Lab 3

W203 Statistics for Data Science

Section 6 - Team 1 - Stone Jiang, Gabriela May Lagunes, Indrani Bose

Introduction

The objective of this exercise was to create a regression model with crime rate as independent variable. The purpose of this process was to identify the variables that most affect the crime rate in different counties of North Carolina, in order to then derive appropriate policy recommendations for a political campaign. The focus of this work is answering the following three research questions. Firstly: *What are the best independent predictors of crime rate?* The analysis and discussions derived from this question gave origin to the following two: *How does wage influence crime rate?* and, *Does fear of getting in trouble with the police deter crime?* The attempt to answer these questions derived the presented development of the regression model for crime rate, and justified the recommendations given at the end of the report. The development of the final model was 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 variable of interests is crime rate. This is because being able to model crime rate can indicate policy makers which metrics should they focus on in order to improve the level of security of their states. Therefore, the goal of the developed models is to find the best possible causal predictors for crime rate.

Before choosing the best dependent variables for our models, the data was cleaned as follows. First, we omitted the last rows of the csv since they do not contain data. Second, we eliminated duplicated entries. Here there was just one county (193) which was repeated twice. Then, we verified the datatype of our variables. Here, the prbconv variable was the only non-numerical variable because it was stored as a factor due to omitted rows having non-numerical values. This was converted to numerical.

```r
library(dplyr)
library(ggplot2)
library(tidyr)
library(caret)
library(MASS)
select <- dplyr::select # Unmask select from dplyr
library(stargazer)
library(tibble)
library(grid)
library(gridExtra)
library(usmap)
library(car)
library(sandwich)
library(lmtest)
library(reshape2)

setwd("~/Desktop/Stone/Berkeley_MIDS/Statistics/Labs/Lab_3")
#setwd("~/Desktop/Berkeley/W203 - Statistics/github/W203_Lab_3/")
full_data <- read.csv('crime_v2.csv')
data <- na.omit(full_data)
```

```r
# Check for duplicated data and remove duplicates
sum(duplicated(data))
```

```
## [1] 1
```

```r
data <- distinct(data, .keep_all=T)

# Convert prbconv factor in numeric
data$prbconv <- as.numeric(levels(data$prbconv))[data$prbconv]

# Check that all fields are numerical
for (field in names(data)) {
  stopifnot(class(data[,field]) %in% c("numeric", "integer"))
  }
```

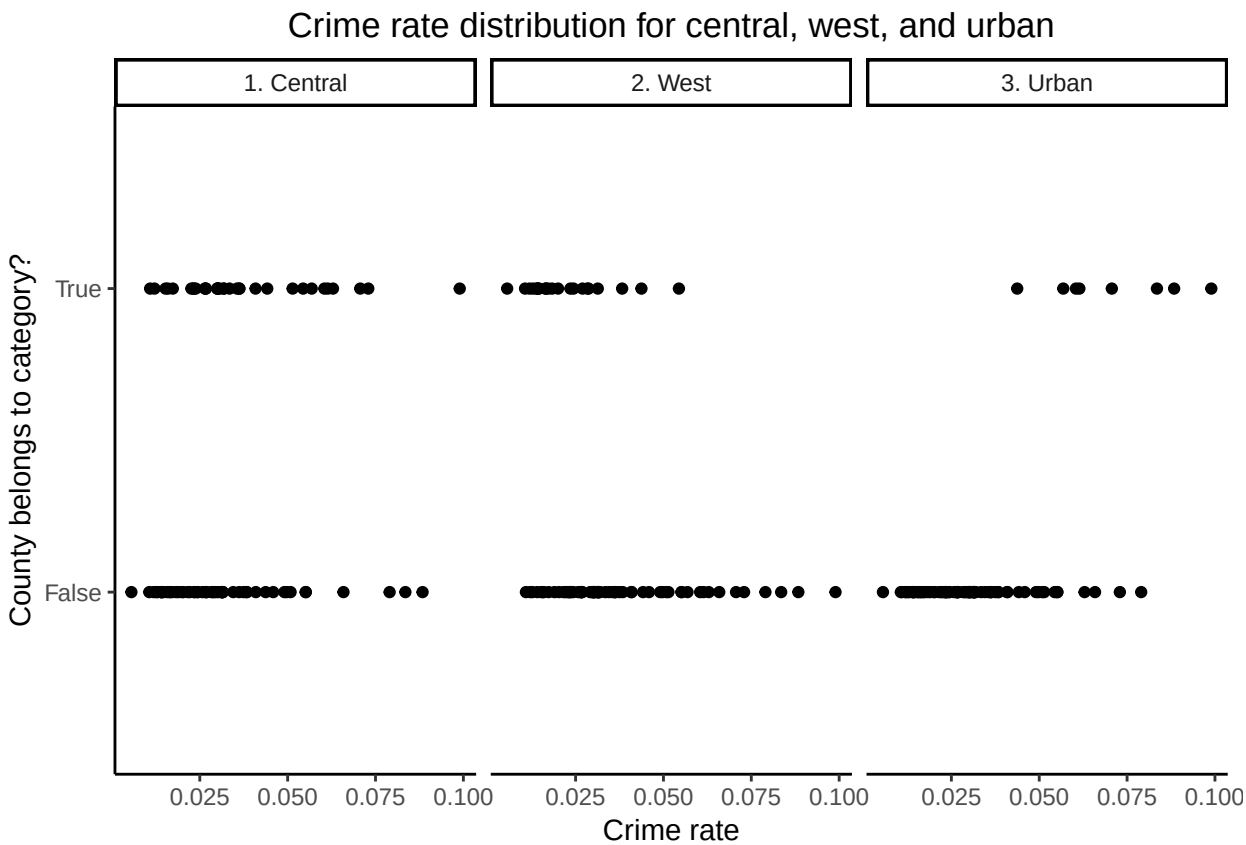
After this initial data cleaning, we identified columns we believed could be potential casual predictors of crime rate. N.B. county and year were disregarded in our models because they are identifiers.

The variables west, central and urban are categorical, which have been one-hot encoded. These can be used directly in the regression. We first saw whether the distribution of crime rate is different depending on the location (west vs central) and whether the county was urban.

```r
C <- data[, c('west', 'central', 'urban')]

df.categorical <- data[, c('crmrte', 'west', 'central', 'urban')]
colnames(df.categorical) <- c('crmrte', '2. West', '1. Central', '3. Urban')
dt_long <- gather(df.categorical, key, value, -crmrte)

ggplot(dt_long, aes(x = crmrte, y = factor(value))) +
  geom_point() +
  facet_grid(. ~ key) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Crime rate distribution for central, west, and urban",
       x='Crime rate',
       y = "County belongs to category?") +
  scale_y_discrete(breaks=c(0,1), labels=c("False","True"))
```



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 than those not. We see definitive summaries below for the urban and west variables below.

```
u <- data %>%
  group_by(urban) %>%
  summarise(mean_crime_rate = mean(crmrte)) %>%
  as.data.frame()
u$urban <- c('N', 'Y')
w <- data %>%
  group_by(west) %>%
  summarise(mean_crime_rate = mean(crmrte)) %>%
  as.data.frame()
w$west <- c('N', 'Y')
print(u)
```

```
##   urban mean_crime_rate
## 1     N      0.02990170
## 2     Y      0.07049427
```

```
print(w)
```

```
##   west mean_crime_rate
## 1    N      0.03720145
## 2    Y      0.02209975
```

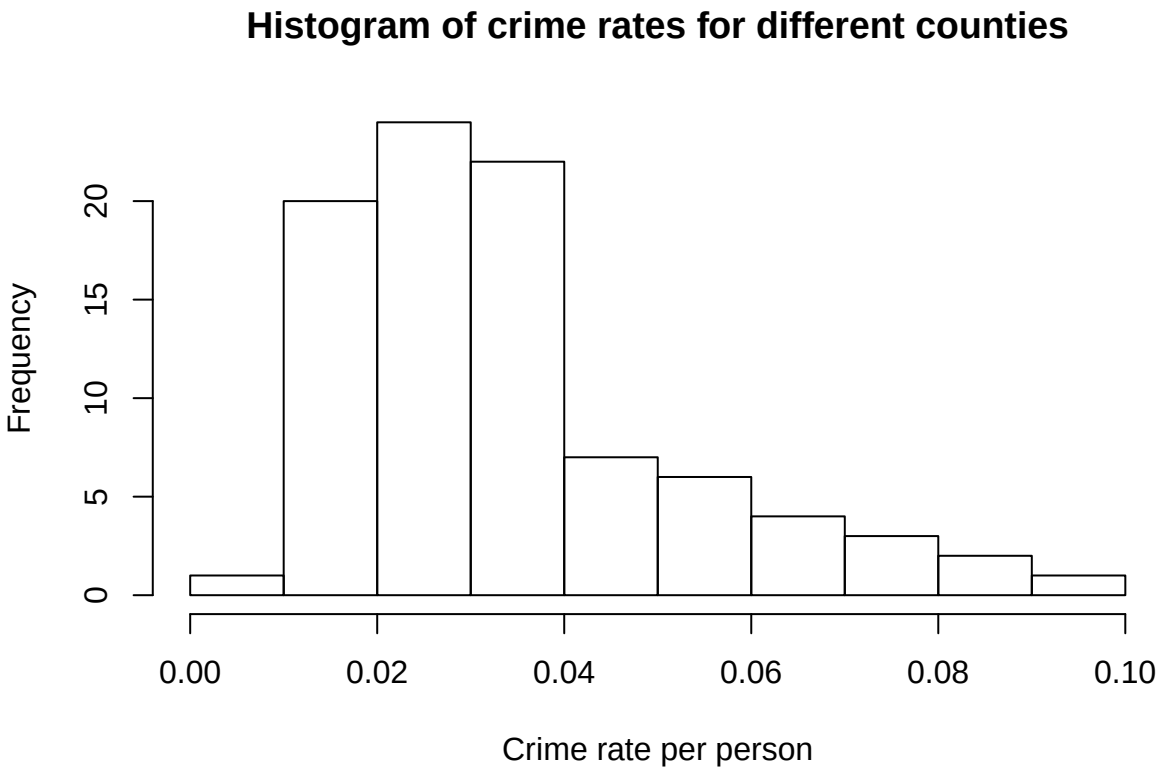
All other variables are numeric, and are also possible candidates for influencing crime rate.

Then, we examined the crime rate variable on its own.

```
summary(data$crmrte)
```

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

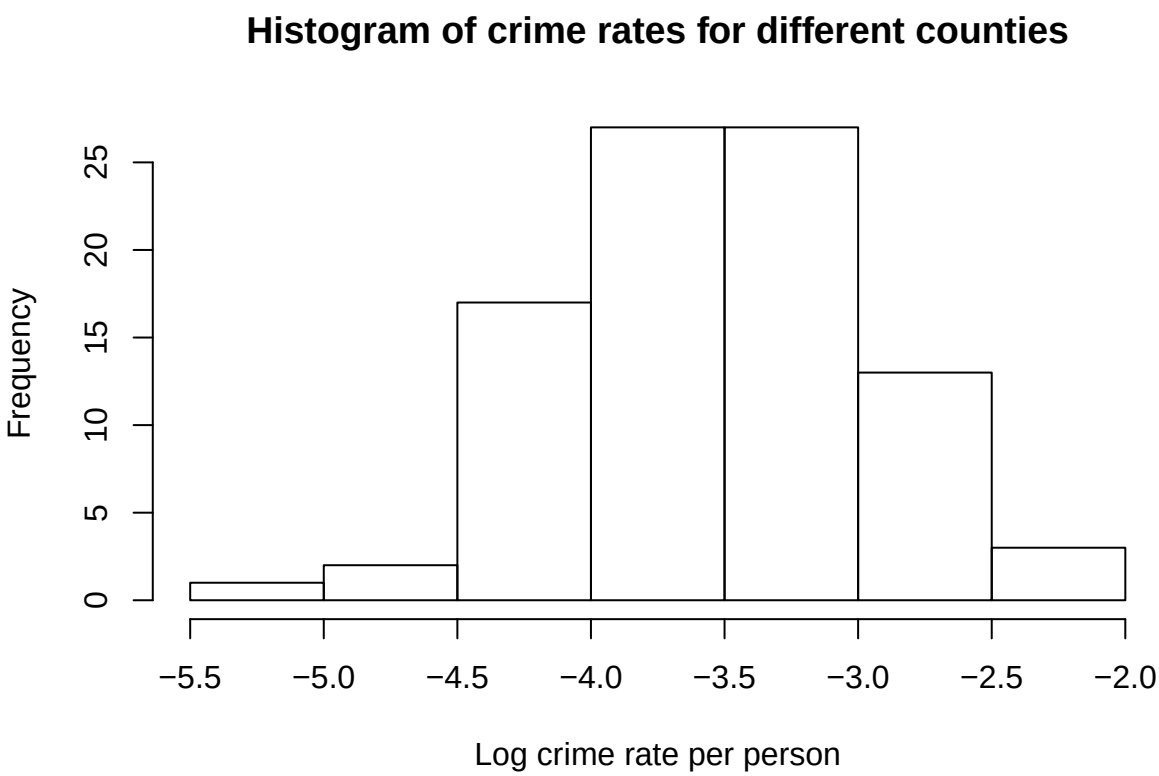
```
hist(data$crmrte,
     main='Histogram of crime rates for different counties',
     xlab = 'Crime rate per person')
```



From this plot, it is possible to see that crime rate is greater than 0 for all counties and highly skewed toward larger values. Since inference analysis was performed 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, however, it would be beneficial to transform crime rate into the log of crime rate. This changed the interpretation from absolute changes in crime rates to percent changes (at least for small changes in crime rate since the percent interpretation is only accurate for differentially small changes), which makes comparisons across counties more accurate. For example, a 0.01 change in crime rate for the lowest county (0.005) is a much larger change than for the largest county (0.099), but a 1% change is comparable regardless of the crime rate starting point. As a result, we performed inference on the log of crime rate.

