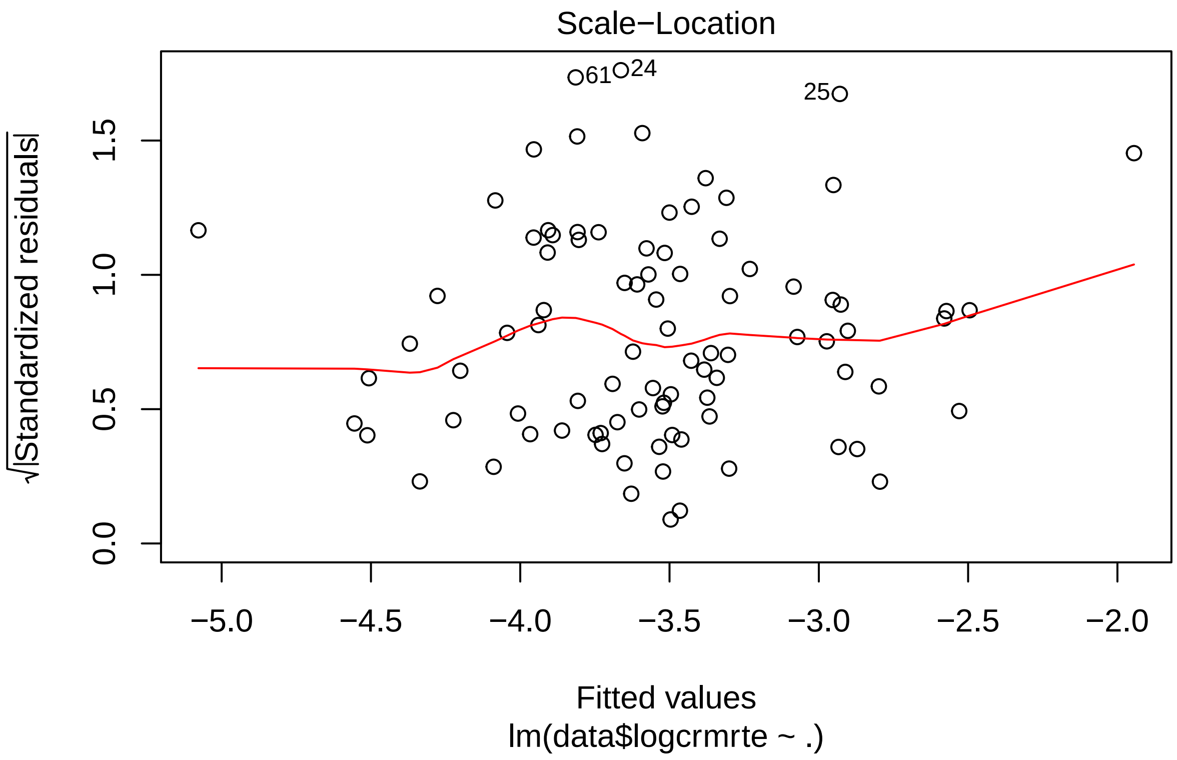
We plot the residual against the fitted values for our set.



We see that the line is very much linear except at the one extreme value of -2 fitted values. This is due to a single data point. At all other fitted values, the model looks decent. As a result, Zero Conditional Mean is met for our model of all variables.

CLM 5: Homoskedasticity

Examining the fitted values versus residuals plot above, we see that the spread appears to be slightly larger around fitted values of around -3.75 than around -4.25. We can also check the scale-location plot:



We see that this line is wavy from -5 to -2. This indicates that we most likely do not have homoskedasticity.

We check with the Breusch-Pagan Test at a standard significance level of 0.05.

