Lab 3 Submission

Siddhartha Jakkamreddy, Neha Kumar, Brian Musisi

12/09/2018

Introduction

The motivation for this analysis is to determine the factors that lead to crime rate in North Carolina counties in 1980. We are assuming the role of data scientists for a political campaign around the same era within North Carolina to determine methods that can be employed to reduce the crime rate. Note, that this requires our analysis to look for causal variables so we can provide concrete and actionable resolutions.

Initial EDA

```
crime = read.csv("crime_v2.csv", header = TRUE)
summary(crime)
```

```
##      county           year         crmrte              prbarr
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6
##         prbconv      prbpris           avgsen           polpc
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6
##     density            taxpc             west           central
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6
##      urban            pctmin80          wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd             wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
```

```
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed            wsta            wloc            mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

Here, we see something interesting. prbconv is a factor due to a rogue ‘ character being added to the bottom of the file. When we view the dataframe, we actually see that this rogue tick mark has also introduced 6 null values in the file. We take steps to remove these records from the file to clean our dataset. We will remove the duplicate value for county 193. De-duplicating was a step we noted from the EDA from Thomas Drage, Venkatesh Nagapudi, Miguel Jaime.

```
crime <- crime[!is.na(as.numeric(as.character(crime$prbconv))),]
crime$prbconv <- as.numeric(as.character(crime$prbconv))
crime[duplicated(crime), ]
```

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle
## 89 0.07819394
```

```
crime = crime[!duplicated(crime), ]
```

The minimum of the density variable is 0.00002. Let us examine the smallest values for the density variable and determine if there are other similarly low values.

```
head(sort(crime$density), 10)
```

```
##  [1] 0.0000203422 0.3005714420 0.3009985690 0.3167155390 0.3503981830
##  [6] 0.3858093020 0.3887587790 0.4126394090 0.4127659500 0.4349593520
```

There is only one density value which is lower than is plausible (corresponds to an average of 2 people in a 10,000 square mile area) and must be an erroneous value that we shall remove. (This was added after giving feedback on the draft by Thomas Drage, Venkatesh Nagapudi, Miguel Jaime)

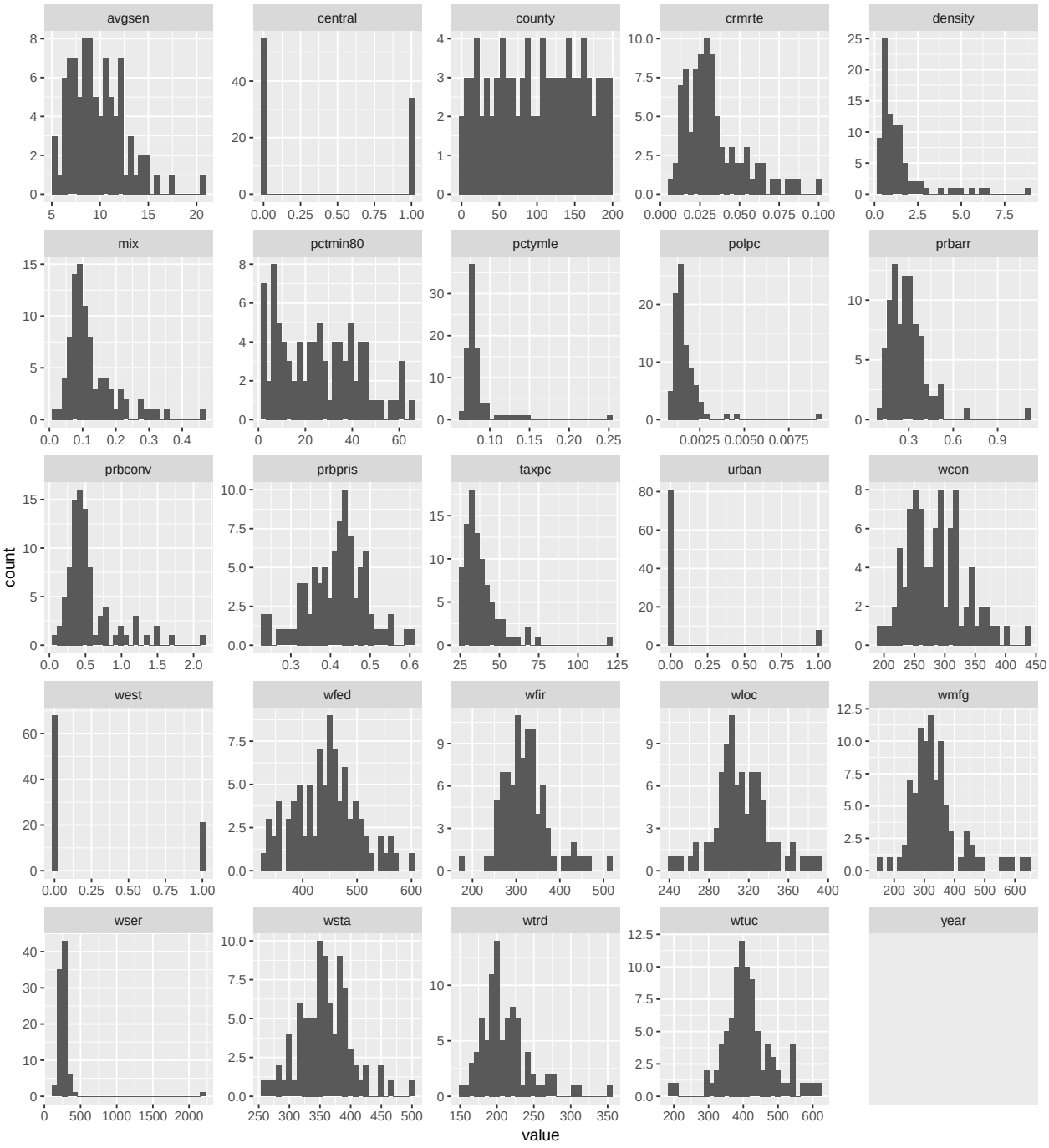
```
crime = crime[crime$density>0.0000203422, ]
```

Something we notice here is that both the probability of arrest and probability of conviction have at least 1 record that is over 1. Thinking through this further, this is not impossible. Multiple people can participate in a crime together, leading to multiple arrests and/or convictions per criminal offense and so this anomaly may be a product of the operationalization of the particular variables. Thus, we will not discard this variable.

Our data file is now clean and ready for further analysis.