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

```
data$crmrte_abs <- data$crmrte
data$crmrte <- log(data$crmrte)
hist(data$crmrte,
     main='Histogram of crime rates for different counties',
     xlab = 'Log crime rate per person')
```



As it can be observed, the distribution of the logarithms of crime rate follow a closer to normal distribution. This is 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. We will call these variables the "fear factors;" namely, 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 they believe the punishment to be (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. Many criminals are likely not thinking about possibility or severity of prison sentences because they might feel even if arrested they can talk their way out of it. For heavy repeat offenders, they will likely prison sentence into account, but possibility of arrest is still a heavy influencer. As a result, we believe arrest and conviction are the most relevant variables.

The wage variables can either deter or motivate individuals to commit a crime and were excluded from this base model. 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 having to pursue 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, 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. As a result, we believe wage can go either way. For our base model, we will look at only what we consider traditional blue collar jobs: construction and manufacturing. We also take the log of these variables: this is common practice as we want to measure the effect of a percent changes in salary, and not absolute changes.

```
nonwage_variables <- c('prbarr', 'prbconv', 'prbpris', 'avgsen',
                       'polpc', 'density', 'taxpc',
                       'pctymle', 'pctmin80', 'mix',
                       'urban', 'central', 'west')

wage_variables <- c('wtrd', 'wfir', 'wser', 'wfed', 'wsta', 'wloc',
                    'wcon', 'wtuc', 'wmfg')

X_non_wage <- data[, names(data) %in% nonwage_variables]
X_wage <- lapply(data[, names(data) %in% wage_variables], log)

X_wage_transformed <- cbind(X_non_wage, X_wage)
```

Before performing EDA on the variables listed above, we note why we have chose to exclude the other variables in our base model.

Additional variables for Model 1

We believe that density should be a positive predictor of crime. Highly dense population areas present more opportunities for crime. There also tends to be a larger wage and wealth gap in these areas, which increases the rate of crime as people will be tantalized to use illegal ways to get to the top. However, this may absorb too much of the causal effect.

We also believe that the police per capita is a key variable that would absorb too much of the model. Namely, more police are required for regions of greater crime, and so counties with more crime are more likely to have more police. In addition, the fact that there's more police could mean that more crime is detected and responded to, increasing the recorded number of criminal cases and perceived rate of crime. However, we

also expect there to be a tipping point. If the density of police is extremely high, that likely acts as a major deterrent for criminals. As a result, police and crime rate are very intricately linked and want to avoid this for our base model.

Location could be important an important variable to consider. Different geographic areas may be more prone to crime due to cultural and socioeconomic differences. However, we would like a model that can be generalised, so location is left aside for now.

Male absorbs the causality because then it will only reflect that. It is universally accepted among criminologists that women are always less likely to commit crimes than men, so a higher percentage of male population will be closely linked to crime rate [1].

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.

Variables giving information about demographics (like minority, male) certainly increase prediction, but these are better moderator variables, than mediator variables. In social sciences, this means that these variables affect the relation between an independent variable and the dependent variables of a model [4]. We left them aside on the first model in order to have a common ground for all our key variables.

Finally, we also assume that the variable mix does not have a great effect because we are interested on crime rate regardless of the nature of the offence for the first 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.

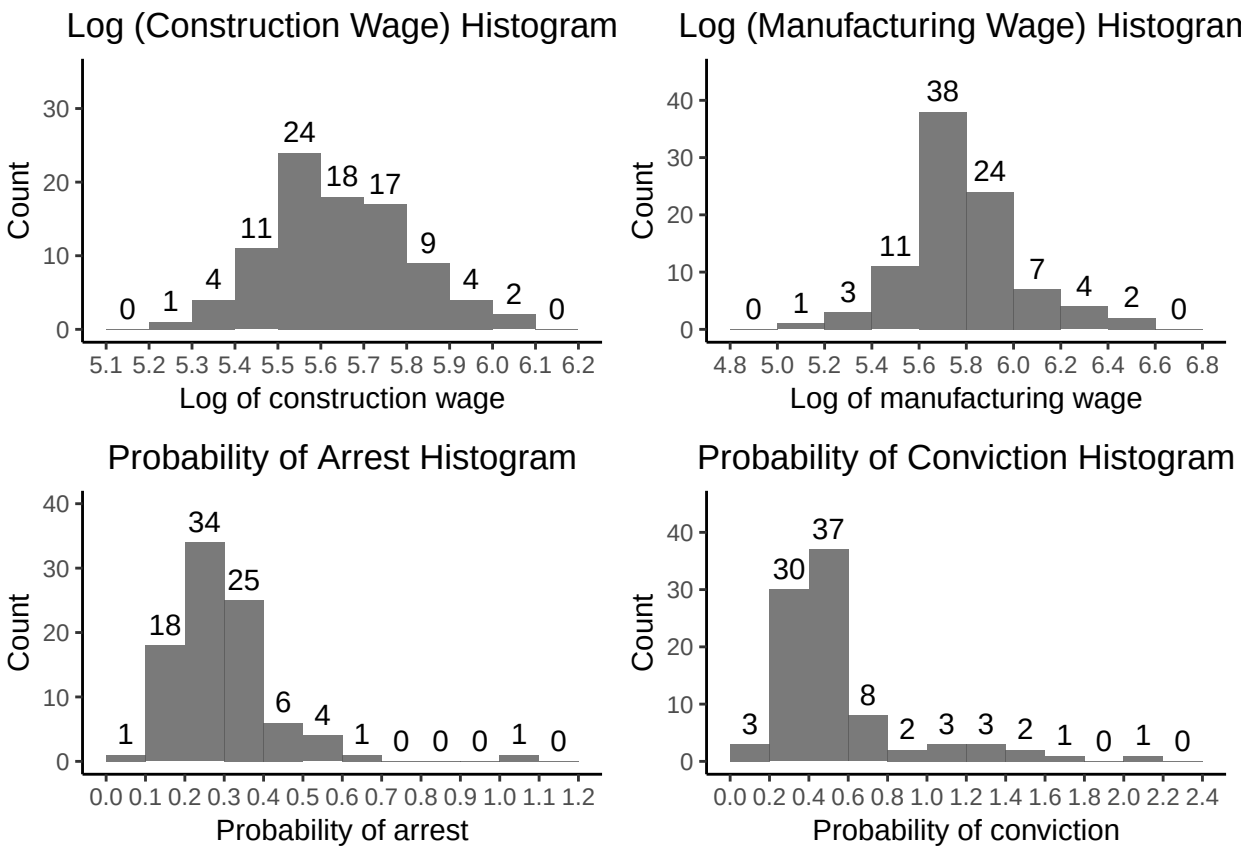
Exploratory Data Analysis

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

```
hist.wcon <- ggplot(data = X_wage_transformed, aes(x=wcon)) +
  geom_histogram(alpha=0.8, breaks=seq(5.1, 6.2, by=0.1)) +
  labs(title='Log (Construction Wage) Histogram',
       x='Log of construction wage',
       y = "Count") +
  theme_classic() +
  ylim(0,35)+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(5.1, 6.2, by=0.1)) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=seq(5.1, 6.2, by=0.1))

hist.wmfg <- ggplot(data = X_wage_transformed, aes(x=wmfg)) +
  geom_histogram(alpha=0.8, breaks=seq(4.8, 6.8, by=0.2)) +
  labs(title='Log (Manufacturing Wage) Histogram',
       x='Log of manufacturing wage',
       y = "Count") +
  theme_classic() +
  ylim(0,45)+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(4.8, 6.8, by=0.2)) +
  stat_bin(aes(y=..count.., label=(..count..)),
```

```
           geom="text",
           vjust=-.5,
           breaks=seq(4.8, 6.8, by=0.2))
hist.prbarr <- ggplot(data = X_wage_transformed, aes(x=prbarr)) +
  geom_histogram(alpha = 0.8, breaks=seq(0,1.2,0.1)) +
  labs(title = "Probability of Arrest Histogram",
       x = "Probability of arrest",
       y = "Count") +
  theme_classic() +
  ylim(0,40)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(0,1.2,0.1)) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=seq(0, 1.2, by=0.1))
hist.prbconv <- ggplot(data = X_wage_transformed, aes(x=prbconv)) +
  geom_histogram(alpha = 0.8, breaks=seq(0,2.4,0.2)) +
  labs(title = "Probability of Conviction Histogram",
       x = "Probability of conviction",
       y = "Count") +
  theme_classic() +
  ylim(0,45)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(0,2.4,0.2)) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=seq(0, 2.4, by=0.2))
grid.arrange(hist.wcon, hist.wmfg,
             hist.prbarr, hist.prbconv,
             nrow=2, ncol=2)
```



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. Both of the probability variables are upward skewed.

For the probability of arrest, defined as the number of arrests to offences, the general trend is a skew to the right, and there is one data point that lies above 1. One possible 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, even though we have one outlier, we will leave in this data point in our regression analysis.

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. The most common form of this is a citation, which are issued in place of arrests for smaller crimes. As a result, we believe it is reasonable for this variable to exceed 1.

Coefficient Interpretation

For our first model, the coefficient in prbarr represents the effect of probability of arrest on crime rate. Specifically, keeping all other variables constant, per unit increase in the probability of arrest leads to a certain percent change in crime rate. Since this variable represents the "fear factor" we presented above, we would hope that this change is negative. An analogous interpretation can be said for probability of conviction. With these variables, we want to measure how a perceived probability of getting in trouble with the legal system deters crime.

The coefficients on the wage variables represents how a small percent change in average wage in that blue collar industry relates to a small percent change in crime rate. This variable can really go both ways: a higher wage could mean that potential criminal are satisfied with their income and would pursue alternative

methods for achieving their goals. Alternatively, higher wage could mean less jobs for potential criminals. We now build our regression model.

```
#first model
model_1 <- lm(data$crmrte ~ prbarr + prbconv + wcon + wmfg, data = X_wage_transformed)
print(model_1)
```

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

We see that the coefficients on the "fear" variables are both negative, meaning that these factors negative influence crime rate. Since the coefficients are all rather large, we cannot directly interpret this in terms of a percent change. We can instead note that increasing either of independent variables by a single unit is a significant amount. For example, increasing the probability of arrest from 0.5 to 1.5 means a 3x increase in probability of arrest, which would require drastic changes and efforts on the part of law enforcement. As a result, we will interpret the coefficients in terms of a 0.1, or 10% increases.

Namely, keeping all other explanatory variables constant, we see that a 0.1 unit increase in the probability of arrest, or a 10% change, leads to a 0.17 units of decrease in the the log of crime rate (or very roughly 17% decrease in crime rate). Similarity, a 0.1 unit increase in the probability of arrest leads to a 0.07 units of decrease in the log of crime rate, or about a 7% decrease in crime rate.

For the wage variables, the pattern is in the opposite direction. Keeping all other explanatory variables constant, for each 0.1 log unit of increase in wage of construction, we see approximately a 0.047 log units of increase in crime. Similarity, for each 0.1 log unit of increase in wage of manufacturing jobs leads to 0.054 log units of increase in crime.

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.