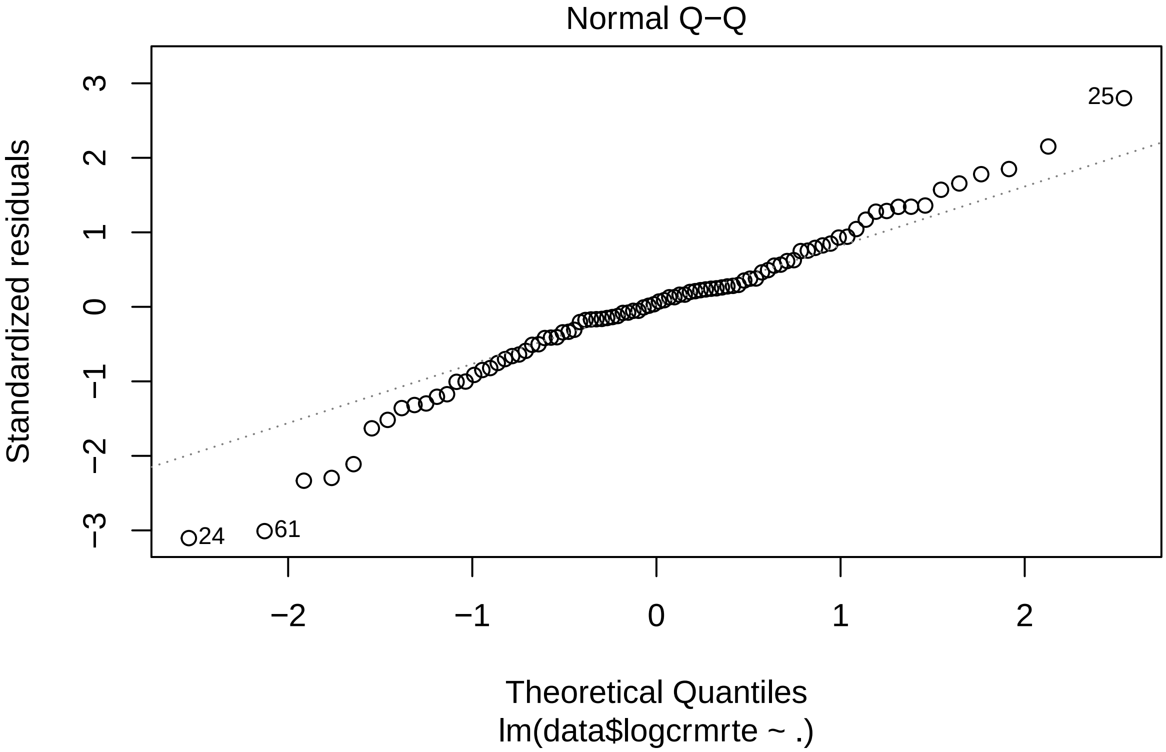
$$H_0 : \text{Homoskedasticity}$$
$$H_a : \text{Heteroskedasticity}$$

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 30.19, df = 22, p-value = 0.1139
```

Since the $p - value > 0.05$, we fail to reject the null hypothesis that we have homoskedasticity, which gives us conflicting results. For our purposes, we will report heteroskedastic robust errors since we are not entire sure whether this assumption is fulfilled.

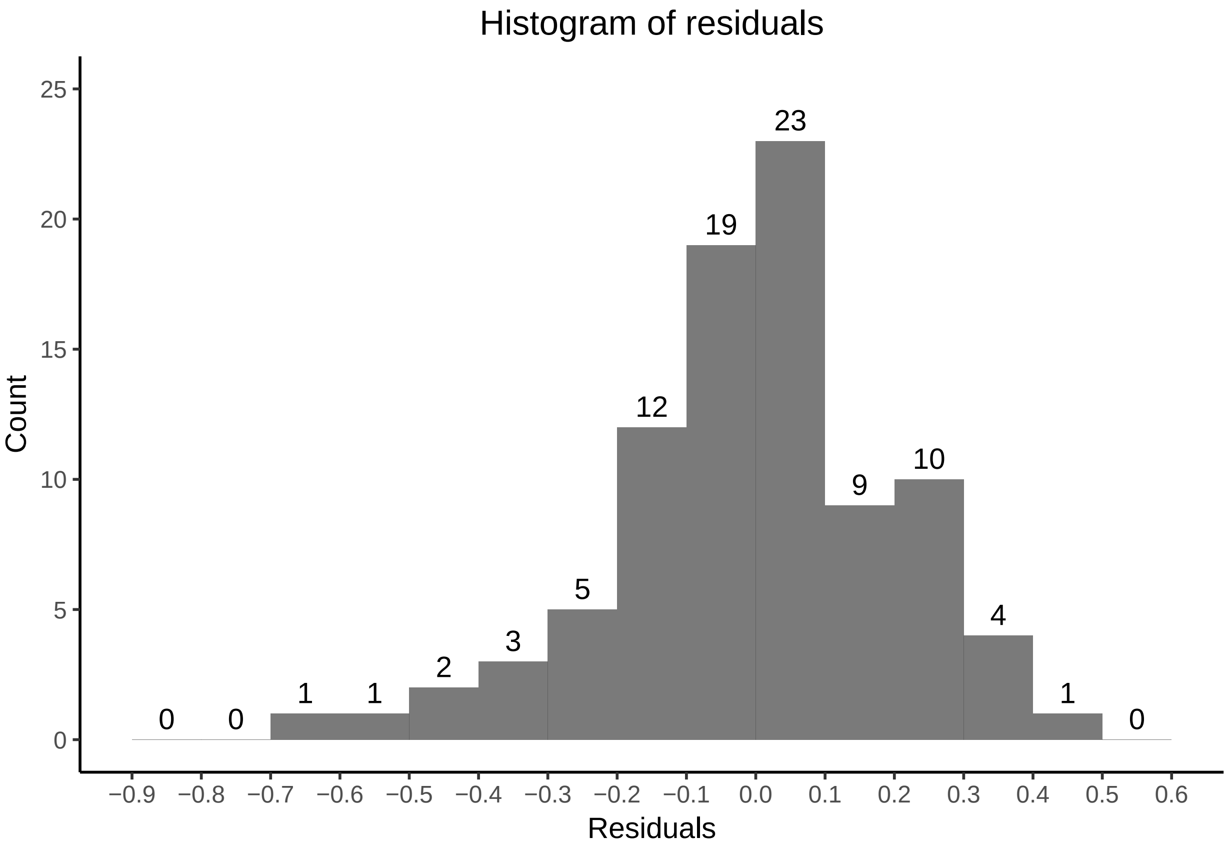
CLM 6: Normality

CLM 6 assumes that population error is independent of the explanatory variables x_1 through x_k, and that the error term is normally distributed with mean 0 and constant variance. We can check this with the qqplot of the fitted values versus residuals plot.



Not even counting the exception of extreme values, the data points generally do not lie on the line except toward the 0 quantiles. This indicates we most likely do not have normality of errors.

We can visualise the residuals in a histogram.



Based on the histogram, the data does not appear very normal with possibly two peaks close to each other. Since our sample size 90 is much greater than 30, asymptotics also kicks in, ensuring that the sampling distribution of our coefficients are approximately normal.

Interpretation

At this point, we can perform hypothesis testing with robust standard errors to see which variables are significant when all co-variates are included in the model.

For all the betas, we will use a 2-sided test as significance level 0.05:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -8.2742468   4.3051934 -1.9219  0.058870 .