Exploratory Data Analysis

```
crime %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```



Initial Tranformation Decisions

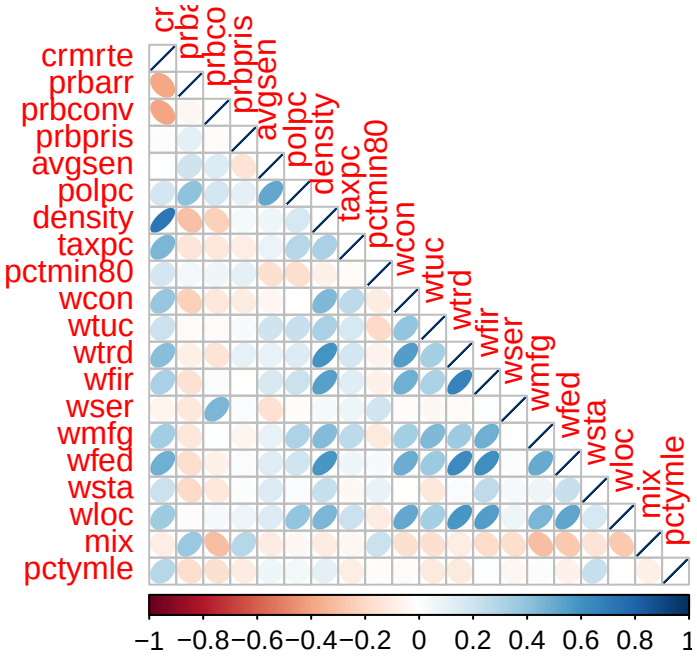
From the above, the distribution of most of the numeric variables presented are approximately normal, with the exception of crime rate, taxpc, density, and polpc that stand out as skewed right and may require a log transformation.

Additional Comments on Correlation

Looking at the variables, we have a few initial thoughts. There is suspected collinearity between wage variables and tax revenue per capita. We also note that higher tax per capita allows counties to spend more money on police forces, possibly having an effect on the police per capita.

We create a correlation matrix to get a high level overview of the correlations between the untransformed variables.

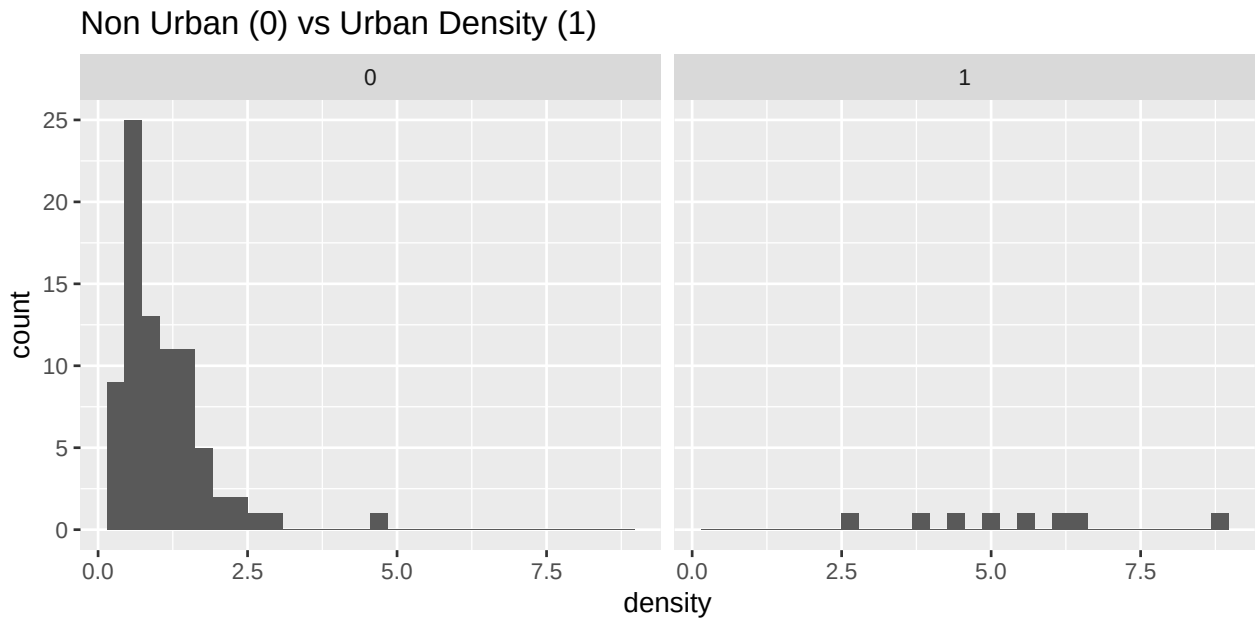
```
corr_crime = cor(crime[, c(-1,-2,-11,-12,-13)])
corrplot(corr_crime, method = "ellipse", type="lower")
```



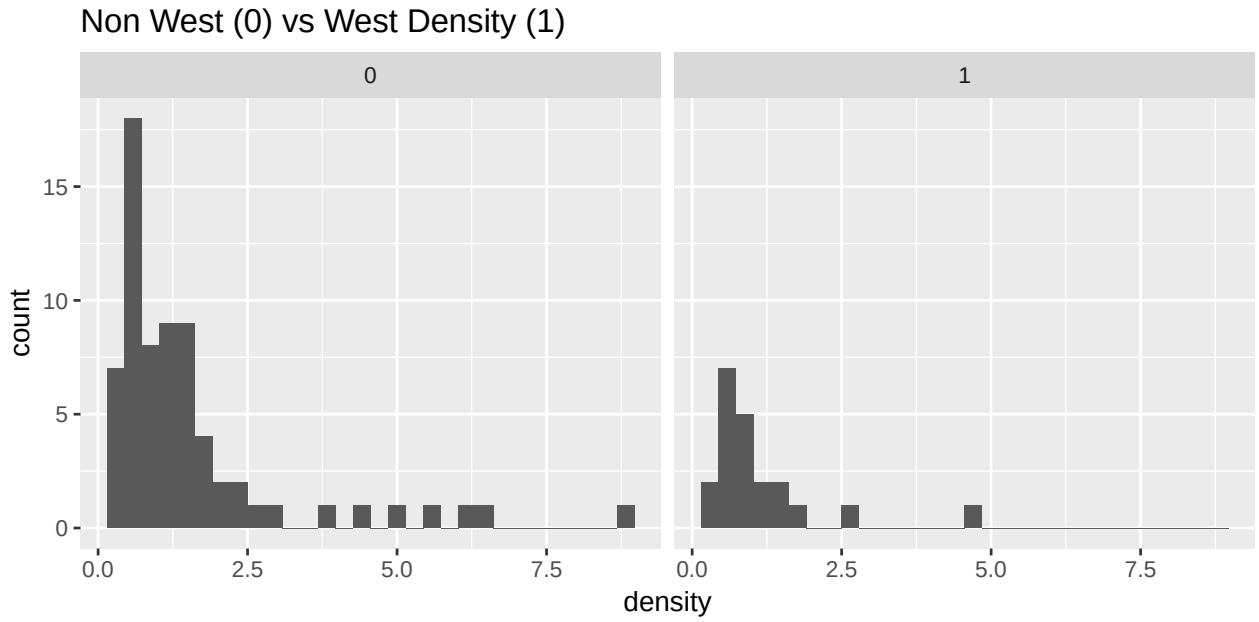
Examining density

The first relationships we decide to investigate further are the density variables against the binary urban, west and central variables. From the correlation matrix and our intuition, we expect that urban areas are more dense.

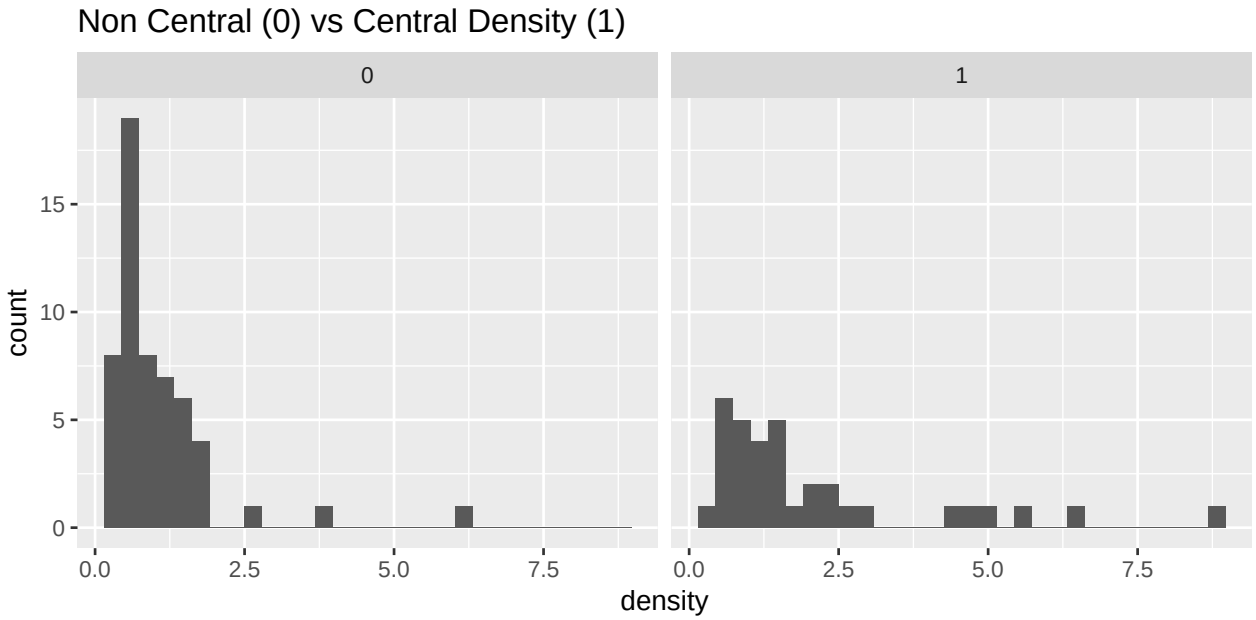
```
crime %>%
  ggplot(aes(density)) +
    facet_wrap(~ urban) +
    geom_histogram() +
    ggtitle("Non Urban (0) vs Urban Density (1)")
```