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

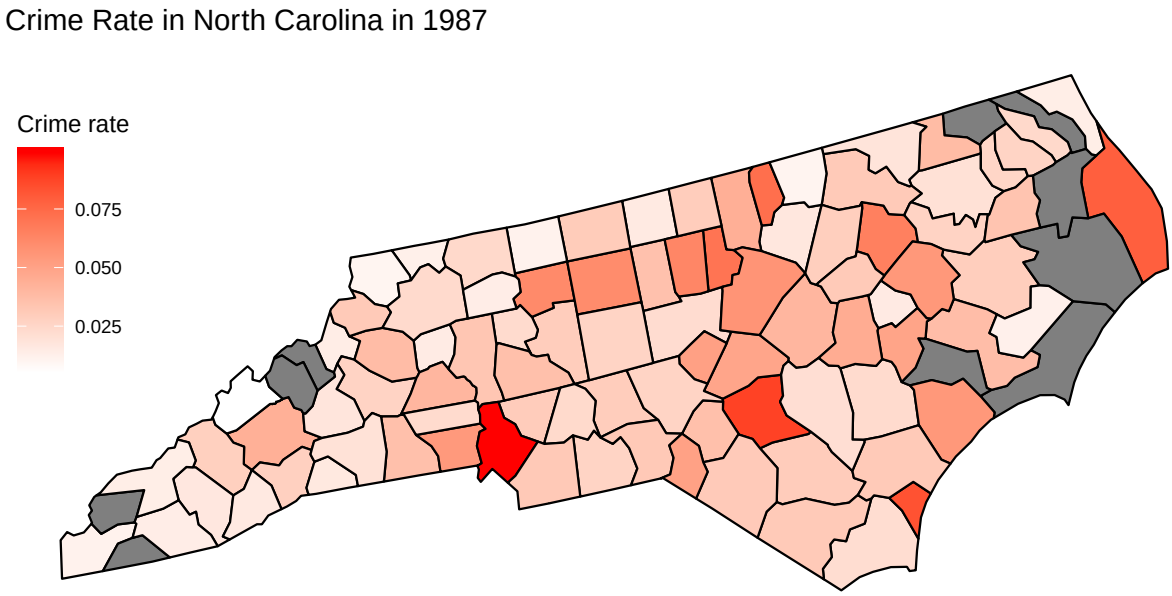
CLM 2: Random Sampling

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.

```
data$fips <- 37*1000 + data$county
plt_data <- select(data, fips, crmrte_abs)

plot_usmap('counties', include = 'NC', data = plt_data, values='crmrte_abs')+
  scale_fill_continuous(low='white', high='red') +
```

```r
  labs(title = "Crime Rate in North Carolina in 1987", fill="Crime rate")+
  theme(legend.position = c(0,0.4))
```



We see that the 10 counties without data (in black) are somewhat clustered along the eastern and western/north western boarders of North Carolina. This can certainly skew our analysis to that of central 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 of 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 drastic 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 that we had perfect multi-collinearity, so in this case we have fulfilled this requirement. We can this using the VIF for each coefficient to evaluate whether some degree of multicollinearity should be of worry. This is done as follows.

```r
vif(model_1)
```

```
##   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.

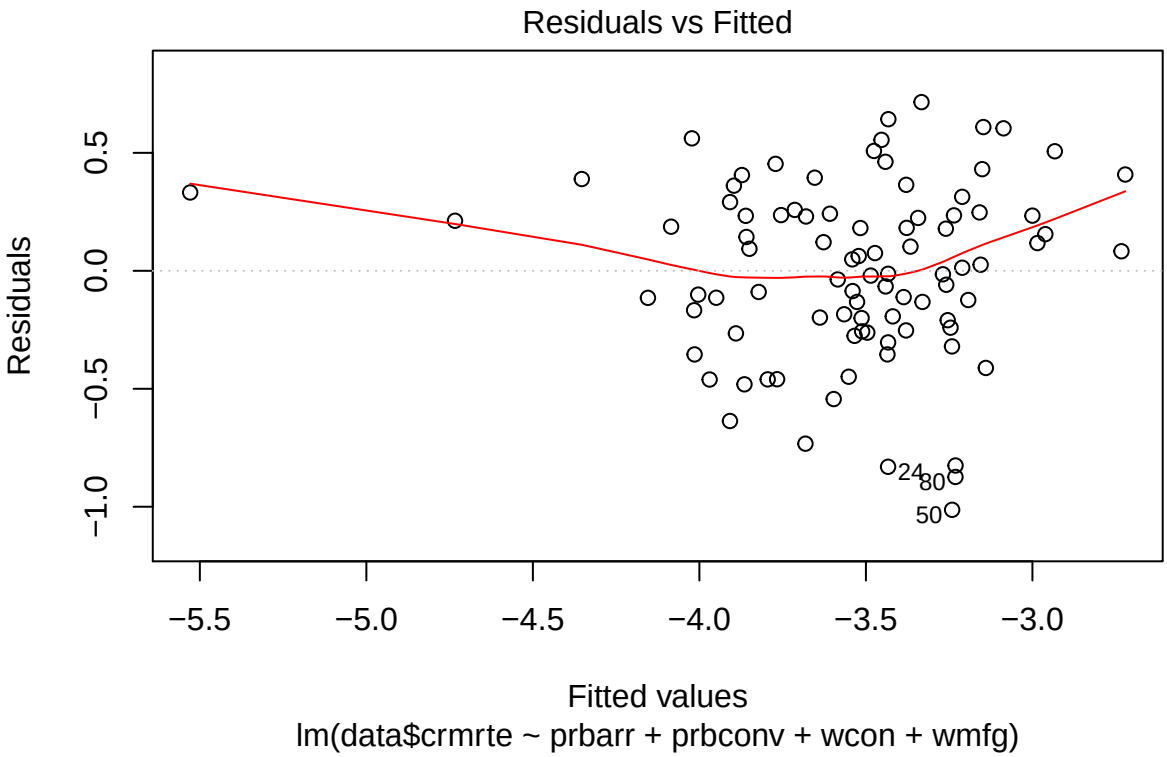
CLM 4: Zero Conditional Mean

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.

```
plot(model_1, which = 1)
```



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 centred 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

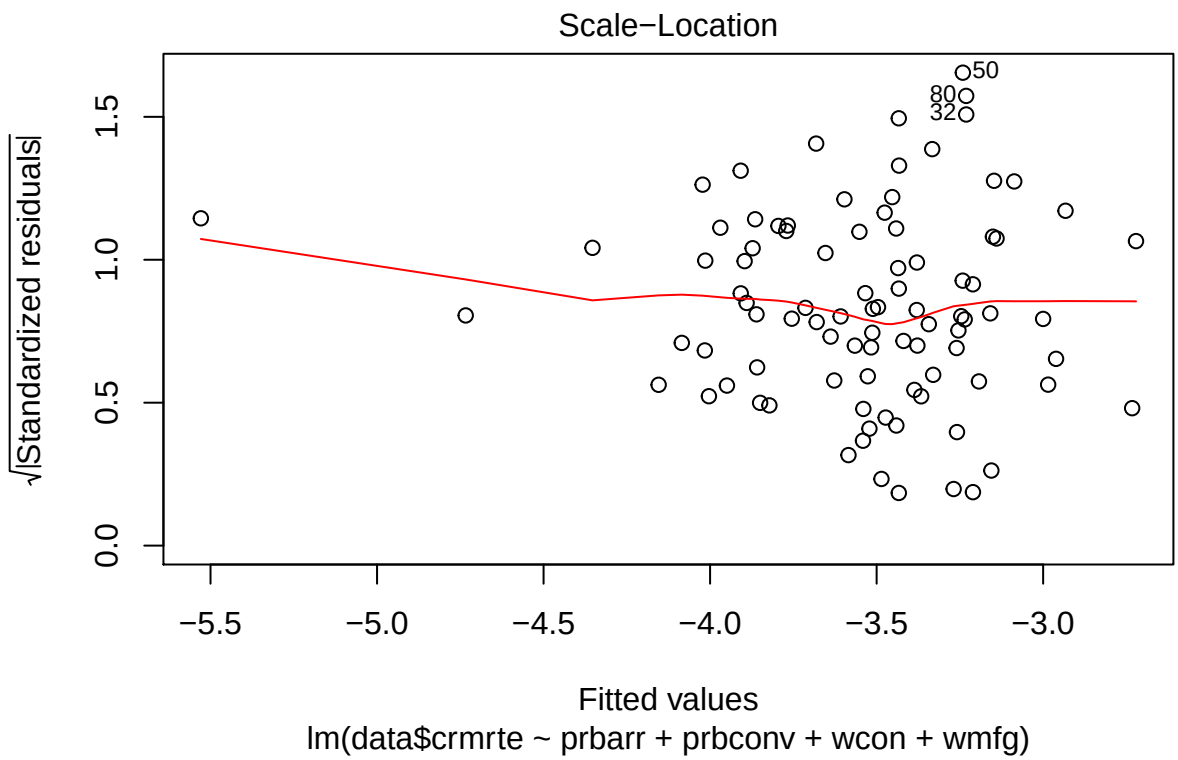
Homoskedasticity assumption is that the variance of the error terms are constant for any combination of x_k values.

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

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:

```
plot(model_1, which=3)
```



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.

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

```
bptest(model_1)
```

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

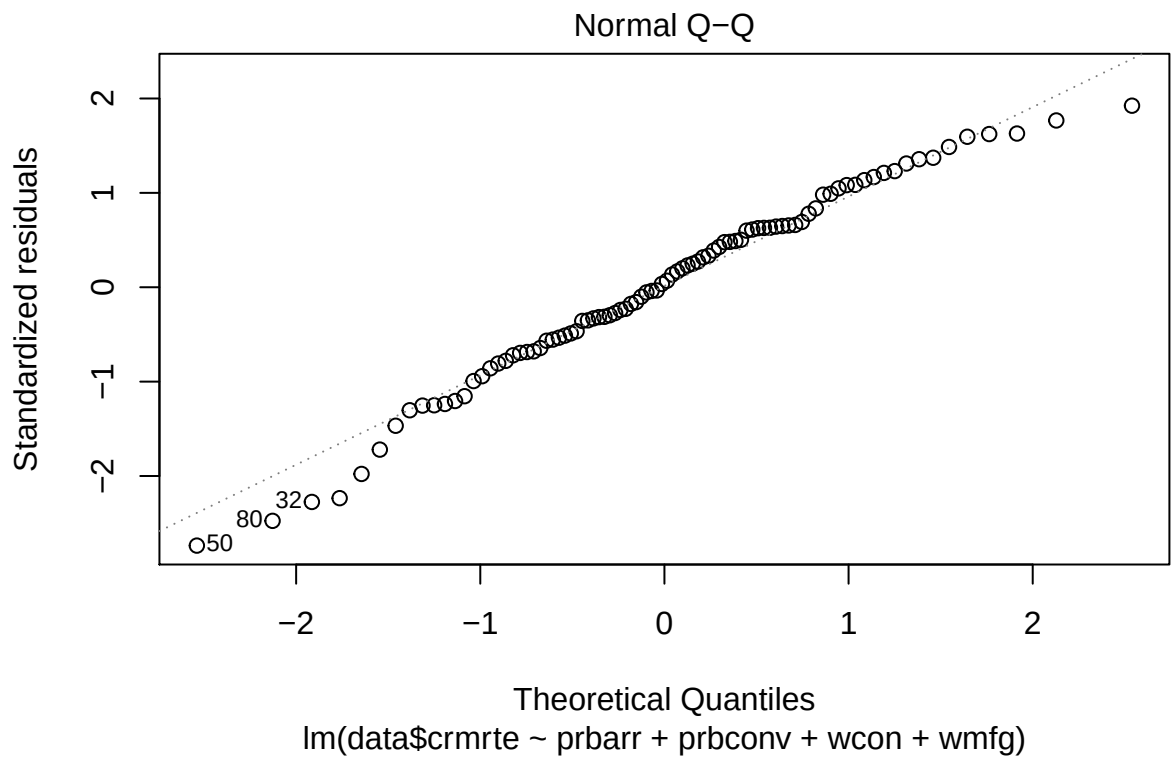
Since the $p - value >> 0.05$, we fail to reject the null hypothesis that we have homoskedasticity.

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.

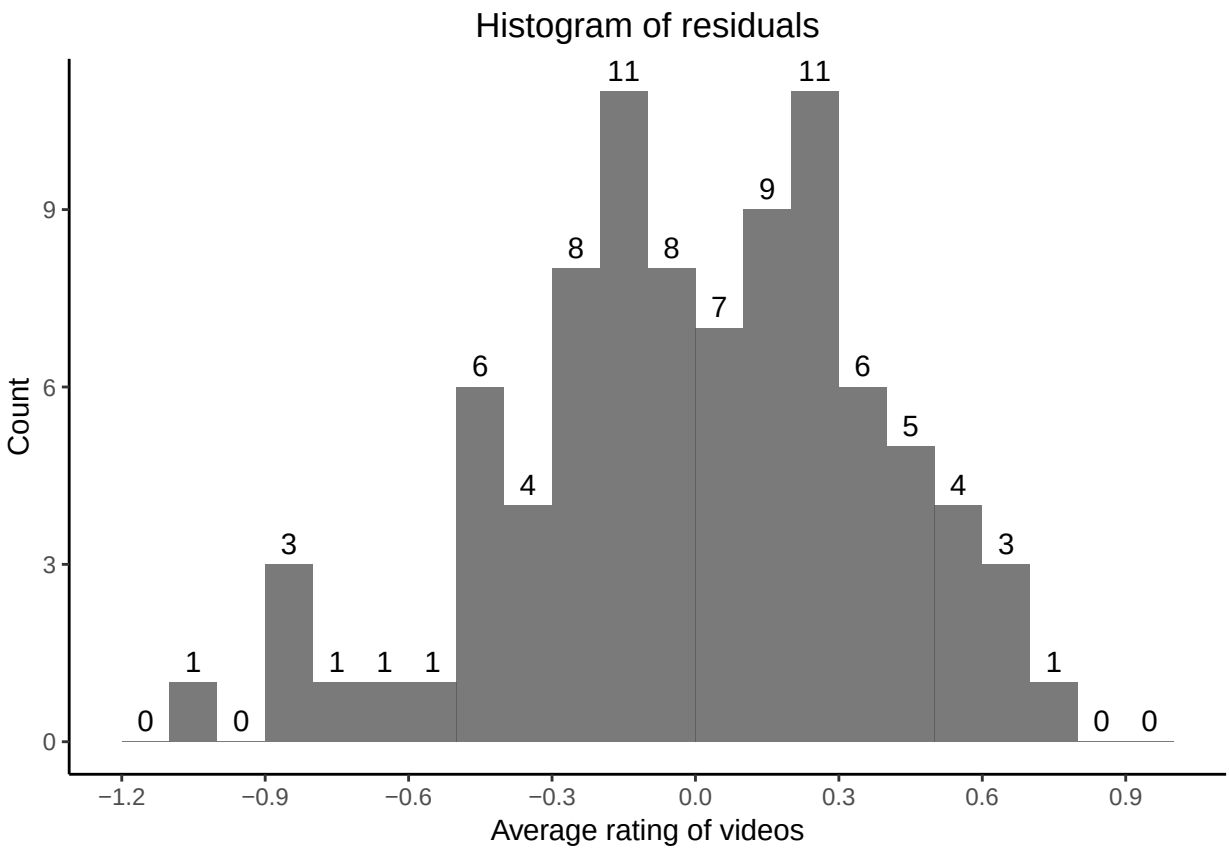
```
plot(model_1,which=2)
```



Not even counting the exception of 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.