## prbarr       -1.9149339   0.3866997 -4.9520 5.246e-06 ***
## prbconv      -0.6870780   0.1312302 -5.2357 1.785e-06 ***
## prbpris      -0.1409586   0.3522110 -0.4002  0.690275
## avgsen       -0.0115162   0.0144097 -0.7992  0.426998
## polpc       156.3939355  87.3184840  1.7911  0.077798 .
## density       0.1122222   0.0587685  1.9096  0.060472 .
## taxpc         0.0033990   0.0071706  0.4740  0.637023
## west         -0.1232345   0.1133665 -1.0870  0.280912
## central      -0.1265106   0.0858633 -1.4734  0.145327
## urban        -0.1678000   0.2261444 -0.7420  0.460678
## pctmin80      0.0091632   0.0027567  3.3240  0.001441 **
## mix          -0.1688287   0.6232296 -0.2709  0.787306
## pctymle       3.1656293   1.3998550  2.2614  0.026986 *
## wcon          0.1976172   0.2421260  0.8162  0.417292
## wtuc          0.1343003   0.2817222  0.4767  0.635119
## wtrd          0.1199324   0.3726951  0.3218  0.748608
## wfir         -0.2044104   0.3918277 -0.5217  0.603610
## wser         -0.4304160   0.2936497 -1.4657  0.147395
## wmfg          0.0351942   0.1812773  0.1941  0.846649
## wfed          1.0268212   0.4488481  2.2877  0.025319 *
## wsta         -0.3947069   0.3223806 -1.2244  0.225108
## wloc          0.3070301   0.6867545  0.4471  0.656264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this model, probability of arrest and probability of conviction are still both highly statistically significant. Percent minority, federal wages, and interestingly, percent young male, are both statistically significant at the 0.05 significance level. Police per capita, and density, are only significant at the 0.1 significance level.