```
crime %>%
  ggplot(aes(density)) +
    facet_wrap(~ west) +
    geom_histogram() +
    ggtitle("Non West (0) vs West Density (1)")
```



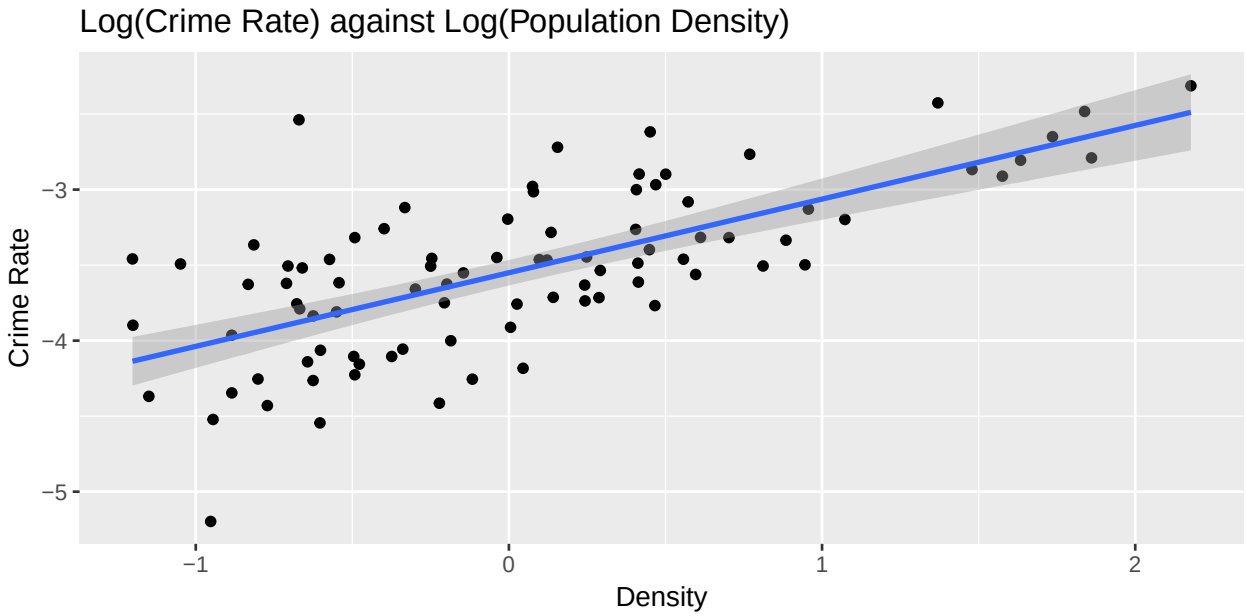