```
bins <- seq(-1.2,1,0.1)
ggplot(data = as.data.frame(model_1$fitted.values), aes(x=model_1$residuals))+
  geom_histogram(alpha=0.8, breaks=bins)+
  labs(title='Histogram of residuals',
       x='Average rating of videos',
       y = "Count") +
  theme_classic() +
  #ylim(0,2200)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(-1.2, 1, 0.3)) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=bins
           )
```

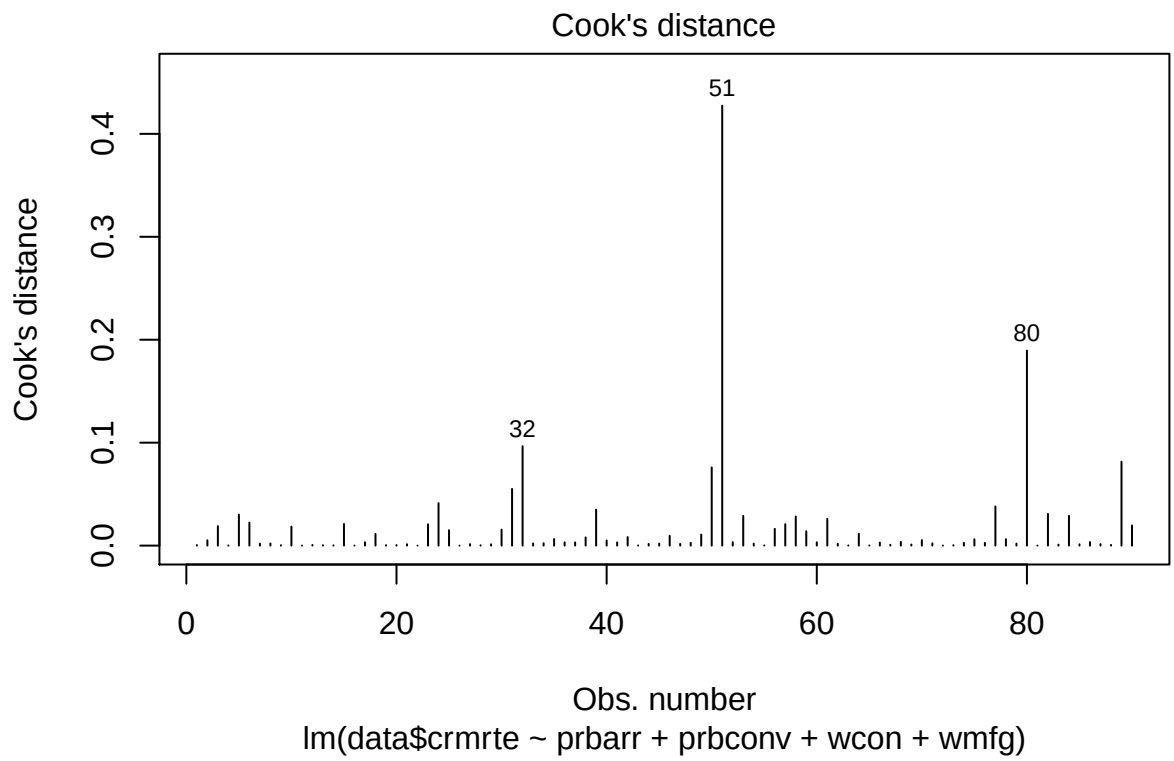


Based on the histogram, the data does not appear very normal. In fact, it is somewhat bimodal around -0.4, and 0.2.

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:

```
plot(model_1, which=4)
```



We see that data point 51 could be problematic. While its Cook's distance is stil below 0.5, which is typically considered to be large, it does deviate significantly from the average. We will examine this data point further in our future models.

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.

```
se.model_1<-sqrt(diag(vcovHC(model_1)))
se.model_1
```

```
## (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, so we are least certain about their estimates from the model.

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 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.

```r
model_1.tests<-coeftest(model_1, vcov = vcovHC, df=dim(X_wage_transformed)[1] - 4 - 1)
model_1.tests
```

```
##
## 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, while the wage of construction is not statistically significant. The wage of manufacturing is statistically significant however.

Model 2

Crime Rate Correlations

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)

```
y <- data$crmrte
X <- data[,!names(data) %in% c('county',
                               'year',
                               'crmrte',
                               'west',
                               'central',
                               'urban',
                               'crmrte_abs')]


#correlate all variables and store in new dataframe
cor_df <- data.frame(variable = character(),
                     crmrte_cor = numeric())
for (x in names(X)) {
  crmrte_cor <- cor(y, data[,x])
  corr <- as.data.frame(crmrte_cor,
                        col.names = c('crmrte_cor')) %>%
                        add_column(variable = x, .before = 1)
  cor_df <- rbind(cor_df, corr)
}

cor_df <- arrange(cor_df, desc(crmrte_cor))
print(cor_df)
```

```
##    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. As stated before, highly dense population areas present more opportunities for crime, and also have larger wealth gaps. In fact,

since we want to predict crime, density in some ways may be viewed as an output variable. Crime is in terms of per person, and a person ability to commit crime, even perhaps unknowingly, increases as the density of population increases, simply due to more opportunities. For example, imagine the thought experiment where we randomly sample some group of people from the entire population of North Carolina state. Then, we randomly assign each person to live in a rural area or a densely populated area. We believe that every time, the group assigned to the densely populated area will commit more crime on average, simply because each person has more opportunity to do this. As a result, this variable may absorb some of the "causal effects" of other variables, and we would like to exclude this from our regression models. Instead, we will save this variable for model 3 in order to check the robustness of our model 2. Furthermore, urban is a similar categorical variable that is directly related to density, and will serve the same purpose in model 3.

We will now take a closer look at the rest of the variables.

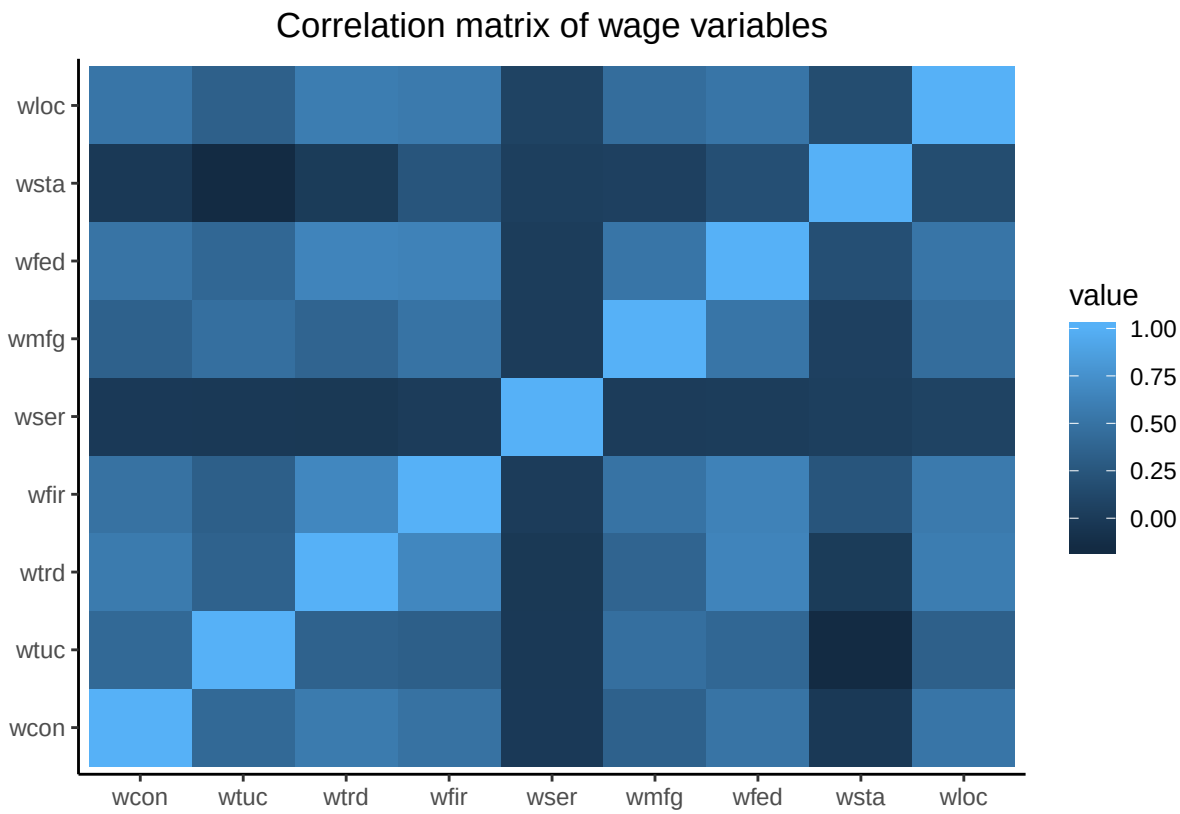
Wages Variables

```
nonwage_variables <- c('prbarr', 'prbconv', 'prbpris', 'avgsen',
                       'polpc', 'density', 'taxpc',
                       'pctymle', 'pctmin80', 'mix',
                       'urban', 'central', 'west')

wage_variables <- c('wtrd', 'wfir', 'wser', 'wfed', 'wsta', 'wloc',
                    'wcon', 'wtuc', 'wmfg')

X_non_wage <- data[, names(data) %in% nonwage_variables]
X_wage <- data[, names(data) %in% wage_variables]

heatmap.data <- melt(cor(X_wage))
ggplot(data = heatmap.data, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+
  labs(title='Correlation matrix of wage variables',
       x='',
       y = "") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
```

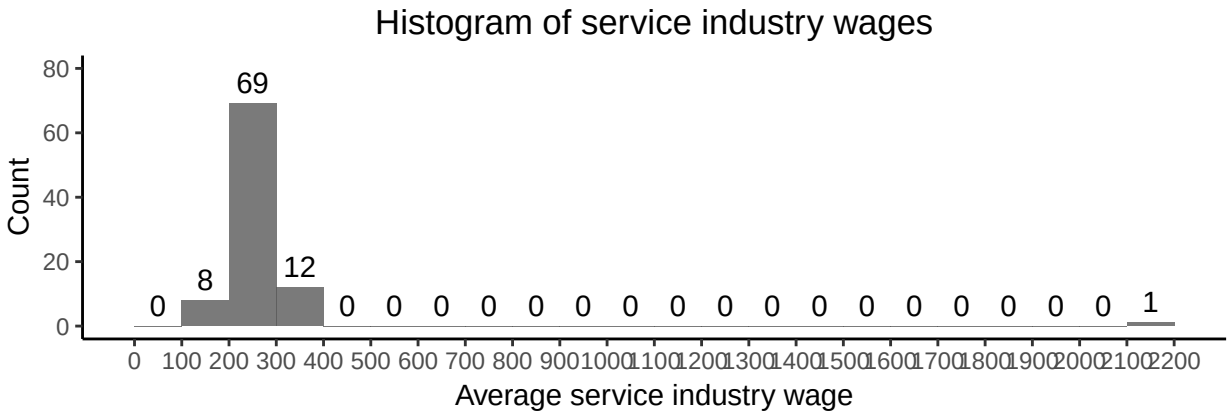


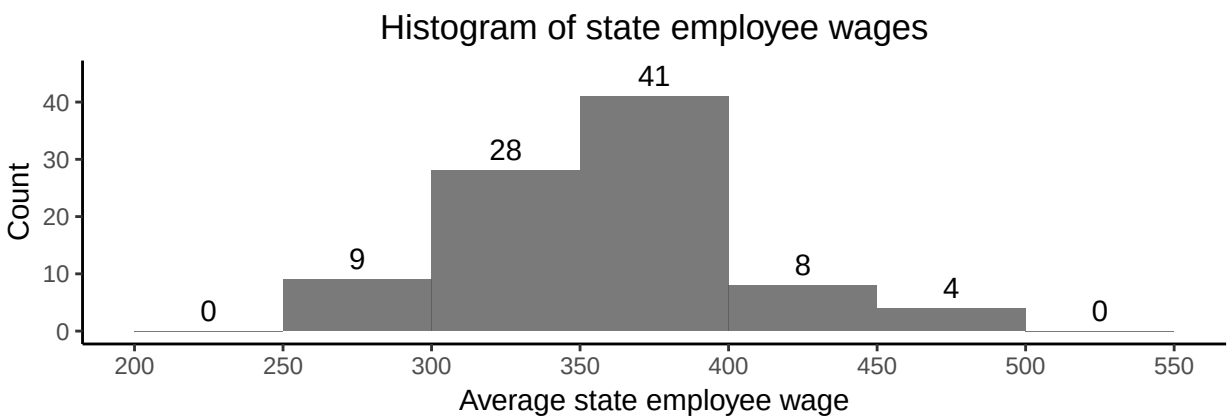
Interestingly, wser and wsta are both relatively dark compared to the rest of the data set. If the data was accurate, then that's a good indication of strong independent predictors within the wage category. wfed is the largest single univariate predictor from the correlation table. Both wfed and wsta are government (federal and state) jobs with the potential to influence social and political change.