The only variable not in model 2 but is significant in model 3 is percent young male. Aside from this, no additional variables were individually significant. We now perform the Wald test to see if the addition variables included in model 3 are jointly statistically significant. We use the robust standard errors. We use a significance level of 0.05, and our hypotheses are as follows:

$$H_0 \text{ :additional model 3 variables are not jointly signficant}$$
$$H_a \text{ :additional model 3 variables are jointly signficant}$$

```
## Wald test
##
## Model 1: data$logcrmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed +
##     wsta
## Model 2: data$logcrmrte ~ prbarr + prbconv + prbpris + avgsen + polpc +
##     density + taxpc + west + central + urban + pctmin80 + mix +
##     pctymle + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
##     wsta + wloc
##   Res.Df Df     F Pr(>F)
## 1     83
## 2     67 16 1.293 0.2279
```

Based on the Wald test, since the p-value > 0.05, we fail to reject H0 and have found evidence that the additional variables in model 3 are not jointly significant. This would indicate that model 2 is as good as model 3 in terms of capturing statistically significant predictors, both individually and jointly.

With these results, we answer our final research question:

3. By including only 6 co-variates, prbarr, prbconv, polpc, pctmin80, wfed, wsta, we have built a model that is robust to a model with 22 co-variates. Our model has very similar measures of fit. While the magnitude of the coefficients change by 20-30% across the board, both models provide the same conclusions for policy due to the fact that the sign on our coefficients are the same across the board. Finally, we see from hypothesis testing that the additional covariates are not jointly significant, and all but one of the covariates are actually statistically significant on their own. As a result, we believe that our model 2 is both predictive and robust, generalizable to both regions with and without high population density. Model 2 is our preferred model in terms of making policy recommendations.

Regression Table

We now present a regression table for our 3 models. We have discussed the practical significance of our variables are part of the model building process, and concluded that prbarr, prbconv, polpc are both practical significant, while pctmin80 was not practically significant. Please see model sections for details.

Table 1: Regression to predict log of crimes per person from co-variates

	Dependent variable:		
	Log of crimes per person		
	(1)	(2)	(3)
Constant	−8.434*** (1.905)	−7.004** (2.592)	−8.274 (4.305)
prbarr	−1.681*** (0.457)	−2.633*** (0.356)	−1.915*** (0.387)
prbconv	−0.707*** (0.145)	−0.901*** (0.101)	−0.687*** (0.131)
wfed		1.164*** (0.351)	1.027* (0.449)
wsta		−0.520* (0.223)	−0.395 (0.322)
polpc		230.049*** (57.520)	156.394 (87.318)
pctmin80		0.012*** (0.001)	0.009*** (0.003)
wmfg	0.541* (0.249)		0.035 (0.181)
wcon	0.470 (0.316)		0.198 (0.242)
prbpris			−0.141 (0.352)
avgsen			−0.012 (0.014)
wtuc			0.134 (0.282)
wtrd			0.120 (0.373)
wfir			−0.204 (0.392)
wser			−0.430 (0.294)
density			0.112 (0.059)
taxpc			0.003 (0.007)
west			−0.123 (0.113)
central			−0.127 (0.086)
urban			−0.168 (0.226)
mix			−0.169 (0.623)
pctymle			3.166* (1.400)
wloc			0.307 (0.687)
Observations	90	90	90
R^2	0.544	0.792	0.861
Adjusted R^2	0.522	0.777	0.816
Note:			*p<0.05; **p<0.01; ***p<0.001

Additional Omitted Variable Bias

While omitted variables have been discussed throughout our report, we focus our attention to global variables that might be influencing all of our models here.