```
crime %>%
  ggplot(aes(density)) +
    facet_wrap(~ central) +
    geom_histogram() +
    ggtitle("Non Central (0) vs Central Density (1)")
```



Urban areas tend to have higher densities, as expected. Thus, the density and urban variables are collinear. The other 2 variables don't show as strong of a relationship with density, and from the scatterplot matrix are not particularly correlated with crime rate. Thus, these indicator variables will likely not be included in our first model specification.

From the distribution of the variables above, we see that taking the log of the density and crime rate would be advisable. Taking a closer look at log(crime rate) against log(density), this does look like a promising variable to include in all our specifications.

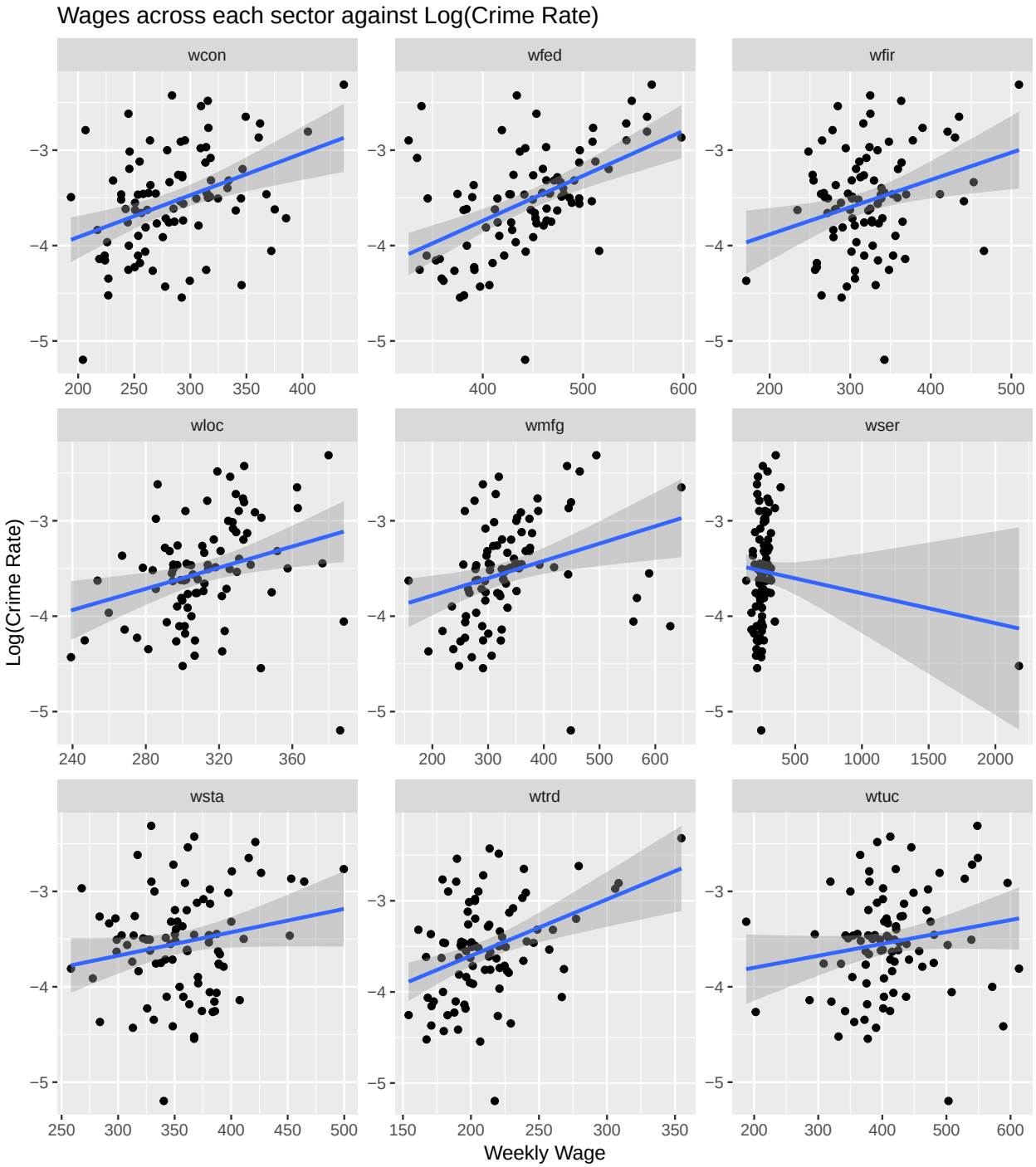
```
ggplot(crime, aes(x=log(density), y = log(crmrte))) +
  geom_point() +
  ggtitle("Log(Crime Rate) against Log(Population Density)") +
  xlab("Density") +
  ylab("Crime Rate") +
  geom_smooth(method = 'lm')
```



Examining Income-related variables

Next, we notice from the scatterplot matrix that each wage variable seems to be highly correlated with each other, and there is some positive correlation with the crime rate. We investigate these closer to see any opportunities for transformation.

```
crime_wage <- crime %>%
  select(crmrte, wcon, wtuc, wtrd, wfir, wmfg, wfed, wser, wsta, wloc) %>%
  gather(sector, wkly_wage, -crmrte)
ggplot(crime_wage, aes(x=wkly_wage, y=log(crmrte))) +
  facet_wrap(~sector, scales = "free") +
  geom_point() +
  ggtitle("Wages across each sector against Log(Crime Rate)") +
  xlab("Weekly Wage") +
  ylab("Log(Crime Rate)") +
  geom_smooth(method = 'lm')
```



Many of the wage variables have a slight positive relationship with log(crime rate). The variables wtrd, wmfg, wfed, and (to a lesser degree) wsta and wloc seem to have a good amount of correlation with crime rate. These relationships explain the phenomenon that more burglaries / thefts / kidnappings are likely to be targeted on wealthier victims, so areas with higher incomes would end up having higher crime rates.

We notice that wser has a significant outlier but other than that the spread of the wage variables is pretty even and does not warrant transformation for our specifications. This outlier reflects a county that has a substantially high wage for service workers, and has a lower crime rate. This is likely an area that has been highly gentrified and is predominantly populated my members of in service roles (with fewer individuals who are in the other, lower paying industries.) We don't believe that removing this outlier is valid, as this county

could have another attribute that is worth investigating as a case study of a "successful" community with a low crime rate, and would be of high interest to our political campaign.

Taking a closer look at this outlier:

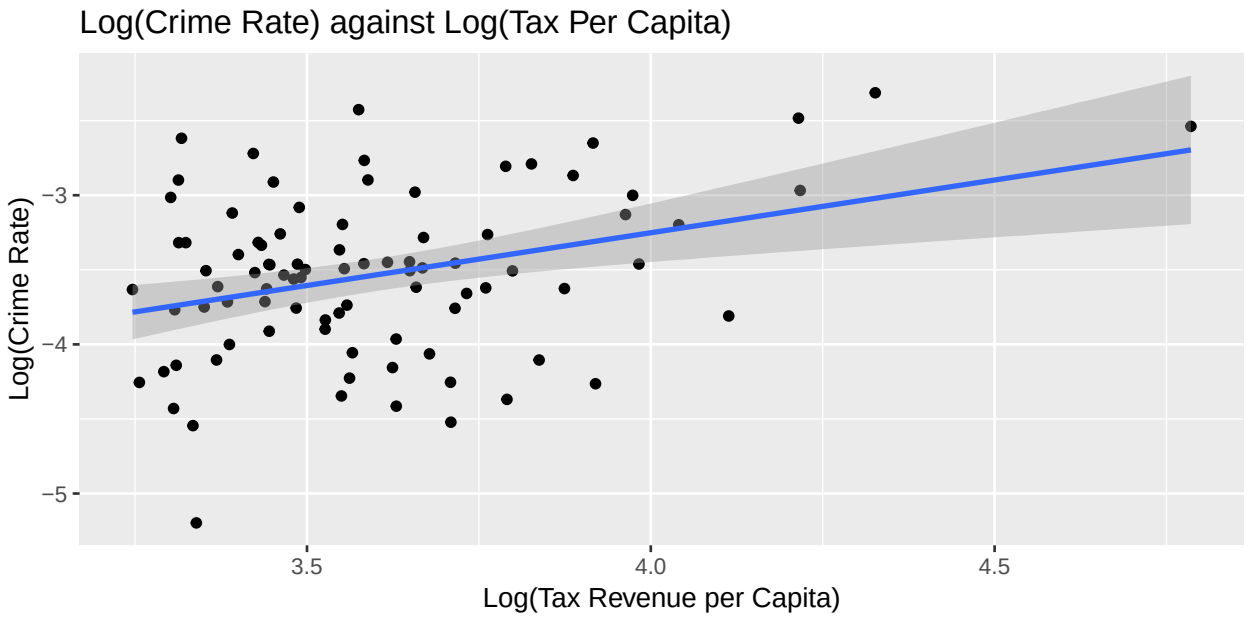
```
crime_outlier <- crime %>% filter(wser>2000)
head(crime_outlier)
```