We start by performing some EDA to ensure that these variables are reasonable:

```
hist.wser <- ggplot(data = data, aes(x=wser)) +
  geom_histogram(alpha = 0.8, breaks=seq(0,2200,100)) +
  labs(title = "Histogram of service industry wages",
       x = "Average service industry wage",
       y = "Count") +
  theme_classic() +
  ylim(0,80)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(0,2200,100)) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=seq(0,2200,100))
hist.wsta <- ggplot(data = data, aes(x=wsta)) +
  geom_histogram(alpha = 0.8, breaks=seq(200,550,50)) +
  labs(title = "Histogram of state employee wages",
       x = "Average state employee wage",
       y = "Count") +
  theme_classic() +
  ylim(0,45)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(200,550,50)) +
```

```
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=seq(200,550,50))
grid.arrange(hist.wser, hist.wsta,
             nrow=2, ncol=1)
```

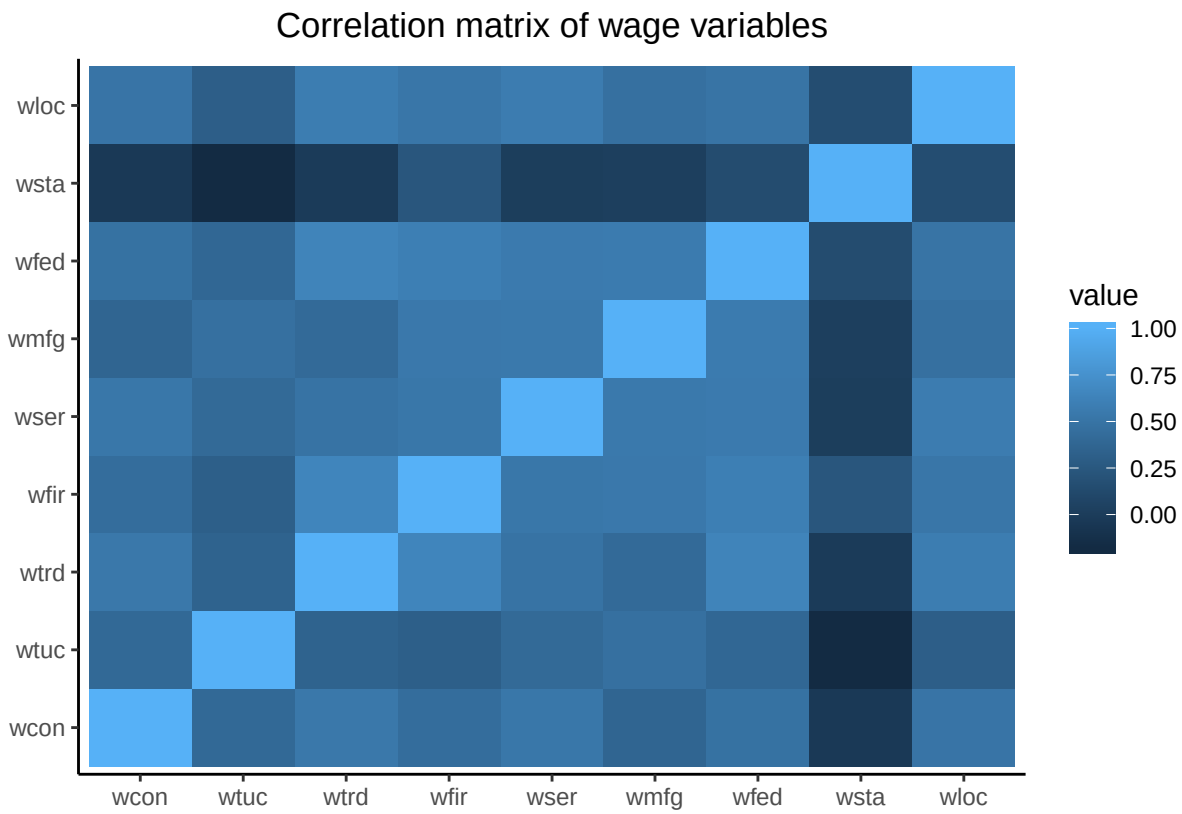




The state wage is somewhat right skewed, likely because high ranking officials make more than most average employees. However, there is not any red flags. However, we see that there's a huge outlier in service industry wages, more than 10 fold. Very likely this is an error in which the decimal was shifted by 1. The county is Warren County, which is not known to have such high service industry wages. Even if the data point is accurate, it may be significantly skewed for example due to non-random sampling of CEOs of service industry. In this case, this data point would not represent our target population, which is all employees in the service industry. We choose to fix this point by imputing the value to the average across the state.

```
data$wser <- ifelse(data$wser>1000, mean(data$wser), data$wser)
X_wage_transformed$wser <- log(data$wser)

X_wage <- X_wage_transformed[, names(X_wage_transformed) %in% wage_variables]
heatmap.data <- melt(cor(X_wage))
ggplot(data = heatmap.data, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+
  labs(title='Correlation matrix of wage variables',
       x='',
       y = "") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
```

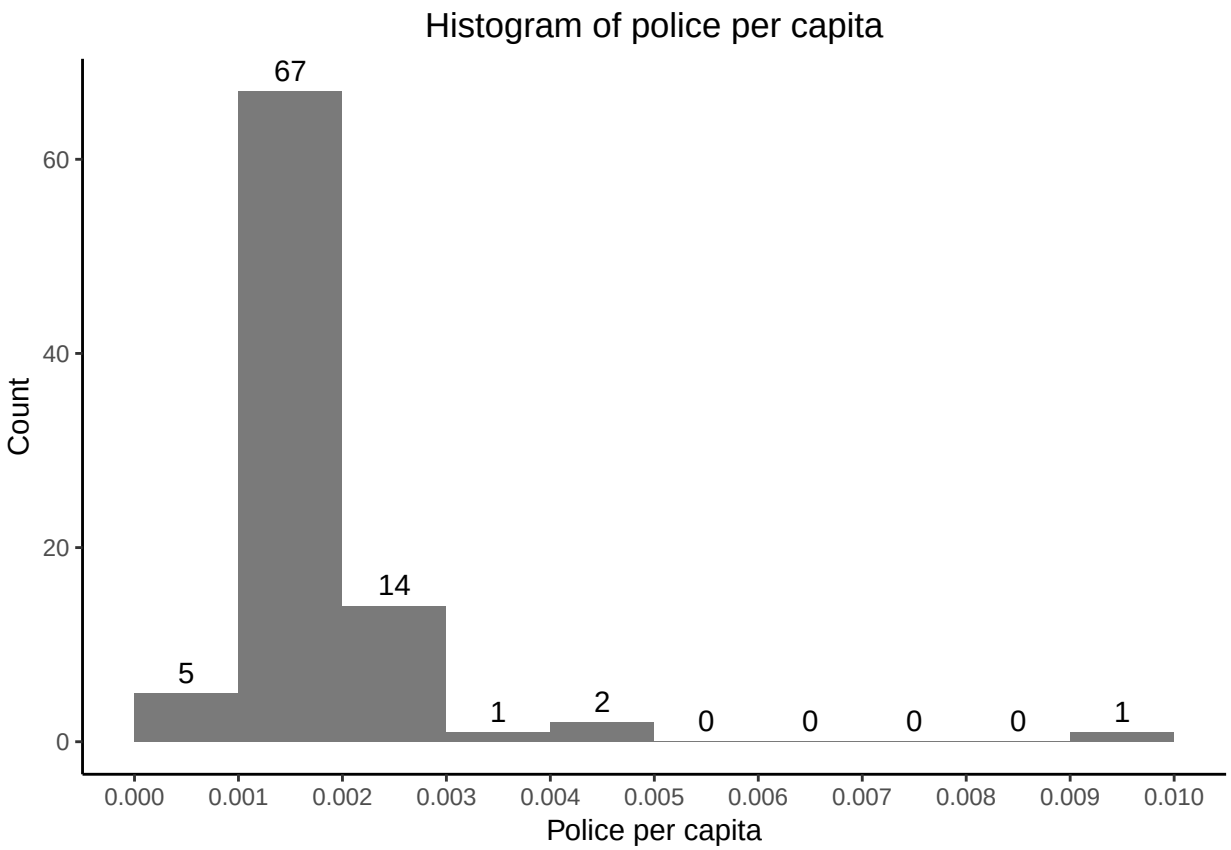


Unfortunately, with the imputation of the incorrect data point. However, wser still has weak correlation with all other wage variables and represents a good independent predictor in the wage category.

Police per Capita

It was previously stated that more police could result in stronger detection of crime. Therefore we look into the effect of the police per capita variable.

```
breaks = seq(0,0.01,0.001)
ggplot(data = data, aes(x=polpc)) +
  geom_histogram(alpha = 0.8, breaks=breaks) +
  labs(title = "Histogram of police per capita",
       x = "Police per capita",
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=breaks) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=breaks)
```



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

```
report <- rbind(data[51, c('crmrte_abs', 'prbarr', 'prbconv', 'polpc')],
                as.data.frame(lapply(select(data, crmrte_abs, prbarr, prbconv, polpc), mean)))
rownames(report) <- c('outlier county', 'state average')
stargazer(report,type='text', summary = FALSE)
```

```
##
## ==============================================
##                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 and 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 more police means that more crime gets responded to in a timely fashion. Interestingly, the rate of crime is about 5-6x less than state average. This is actually consistent with the analysis above. The capacity of the police force to respond to crime likely exceeds the rate of crime. As a result, people are less likely to commit crime because they know they will be caught, and at the same time any crime that does get committed is likely dealt with leading to arrests and convictions. This is to say that we actually believe this is a legitimate data point, and as a result will keep it in our regression analysis.

Demographic and Geographical Variables

We believe percent minority will be a good independent predictor of crime. Unfortunately, the truth is that minorities are more likely to targets for the police, especially in rural regions. Their behavior tends to be

scrutinized more, and sometimes even activities considered non-criminal could be considered criminal for minorities. In addition, due to their socialeconomic position, minorities statistically are poorer, which would be an exogenous variable correlated with crime. In our case, young male won't be considered a priority because the nature of the crime is not specified. Young male can't commit some of the white collar crimes for example. If the data was only for example, petty theft, perhaps young male would be a good predictor.

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. We did not believe these were key determinants for model 1 because for many types of crime prison would be considered too harsh of a punishment over say a fine or community service. However, we do know what crime is exactly in this particular report. We would like to evaluate these variables as potential important co-variates.

Finally, while we see that geography is important in some cases, we are uncertain whether it will be a strong predictor here.

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. The discussion presented above seems to indicate that it would be beneficial to include the variables wsta, wser, police and minorities.

In order to confirm this, and to further narrow down the variable selection further, we begin with a global AIC optimization using a combination of forward and backward selection, and then apply knowledge from the EDA above to further adjust our model. In order to perform automatic feature selection using the AIC criteria, we will use the MASS package.

```
names_not_include <- c('density', 'urban')

y <- data$crmrte
X_stepwise <- X_wage_transformed[, !(names(X_wage_transformed) %in% names_not_include)]

model_upper <- lm(y ~ ., data=X_stepwise)
model_lower <- lm(y ~ 1, data=X_stepwise)

AIC.mixed <- stepAIC(model_upper,
                     trace = FALSE,
                     direction = 'both',
                     scope = list(upper=model_upper,lower=model_lower))
AIC.mixed
```

```
##
## 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
```

```
model_2 <- AIC.mixed
```

First, we see that the insignificant variables in our first model did not show up in this AIC optimized model, which is a good sign as we likely would not want to specifically include wages of blue collar jobs in our second

model. The AIC mixed model also suggests that a lot of the same variables are from our EDA above, with addition of two variables we did not expect to include, which are percent young male, and taxpc, or tax revenue per capita. The model did not include the probability of conviction and length of prison sentence, which we thought would be important. We will evaluate these with separate specifications.

First, we test the significance of the coefficients generated by the AIC minimized model. For all the betas, we will use a 2-sided test as significance level 0.05:

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

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.