In our first regression model, we found that non-customer facing blue collar wages in construction and manufacturing were not significant in predicting crime rate. An omitted variable influencing these variables is education. We believe that counties with high educations will have higher wages in these sectors, perhaps because more of these individuals make it to management level, or are more efficient at their jobs. Education should also be negatively correlated with crime. One reason could be that those individuals have more exposure to ethics as education level increases. For the two specifications below, imagine that the second model is the true causal model:

$$crmrte = \delta_0 + \delta_1 * wconcrmrte = \beta_0 + \beta_1 * wcon + \beta_2 * edlvl$$

The omitted variable bias can be derived from the regression of edlvl on wcon :

$$edlvl = \alpha_0 + \alpha_1 * wcon$$

In this case, the omitted variable bias is beta2 * alpha1. We believe that alpha1 is positive and beta2 is negative, so the bias is negative. This means that the observed coefficient is smaller than the actual. For wcon (positive), the analysis suggests the size of the coefficient is even more positive, so this variable might in fact become significant as a result of including education level. Of course, this analysis is greatly simplified. In reality, education levels will likely correlate with many of the independent variable and can have much more profound effects.

Another source of omitted variable bias is what we will call family happiness. Counties with more individuals who are happily married with kids might receive a very high score, while those with more individuals who recently divorced may have a low score. We expect the happier someone is with their family, the less likely they will commit any type of crime. Counties with happier families could see higher overall wages as well because those who are happy at home tend to be more productive and likable at the work place. This results in a negative omitted variable bias, and a smaller observed coefficient compared to the true casual model. This could influence recommendations for wage changes, since increases might have a stronger effect than anticipated.

We saw that police per capita was a positive predictor of crime, but an omitted variable that strongly correlates with police per capita is the general effectiveness of police officers. This can be measured by how long it takes officers to resolve similar cases across counties, how many officers are required for similar tasks, etc. We expect the more effective the police force, the less police needed to manage crime in a county. Furthermore, the more effective the police force, the lower the crime rate, due to faster and more effective approaches to controlling crime and dealing with criminals. Therefore, the omitted variable bias is positive, meaning the observed coefficient on police per capital is larger than what we would expect with this variable included. This coefficient is currently very large, and unfortunately, we have no way of knowing whether the true coefficient is even positive given this bias. As a result, effectiveness of the police force is a necessary metric to even more accurately measure the effect of police per capita on crime rate.

Cost of living is another omitted variable. We would expect a positive correlation between wage and cost of living (jobs in the cities tend to pay more). Additionally, cost of living likely is positively correlated with crime rate, as people are more likely to be below the poverty line and in need of more income. This omitted variable would have a positive bias on the coefficient of our wage variables, artificially inflating them in our model. Since wage of federal employees is a strong predictor of crime in our model, this effect may be diminished if all federal employees also work in regions of high cost of living.

Substance abuse is a major driver behind certain forms of crime. A closely related factor is the ability to access rehabilitation. The higher the ability, the greater chance the individual can re-enter society, and lower the crime. Studies show that while minorities are much less likely to find accessible care due to financial and social factors [8]. As a result, we expect the ability to recover from substance abuse is negatively correlated with percent minority. The omitted variable bias is positive, meaning the observed coefficient on percent minority is smaller than the true causal model. We showed this variable was statistically significant, but not highly significant. Adding this omitted variable could drive the coefficient toward 0. We would certainly like to make policy changes such that in the true causal model of the world, race is not a significant predictor of crime rate.

Policy Recommendations and Concluding Remarks

Full discussion of our research questions 1 and 2 were included in the discussion of model 2. Our research question 3 was discussed as part of model 3. Below is a brief summary of our conclusions.

How does fear of arrest and convictions deter crime across North Carolina?

We showed that in all models, arrest and conviction probabilities were highly statistically significant deterrents of crime, and that arrest was statistically even more significant than convictions. Within this, police per capita had a moderating effect. The more police, the greater the rate of arrest and convictions, and greater (what we propose) the crime detection rate.

How does wage for all types of employees influence overall crime rate?

Federal wages and state wages were statistically significant. State wages were negatively correlated with crime rate, and we believe this could in part be due to that fact that state employees who are effective at combating might be rewarded with monetary gains, leading to higher compensation for regions with lower crime due to the presence of these individuals. The positive correlation of federal wages with crime is likely due to the fact that high ranking federal officials tend to aggregate in wealthier areas.

What variables are strong predictors of crime, and at the same time are robust across the entire state, irrespective of location and population density?

According to the developed model, all variables in model 2 are significant, and robust in terms of fit and coefficient signs to adding in all co-variables from model 3.