```
##   county year    crmrte   prbarr prbconv  prbpris avgsen     polpc
## 1    185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.0012221
##     density    taxpc west central urban pctmin80     wcon    wtuc     wtrd
## 1 0.3887588 40.82454    0       1     0  64.3482 226.8245 331.565 167.3726
##       wfir     wser   wmfg   wfed   wsta   wloc        mix    pctymle
## 1 264.4231 2177.068 247.72 381.33 367.25 300.13 0.04968944 0.07008217
```

Interestingly, the tax revenue per capita is lower than what we would expect for such a supposedly affluent county. Further investigation is required here to better understand the exact job market of this population. It is likely that there are only 1 or 2 members of this county who have extremely high paying jobs, driving up wser. In this dataset, we would hope to see a percentage breakdown of workers in each sector. This would allow us to weight each wage parameter accordingly and provide context as to how much of an influence we would expect a sector's wage to have on its crime rate.

As mentioned above, we note that taxpc is related to the wages of workers in each county, as higher taxes are applied to individuals with a higher income. We notice this in the correlation matrix at the top of this report, where taxpc shows at least a light blue relationship with each of the wage variables individually (this is expected as we anticipate that taxpc is more tightly correlated with a linear combination of these variables). From the correlation matrix, we expect to see the same positive relationship between taxpc and crmrte. Taking a closer look at this relationship while taking the log of both variables due to their positive skew in the EDA distribution curves, we can verify that this relationship holds.

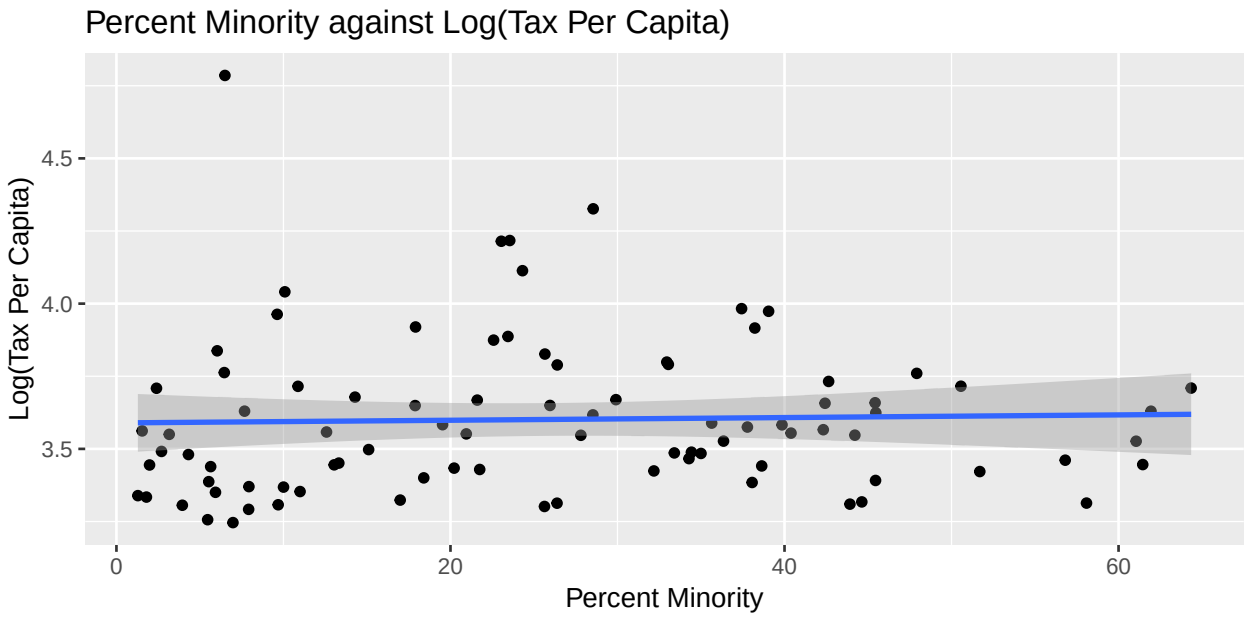
```
ggplot(crime, aes(x=log(taxpc), y = log(crmrte))) +
  geom_point() +
  ggtitle("Log(Crime Rate) against Log(Tax Per Capita)") +
  xlab("Log(Tax Revenue per Capita)") +
  ylab("Log(Crime Rate)") +
  geom_smooth(method = 'lm')
```



Investigating demographic variables

We notice a variable pctmin80, which is the percentage of minority groups in the population. We predict that neighborhoods with higher percent minorities had lower tax revenue per capita, as socio-economic barriers often forced minority groups to take lower paying roles, and racism factors often implied that minority groups would be paid less for the same jobs as white coworkers.

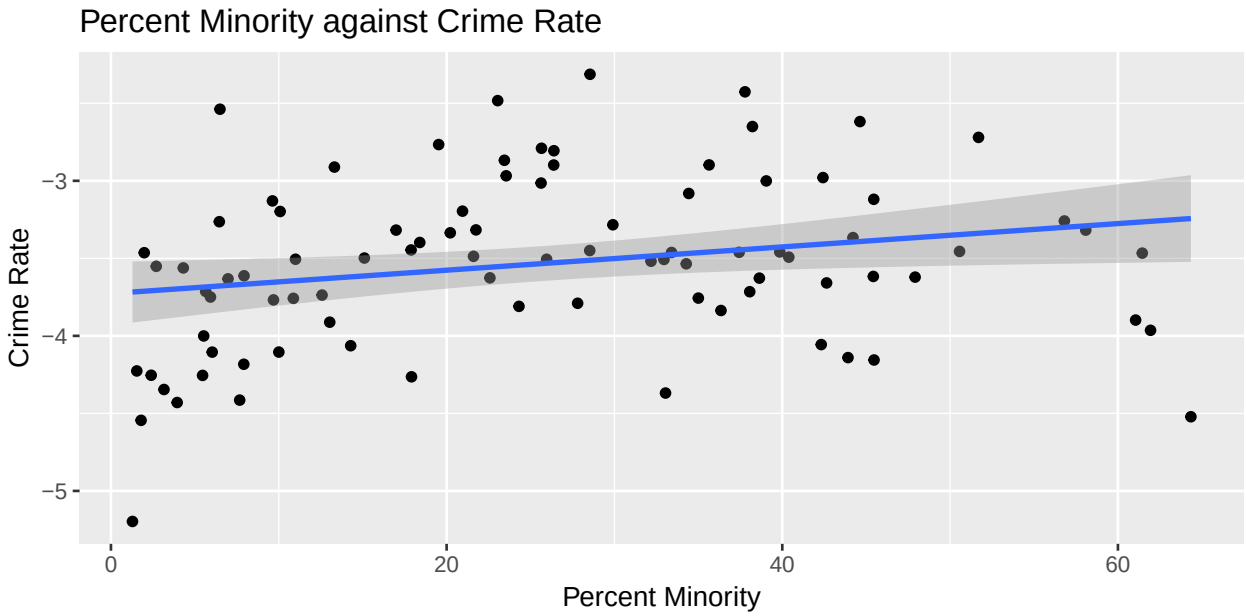
```
ggplot(crime, aes(x=pctmin80, y = log(taxpc))) +
  geom_point() +
  ggtitle("Percent Minority against Log(Tax Per Capita)") +
  xlab("Percent Minority") +
  ylab("Log(Tax Per Capita)") +
  geom_smooth(method='lm')
```



Contrary to the above discussion, Tax Revenue per Capita does not seem to be related to the percentage of minorities in a population. Therefore these variables are likely not multicollinear and can exist within the same specification.

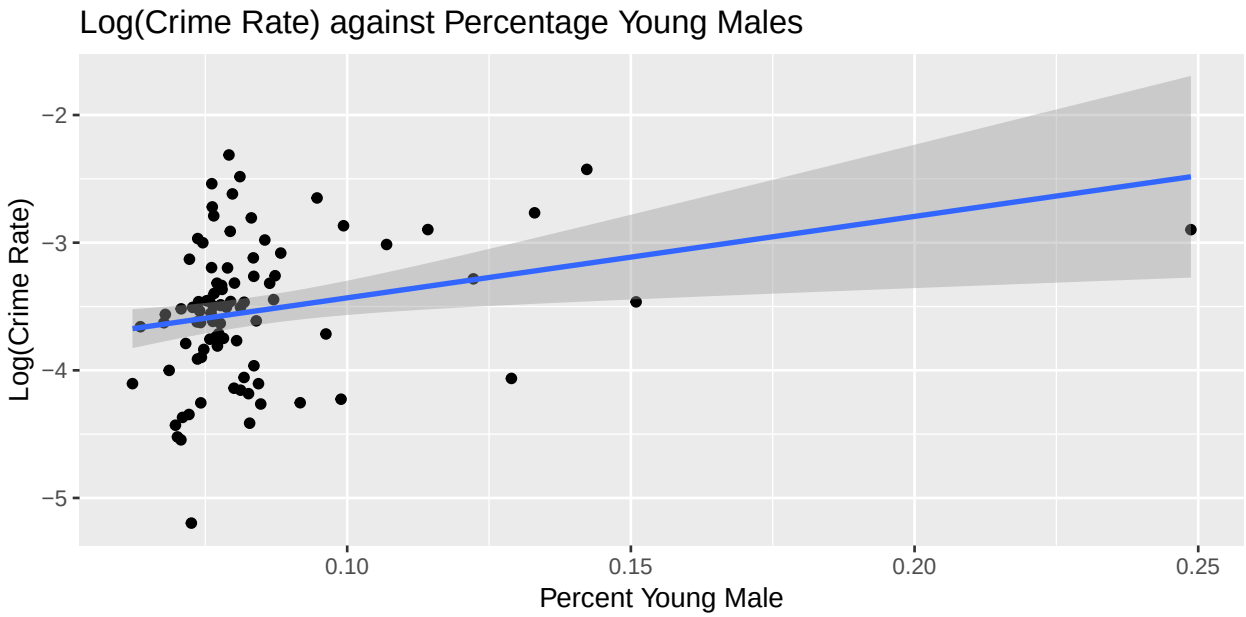
Also, when we look at the relation between Percentage Minority and Crime rate, we some positive correlation between percentage Minority and Crime rate