```
coeftest(AIC.mixed, vcov. = vcovHC, df=dim(X_wage_transformed)[1] - 8 - 1)
```

```
##
## 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
```

Everything is significant (p<0.05) with the exception of taxpc and pctymle. As a result, for all variable coefficient except taxpc and pctymle, we reject the null hypothesis and state that the coefficient is in fact not zero. We fail to do so for taxpc and pctymle. This seems to suggest that our intuition about these variables are correct. However, the test presented above are significance values for each individual coefficient. It could be that these variables are jointly significant in the following fashion: imagine that regions with larger young male populations simply represents regions with more families with children. These family most likely would vote in support of better schools and safety for their families (higher taxes), and as a result, would have a positive collinear relationship with the tax revenue per capita. Alternatively, since the tax variable is measured in terms of per capita, it could also just be that regions with higher percentage of young male generates tax revenue simply because young people generate less income. In both cases, these variables have the potential to be jointly significant. We check this with an F-test specified below at alpha=0.05, two-tailed test:

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

```
linearHypothesis(AIC.mixed, c("taxpc = 0", "pctymle = 0"), vcov = vcovHC)
```

```
## 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 and perform a statistical test, analogous to that above:

```
prison.model <- lm(formula = data$crmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed + wsta + prbpris +
coeftest(prison.model, vcov. = vcovHC, df=dim(X_wage_transformed)[1] - 8 - 1)
```

```
##
## 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 joint significance:

```
linearHypothesis(prison.model, c("prbpris = 0", "avgsen = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbpris = 0
## avgsen = 0
##
## Model 1: restricted model
## Model 2: data$crmrte ~ 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.

We saw from the EDA that many of the wage variables, with the exception of wsta, were correlated with each other. Each though each of the individuals variables did not make into our AIC model, could a set of them be jointly significant? We can test this by generating a model specfication with all wage variables, and testing the joint significance of wage terms not included in our AIC model:

```
model_wage <- lm(data$crmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed + wsta + wcon + wtuc + wtrd +

coeftest(model_wage, vcov. = vcovHC, df=dim(X_wage_transformed)[1] - 13 - 1)
```

```
##
## 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
```

First, checking the significance of each coefficient, we see that none of the wage variables are statistically significant, and including all of the wage variables absorbed the significance from wsta as well. We now check joint signficance:

```
linearHypothesis(model_wage, c("wcon = 0", "wtuc = 0", "wtrd= 0", "wfir= 0" , "wser= 0" , "wmfg= 0" , "
```

```
## 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$crmrte ~ 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
```

To our surprise, all wage variables (except wfed and wsta) were also not joint significantly. Based on experimenting with these alternative specifications, our AIC model, and statistical testing, we now arrive at our model 2.

```
model_2 <- lm(data$crmrte ~ prbarr + prbconv + polpc + pctmin80 + wfed + wsta, data = X_wage_transformed
```

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

$$H_0 \text{ :second order polynomial not needed}$$
$$H_a \text{ :second order polynomial is needed}$$

```
resettest(model_2, power=2, type='regressor')
```

```
##
##  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.

Classical Linear Model Assumptions

CLM 1 and 2 are identical to our original model.

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. We can this using the VIF for each coefficient to evaluate whether some degree of multicollinearity should be of worry. This is done as follows.

```
vif(model_2)
```

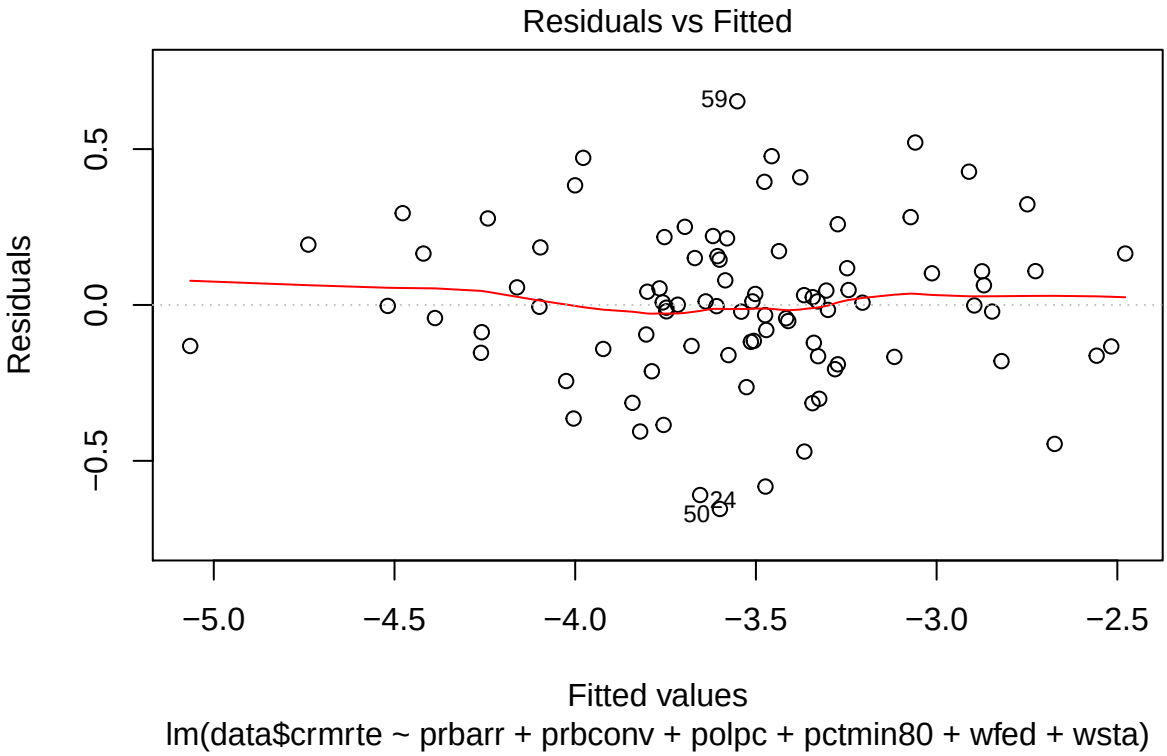
```
##   prbarr  prbconv    polpc pctmin80     wfed     wsta
## 1.481696 1.120632 1.527763 1.103844 1.152948 1.095947
```

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

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.

```
plot(model_2, which = 1)
```

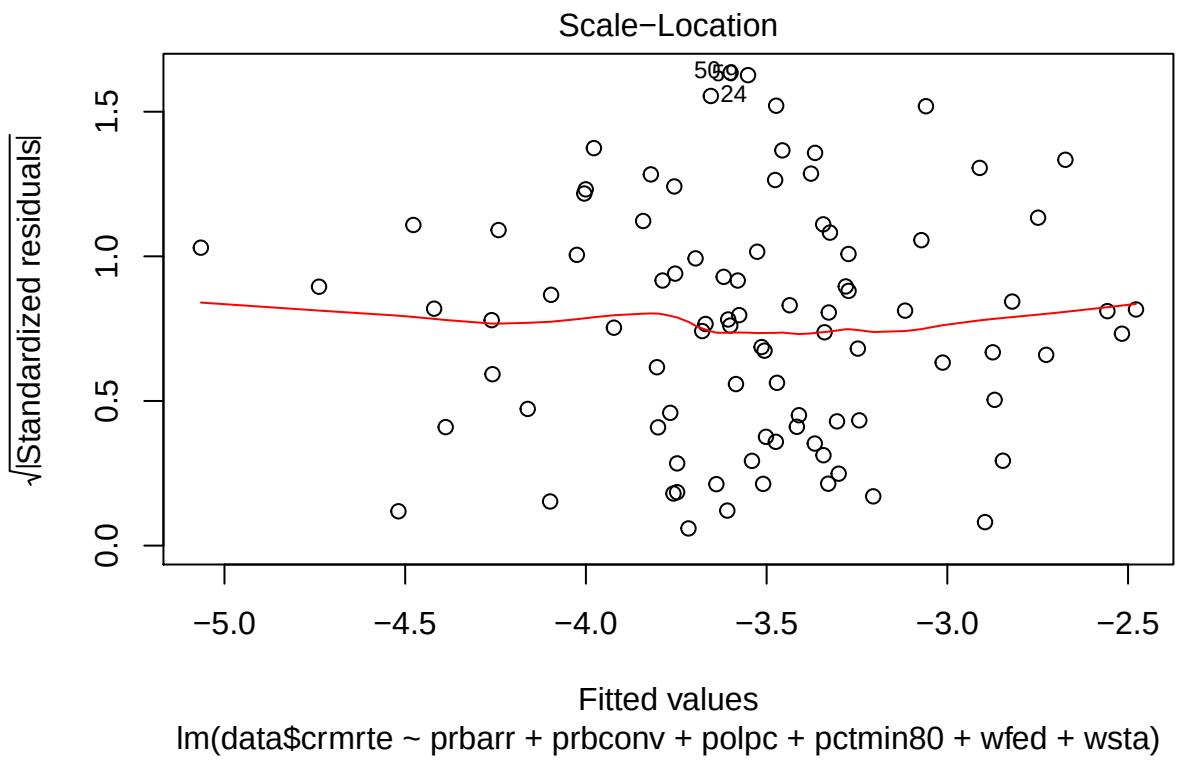


Based on this plot, we see that the line is very much linear even at extreme values. The average value of the residual is approximately 0 for all fitted vales. Zero Conditional Mean is met for our model of 6 variables.

CLM 5: Homoskedasticity

Examining the fitted values versus residuals plot above, we see that the spread appears to be 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. If homoskedasticity were achieved, we would expect a horizontal line across this plot:

```
plot(model_2, which=3)
```



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.

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

```
bptest(model_2)
```

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

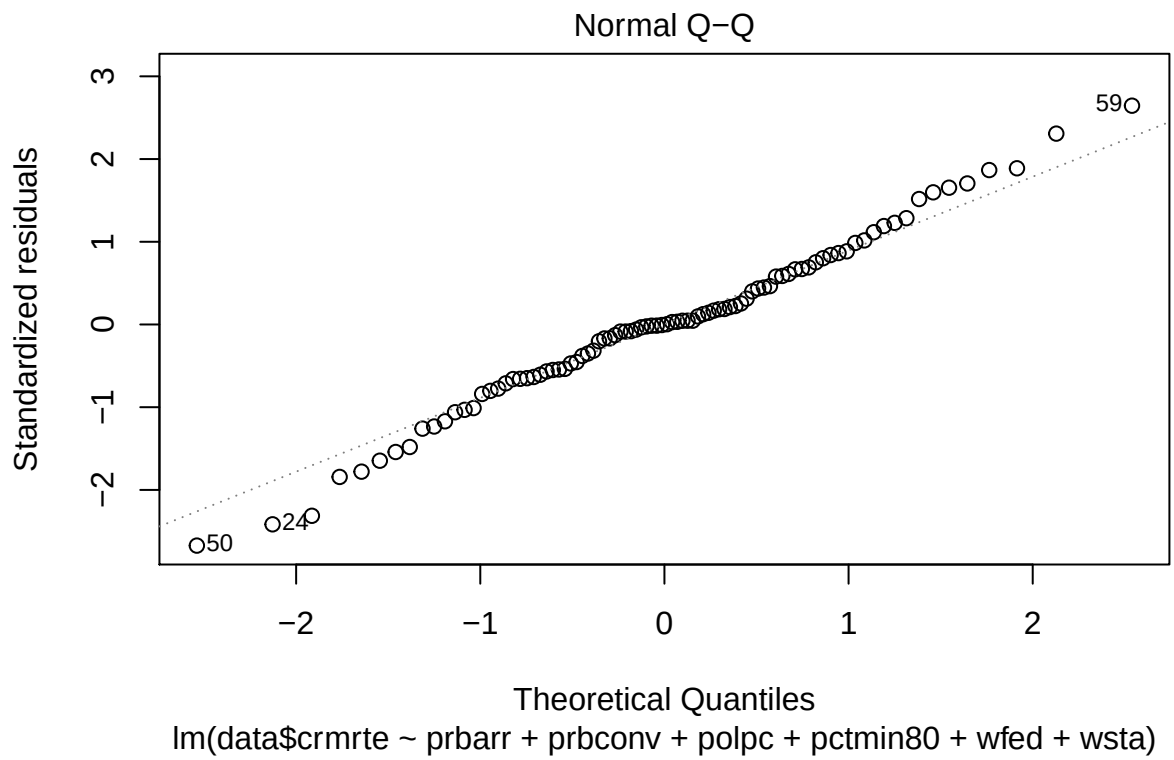
Since the $p - value < 0.05$, we reject the null hypothesis that we have homoskedasticity. Our sample size is relatively small, so the test suggests we have a major deviation from homoskedasticity.

As a result, we will use heteroskedastic robust errors (as we have done by default) 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.

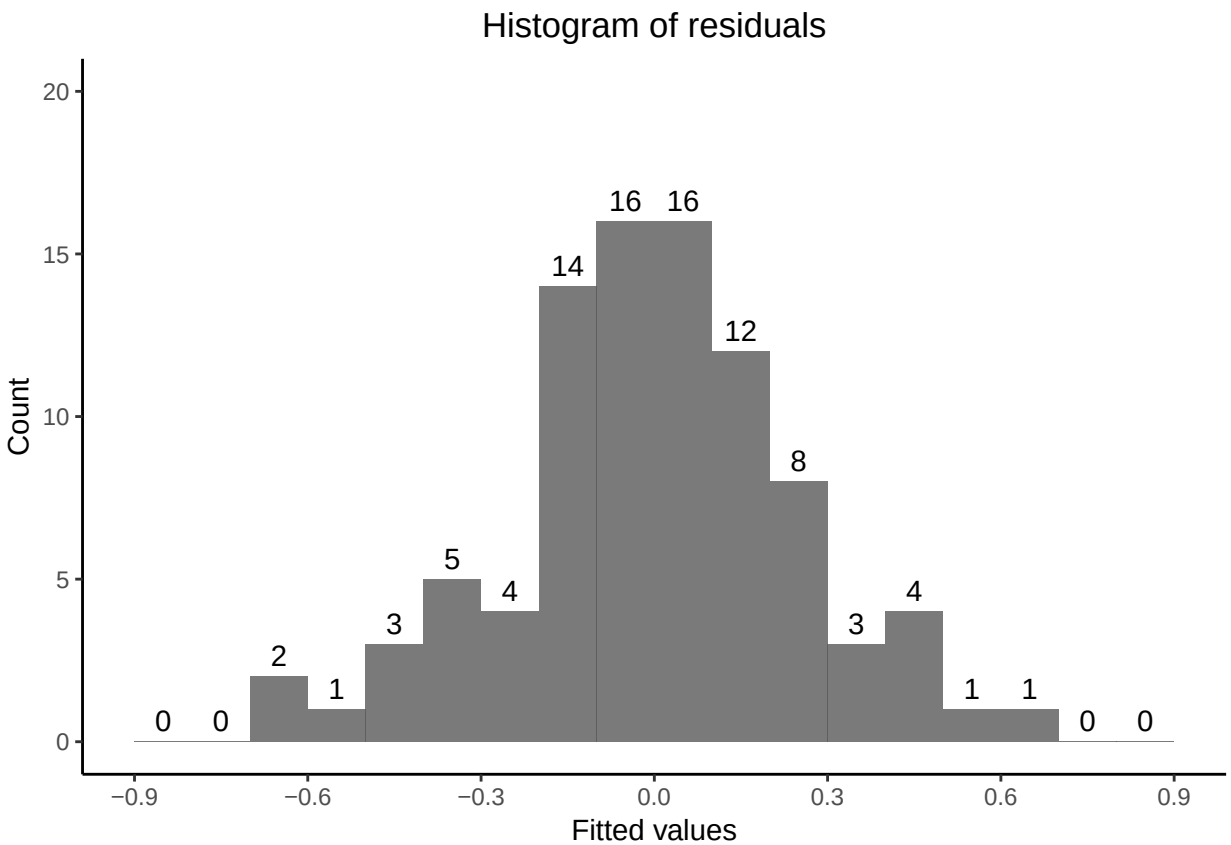
```
plot(model_2,which=2)
```



Not even counting the exception of 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.