From these results, we propose the following policy recommendations for the North Carolina campaigner focusing on crime:

1) Revision of state law to reinforce arrest and conviction. Specially, since the perception of the likelihood of arrest and conviction are both strong deterrents of crime, police and the legal system should be less lenient toward arresting and convicting detected criminals. If the campaign were to focus on a single aspect, arrest probability is more important than convictions. This is good news because arrest is usually cheaper and faster to carry out, requiring less legal workers compared to convictions. Since we believe that these variables are significant due to fear of being caught, greater publicity of arrest and conviction through media may serve the same purpose.

2) We see that the police per capita is a positive predictor of crime in all regions except for one where the police number are extremely high. This is likely due to the fact that increased police force increases the level of crime detection. For the one county with very high police per capita, it is likely that the police levels exceed what is needed to detect and handle crime in a timely manner. We recommend increase the police workforce in order to detect more of the crime, which could also increase the opportunities for recommendation #1 (greater publicity of arrests). In addition, this would increase the level of perceived and measured security, and hence, could be used as an argument for a direct improvement on the life quality of voters.

3) While minorities percent is only somewhat statistically significant, systematic difference in crime rate between racial groups presents a social problem that a political campaign can strongly leverage. We also noted that inability to rehabilitate from drug abuse is a potential omitted variable bias. We recommend that further studies should be conducted on counties with large minority groups, and actions be take appropriately. For example, it may be possible that stronger job placement programs for minorities would decrease the chance that someone that in the population would commit a crime, and better rehabilitation programs might even eliminate race as a predictor for crime rate.

4) We see that state salaries is negatively correlated with crime. One reason could be that state officials, such as the State Patrol, who are effective at combating crime are rewarded with higher pay, resulting in counties with less crime appearing to have higher state pay. We encourage policies that increase the salary and bonuses of workers in the criminal field, including the police and state patrol officials. Such as policy change would also be viewed in a positive light by the public and beneficial for the campaige.

5) Finally, our model 2 with only 6 co-variates retained the predictive power of model 3 with 22 co-variates, which included regional factors such as west versus central, as well as population differences such as density. As a result, we recommend a more state-wide approach to criminal policy change rather than regional changes. Our 6 co-variates are independent of region and density, and are furthermore robust to them, indicating that these 6 underlying factors are important predictors of crime regardless of regional differences in crime patterns.

References

[1] 'The Preassure Cooker: Population Density and Crime'. Steinmetz, David (Jul 20, 2016). NYC Data Science Academy. https://nycdatascience.com/blog/student-works/pressure-cooker-higher-population-densities-increase-crime/

[2] 'Taxing, Spending, and Voting: Voter Turnout Rates in Statewide Elections in Comparative Perspective'. Garrick L. Percival, Mary Currin-Percival, Shaun Bowler and Henk van der Kolk State & Local Government Review Vol. 39, No. 3 (2007), pp. 131-143 https://www.jstor.org/stable/25130415?seq=1

[3] 'Criminal Behavior, Sanctions, and Income Taxation: An Economic Analysis'.Avraham D. Tabbach. The Journal of Legal Studies. Vol. 32, No. 2 (June 2003), pp. 383-406

[4] 'Gender and Crime: Toward a Gendered Theory of Female Offending'. Darrell Steffensmeier and Emilie Allan. Annual Review of Sociology Vol. 22 (1996), pp. 459-487 https://www.jstor.org/stable/2083439?seq=1

[5] 'Crime and the minimum wage'. Christine Braun. Review of Economic Dynamics Volume 32, April 2019, Pages 122-152. https://www.sciencedirect.com/science/article/pii/S1094202518302941

[6] 'Why the Police Have an Effect on Violent Crime After All: Evidence from the British Crime Survey'. Ben Vollaard and Joseph Hamed. The Journal of Law & Economics. Vol. 55, No. 4 (November 2012), pp. 901-924 https://www.jstor.org/stable/10.1086/666614?seq=1

[7] 'There's overwhelming evidence that the criminal-justice system is racist. Here's the proof'. Radley Balko. The Washington Post. September 18, 2018. https://www.washingtonpost.com/news/opinions/wp/2018/09/18/theres-overwhelming-evidence-that-the-criminal-justice-system-is-racist-heres-the-proof/

[8] "Addiction Among Different Races." Sunrise House, Editorial Staff. July 9, 2019. https://sunrisehouse.com/addiction-demographics/different-races/.

*I thank Gaby May-Lagunes for her invaluable contributions.