```
ggplot(crime, aes(x=pctmin80, y = log(crmrte))) +
  geom_point() +
  ggtitle("Percent Minority against Crime Rate") +
  xlab("Percent Minority") +
  ylab("Crime Rate") +
  geom_smooth(method='lm')
```



Now looking at percent young male. This variable is of interest as the perpetrators of crime are often thought to come from this demographic group.

```
ggplot(crime, aes(x = pctymle, y = log(crmrte))) +
  geom_point() +
  ggtitle("Log(Crime Rate) against Percentage Young Males") +
  xlab("Percent Young Male") +
  ylab("Log(Crime Rate)") +
  geom_smooth(method='lm')
```

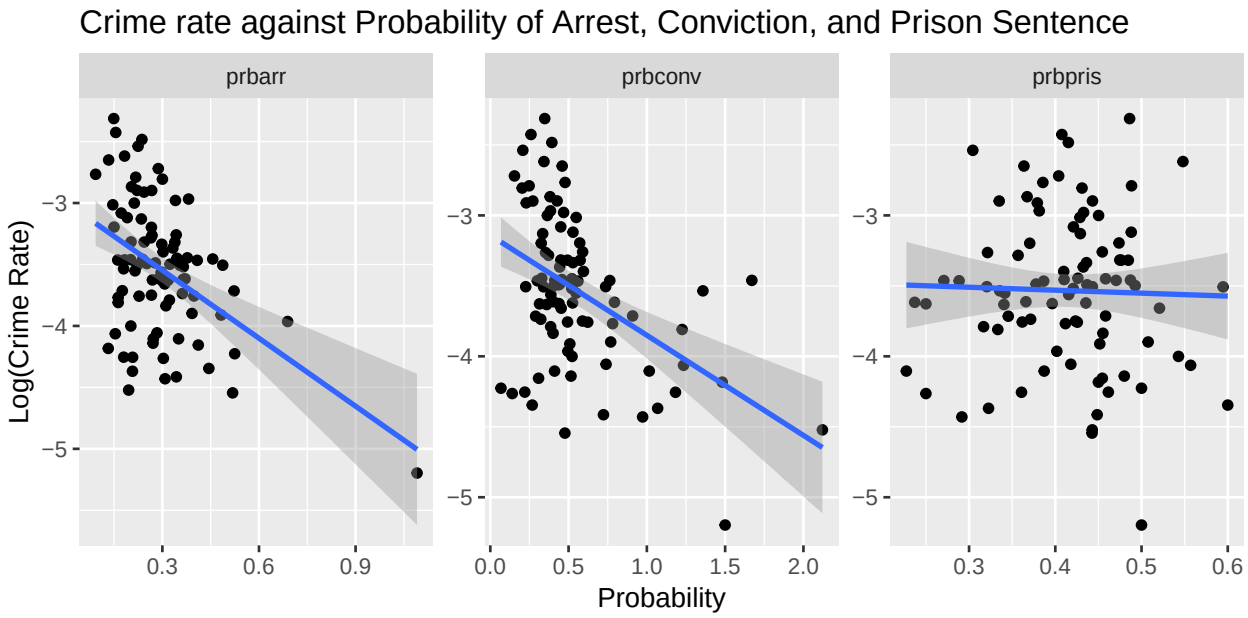


There does appear to be some positive correlation between percent young male and log(crime rate), so we will include this in our model specifications.

The influence of fear - probability of punishment

Now looking at the probabilities associated with arrest, conviction and prison sentence. These 3 probabilites all illustrate the likelihood of being punished for a crime. Therefore, we will onlly use probability of arrest in the first specification and include the others in the remaining specifications.

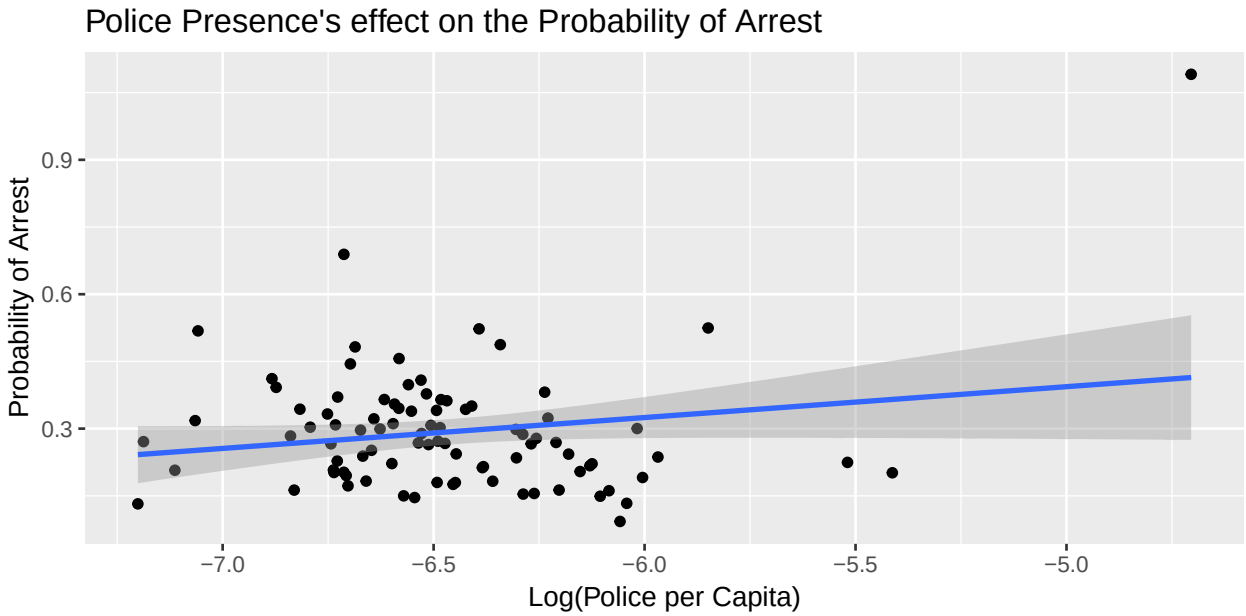
```
crime_prob_punishment <- crime %>%
  select(crmrte, prbarr, prbconv, prbpris) %>%
  gather(punishment, probability, -crmrte)
ggplot(crime_prob_punishment, aes(x=probability, y=log(crmrte))) +
  facet_wrap(~punishment, scales = "free") +
  geom_point() +
  ggtitle("Crime rate against Probability of Arrest, Conviction, and Prison Sentence") +
  xlab("Probability") +
  ylab("Log(Crime Rate)") +
  geom_smooth(method = 'lm')
```