```
bins <- seq(-0.9,0.9,0.1)
ggplot(data = as.data.frame(model_2$fitted.values), aes(x=model_2$residuals))+
  geom_histogram(alpha=0.8, breaks=bins)+
  labs(title='Histogram of residuals',
       x='Fitted values',
       y = "Count") +
  theme_classic() +
  ylim(0,20)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(-1.2, 1, 0.3)) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=bins
           )
```



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.

Tests important for recommendations below

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 can either generate an alternative specification by defining a new variable as the difference of the two we are interested in, or use the linearHypothesis test from the car package. We will perform the ladder.

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

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

```
linearHypothesis(model_2, 'prbarr = prbconv', vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr - prbconv = 0
##
## Model 1: restricted model
## Model 2: data$crmrte ~ 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. Namely,

```
model_2$coefficients
```

```
##  (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, probability of arrest is a better deterrent than probability of conviction.

Model 3

After adding relevant covariant variables to our base model in the previous section, we now evaluate the robustness of the model by comparing its performance when adding further covariates back into the model.

Specifically, we left out the density variable specifically for this matter. From the correlation matrix, we see that all variables are individually somewhat correlated with crime. For model 3, we will include all of these variables.

```
model_3 <- lm(data$crmrte ~ ., data = X_wage_transformed)
summary(model_3)$adj.r.squared
```

```
## [1] 0.8157934
```

```
AIC(model_3)
```

```
## [1] 16.57352
```

```
summary(model_2)$adj.r.squared
```

```
## [1] 0.7769309
```

```
AIC(model_2)
```

```
## [1] 21.07499
```

```
model_2.coef <- model_2$coefficients
model_3.coef2 <- model_3$coefficients[names(model_3$coefficients)  %in% names(model_2$coefficients)]
perc.change <- (model_3.coef2 - model_2.coef)/model_2.coef * 100
report <- cbind(model_2.coef, model_3.coef2, perc.change)
colnames(report) <- c('model_2 coefficients', 'model_all_coefficients', 'percent_change')
stargazer(report, summary=FALSE, header=FALSE, type='text')
```

```
##
## ======================================================================
##             model_2 coefficients model_all_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
## ----------------------------------------------------------------------
```

Adding back variables such as density does not significantly alter our R^2. Our model_2 is already very robust. 23 predictors increase the adjusted r squared by very little compared to using just 6. The AIC is slightly better for model 3 compared to model 2, but not to the AIC.mixed model, which is to be expected since AIC.mixed was optimized for AIC. Taking out percentage young male and tax revenue reduced AIC from AIC.mixed to model 2. Our coefficients however have changed between 20%-30%. However, none of the signs on the coefficients have changed, meaning they correctly predict the direction of change.

We now evaluate the CLM assumptions.

Classical Linear Model Assumptions

CLM 1 and 2 are identical to our original model.

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. We can this using the VIF for each coefficient to evaluate whether some degree of multicollinearity should be of worry. This is done as follows.

```
vif(model_3)
```

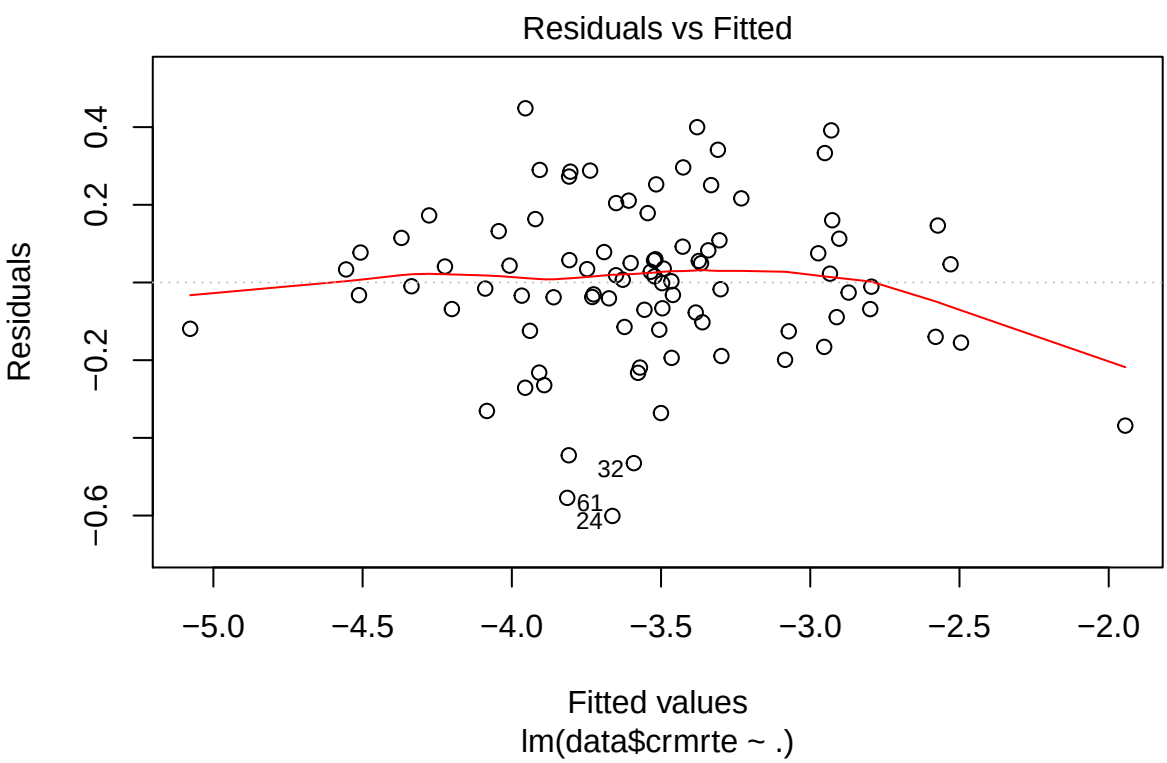
```
##   prbarr  prbconv  prbpris   avgsen    polpc  density    taxpc     west
## 2.364452 1.761170 1.218374 1.825254 2.990818 5.443675 1.972683 3.271060
##  central    urban pctmin80      mix  pctymle     wcon     wtuc     wtrd
## 2.080564 3.957133 2.912945 1.922204 1.503969 2.211167 1.768986 3.065321
##     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 below 4. We tagged these variables in the very beginning as strong predictors, and possibly being viewed as even an output variables, so this is to be expected. Since we do 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

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.

```
plot(model_3, which = 1)
```



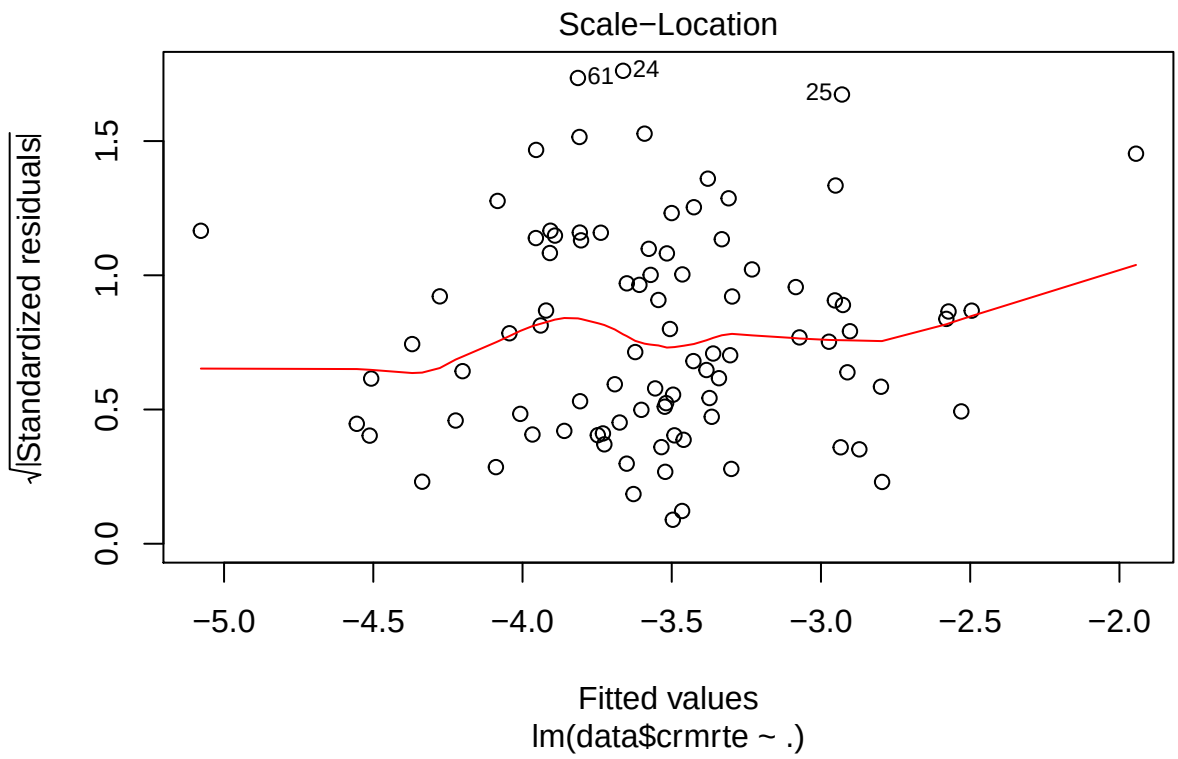
Based on this plot, we see that the line is very much linear except at the one extreme value of -2 fitted values. This is due to a single data point, which is enough to say that the entire model does not satisfy zero conditional mean. At all other fitted values, the model looks decent. As a result, Zero 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 (around -0.6 to 0.4) than around 4 (around -0.3 to 0.4). We can also

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

```
plot(model_3, which=3)
```