As suggested by the correlation matrix and the analysis above, there is a negative relationship between the probability of arrest and conviction with crime rate. However, there seems to be no correlation between the probability of a prison sentence with crime rate.

From intuition, police presence can either be positively related with crime (more police are needed in more crime active areas) or they can be negatively related (higher plice presence serves as a deterrant of crime). In the former case, police presence is an outcome variable of crime, and in the latter case, crime is the outcome variable. We plot the probability of arrest against the log transformation of polpc (due to its positive skew), to assess multicollinearity

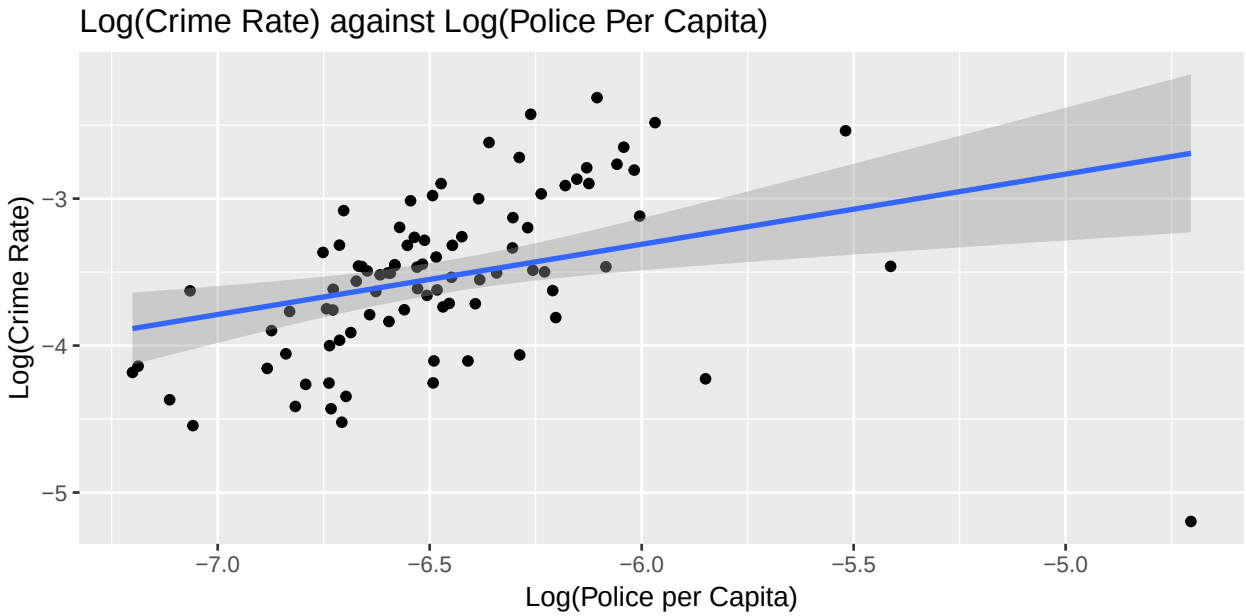
```
ggplot(crime, aes(x = log(polpc), y = prbarr)) +
  geom_point() +
  ggtitle("Police Presence's effect on the Probability of Arrest") +
  xlab("Log(Police per Capita)") +
  ylab("Probability of Arrest") +
  geom_smooth(method = 'lm')
```



There is one outlier where a very high police per capita leads to a high probability of arrest. This likely could be a neighborhood where crime is especially high, so police are typically stationed here. This is not grounds to remove the outlier from our analysis, though we do note that, aside from this point, there doesn't seem to be a relationship between the police per capita and probability of arrest. This suggests that both variables can exist within our model

We question if more police officers could be a deterrent of crime from occuring in the first place. Thus, let us plot the log transformed police per capita and crime rate.