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

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

```
bptest(model_3)
```

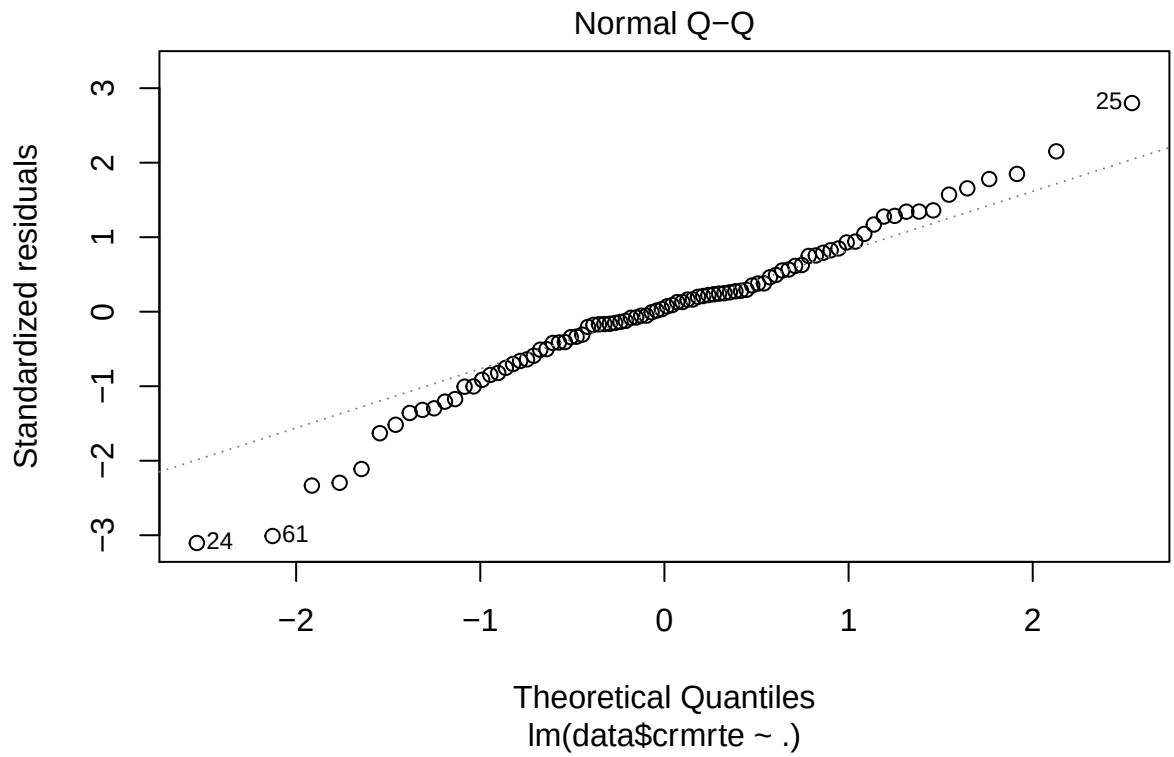
```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 30.19, df = 22, p-value = 0.1139
```

Since the $p - value > 0.05$, we fail the null hypothesis that we have homoskedasticity. Our sample size is relatively small, so this test 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.

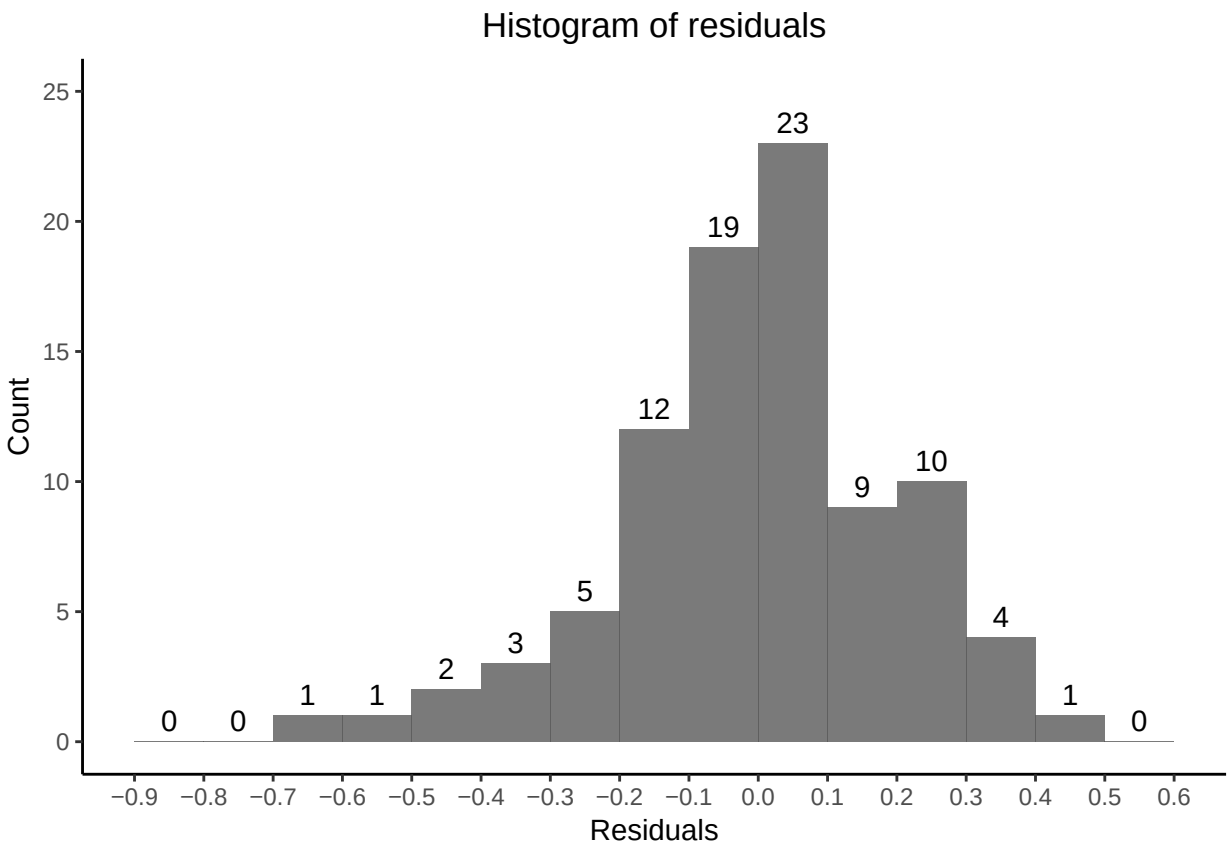
```r
plot(model_3,which=2)
```



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.

```r
bins <- seq(-0.9,0.6,0.1)
ggplot(data = as.data.frame(model_3$residuals), aes(x=model_3$residuals))+
  geom_histogram(alpha=0.8, breaks=bins)+
  labs(title='Histogram of residuals',
       x='Residuals',
       y = "Count") +
  theme_classic() +
  ylim(0,25)+
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=bins) +
  stat_bin(aes(y=..count.., label=(..count..)),
           geom="text",
           vjust=-.5,
           breaks=bins
           )
```



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.

Finally, we perform hypothesis testing to see which variables are significant when all co-variates are included in the model:

```
coeftest(model_3, vcov. = vcovHC, df=dim(X_wage_transformed)[1] - 22 - 1)
```

```
##
## 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.

We explore percent young male further in terms of omitted variable bias. Since model 2 did not pick up young male as statistically significant, this is likely due to omitted variable bias of a variable or set of variables that is in our dataset. We think the most influential variables are density and urban. The primary reason is that the likelihood of young males committing a crime is much higher in regions of larger population density. More negative influences are present for young males in cities, and more opportunities present themselves in cities for the young male population. Since density is such a strong predictor of crime rate, the mean squared residuals (and error estimate of our residuals) should be decreased, which can make male significant due to overall better fit of the data. We can check this with an alternative specification:

```r
model.no.density.urban = lm(formula = data$crmrte ~  prbarr + prbconv + prbpris +
    avgsen + polpc + taxpc + west + central + pctmin80 + pctymle + mix +
    wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
    wloc + wsta, data = X_wage_transformed)
coeftest(model.no.density.urban, vcov. = vcovHC, df=dim(X_wage_transformed)[1] - 20 - 1)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -11.9631453   4.1333202 -2.8943  0.005083 **
## prbarr       -2.1655875   0.3690083 -5.8687 1.385e-07 ***
## prbconv      -0.7384538   0.1233473 -5.9868 8.606e-08 ***
## prbpris      -0.1614770   0.3954814 -0.4083  0.684314
## avgsen       -0.0143041   0.0147468 -0.9700  0.335442
## polpc       174.6579170  77.2523605  2.2609  0.026924 *
## taxpc         0.0050281   0.0041549  1.2102  0.230345
## west         -0.0945866   0.1210298 -0.7815  0.437173
## central      -0.0749321   0.0921962 -0.8127  0.419159
## pctmin80      0.0093780   0.0029348  3.1955  0.002106 **
## pctymle       3.9063360   2.0241575  1.9299  0.057736 .
## mix          -0.0105748   0.5992990 -0.0176  0.985973
## wcon          0.1417335   0.2839191  0.4992  0.619224
## wtuc          0.1919980   0.2869963  0.6690  0.505733
## wtrd          0.3344999   0.3474585  0.9627  0.339058
## wfir         -0.1444032   0.3622615 -0.3986  0.691407
## wser         -0.3845419   0.2719604 -1.4140  0.161869
## wmfg          0.0273960   0.1780251  0.1539  0.878147
## wfed          1.2674176   0.5054542  2.5075  0.014517 *
## wloc          0.3567677   0.6990739  0.5103  0.611440
## wsta         -0.3461718   0.3033311 -1.1412  0.257718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sd(model_3$residuals)
```

```
## [1] 0.2043454
```

```r
sd(model.no.density.urban$residuals)
```

```
## [1] 0.2216842
```

We see that this is indeed the case. The standard deviations of the residuals are lower with inclusion of density and urban, and pctymle is not longer statistically significant at a level 0.05 when these variables are omitted.

Additional Omitted Variable Bias

In our regression model, we found that blue collar wages in construction and manufacturing were not significant in predicting crime rate. The average level of education someone in a county has is an omitted variable that has an effect on blue collar wage. For example, 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 so are paid higher. Let's take only construction wage and tmagine two model specifications below letting edlvl be the average education level:

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

We would think that beta_2 is negative – the more educated a county the less crime overall. One reason for this might be that individuals have more exposure to ethics as education level increases. 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. If we believe that alpha1 is positive as described above, then the product is going to be negative. This means that the observed beta_1 is the true beta_1 minus the absolute value of the omitted variable bias – i.e. observed is lower than actual. The effect of wcon might be more 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 variable affects crime rate itself, which is the fraction of crime commited compared to what is detected and recorded. If this fraction is lower for some counties than other, our data would suggest lower crime rates for those lower counties when reality it is a problem in the county's crime detection system.

For the wage variables, we believe that wage itself is influenced by the overall cost of living for each county. We would expect a positive correlation between wage and cost of living simply because people need to make more to survive in areas that cost 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. Since this is the case, this omitted variable of cost of living would bias the coefficient of our wage variables, artificially inflating them in our model (omitted variable bias is positive, so apparently coefficient is actually larger than the true coefficients). Since wage of state employees is a strong predictor of crime, this effect may be diminished if all state employees work in regions of high cost of living.

Other omitted variables include type of crime – white collar versus blue collar, substance abuse, and religious beliefs.

Policy Recommendations and Concluding Remarks

To conclude, consider once more our research questions:

What are the best independent predictors of crime rate?

As shown in the model development in previous sections, we proposed that the best independent predictors for crime rate are probability of arrest, probability of conviction, police per capita, the percentage of minority groups in the population, and the wages of federal and state employees.

How does wage influence crime rate?

At the beginning of the exercise we thought that higher wages overall would decrease the overall rate of crime because it would imply that the overall population would not have to commit crime to meet their needs. However, it makes sense that the best predictors are wages of people in the public sector, as they have a direct impact in the policy and law enforcement that prevents and fights crime.

Does fear of getting in trouble with the police deter crime?

According to the developed model, a stronger response (police per capita) and consequences to crime (chances of being arrested or convicted) diminish the possibility of crime from happening. This could be either because of and increase fear from potential criminals or due to a more efficient response from the authorities.

From these results, we believe the best possible policy recommendations to give to a North Carolina campaigner focusing on crime would be:

1) Revision of state law to reinforce punishment to certain crimes. Specially, the perception of the likelihood of arrest and conviction are both strong deterrents of crime. Increasing the rate of arrest and conviction can potentially lower crime in a particular area, due to increased fear that individuals will get caught. Since we believe that these variables are significant due to fear of being caught, greater publicity of arrest and conviction may serve the same purpose. The campaign should utilize these findings in order to advocate lower police leniency toward detected crime, and higher public reporting of arrests and convictions.

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 all crime in a timely manner. We recommend increase the police workforce in order to detect more of the crime being committed, and if funds permit, increase police to levels where criminals are afraid they will detected due to rampant police presence.

3) 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. Alternatively, potential reinforcement of police forces in counties with high percentage of minority groups in the population could decrease crime.

4) We see that state salaries is a positive predictor of crime. We ARE NOT SURE WHY. PLEASE HELP on this one. Increment of salaries, benefits, training and overall support of state and public servers in North Carolina.

To dos

Interpret the coefficients more for each model Discuss further omitted variables Make sure policy is consistent with regression analysis Wrap up the report so that introduction and concluding remarks play together with regression analysis.