```
ggplot(crime, aes(x = log(polpc), y = log(crmrte))) +
  geom_point() +
  ggtitle("Log(Crime Rate) against Log(Police Per Capita)") +
  xlab("Log(Police per Capita)") +
  ylab("Log(Crime Rate)") +
  geom_smooth(method = 'lm')
```



From above, the more police there are the more crime there is, with the exception of the outlier at the bottom right, reflecting an area with high police per capita and low crime. This is the same point as the outlier of the previous graph. Therefore, this one county has a high number of police with a high likelihood of arreset, and a low crime rate. This likely is an area that is actively cracking down on crime. Overall though, we cannot make any conclusions on whether police presence reduces crime, or whether crime increases police presence, or whether police presence provokes criminals and incites crime. Therefore, this variable will likely not be included as part of our first model specification, but we will consider it for our other models.

```
crime_outlier2 <- crime %>% filter(polpc>0.0075)
head(crime_outlier2)
```

```
##   county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 1    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##     density   taxpc west central urban pctmin80     wcon     wtuc     wtrd
## 1 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351 217.4908
##       wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 1 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

Bringing our analysis together

As we build our models, we shall create three main specifications. The first specification, model1 consists of the variables that are highly correlated with the log transformed crime rate variable only as per the preceding exploratory data analysis. This is the most parsimonious and consists of the tax per capita, density, percentage of young males and the probability of arrest.

The second specification, model2 consists of the variables in the first specification plus those we think are likely to affect the crime rate as well as those that have some correlation with the target variable. The variables added in this specification are the log transformed police per capita, probability of conviction, probability of arrest and the percentage of minorities.

The third specification contains all the appropriately transformed variables in the dataset. To be noted is that we incorporate all the wage variables.

An exploratory fourth specification investigates if there are any interaction effects between the variables depicting probabiltiy of punishment (arrest, conviction, sentence)

```
model1 <- lm(log(crmrte) ~ log(taxpc) + log(density) + pctymle + prbarr, data = crime)
model2 <- lm(log(crmrte) ~ log(taxpc) + log(density) + pctymle + log(polpc) + prbarr + prbconv + prbpris
model3 <- lm(log(crmrte) ~ log(taxpc) + log(density) + pctymle + prbarr + prbconv +
               prbpris + avgsen + log(polpc) + pctmin80 + wcon +
               wtuc + wtrd + wfir + wser + wmfg +
               wfed + wsta + wloc + mix, data = crime)

crime$arr_sent <- crime$prbarr * crime$avgsen
crime$pris_sent <- crime$prbpris *crime$avgsen
crime$arr_pris <- crime$prbarr * crime$prbpris

model4 <- lm(log(crmrte) ~ log(taxpc) + log(density) + pctymle + prbarr + log(polpc) + prbconv + pctmin8
```

The Regression Table

We generate a regression table displaying the 4 models side by side

```
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
se.model3 = sqrt(diag(vcovHC(model3)))
se.model4 = sqrt(diag(vcovHC(model4)))

stargazer(model1,model2,model3, model4
          , type = "latex"
          , column.labels   = c("Specification 1", "Specification 2 ", "Specification 3", "Specificatio
          , title = "Model Summaries"
          , omit.stat="f"
          , keep.stat = c("rsq","adj.rsq")
          , se = list(se.model1,se.model2,se.model3, se.model4)
          , star.cutoffs = c(0.05,0.01,0.001)
          , font.size = "small"
          , single.row = TRUE
          , omit.table.layout = "n"
          , add.lines=list(c("AIC", round(AIC(model1),1), round(AIC(model2),1), round(AIC(model3),1), r
          c("BIC", round(BIC(model1),1), round(BIC(model2),1), round(BIC(model3),1), round(BIC(model4),
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Dec 09, 2018 - 16:04:04

It is clear from the results that the second model has the best balance between parsimony and explaining the variation in the outcome variable. Additionally, the interaction terms did not seem to benefit our model in Specifcation 4 as AIC and BIC increased, and adjusted R squared decreased.Therefore we will no longer explore Specification 4.

It is interesting to note that both the AIC and BIC point are lower for the third specification that contains nearly all the variables than for the second specification that seems to be more parsimonious and hence could have been expected to be a better model. This also applies for the adjusted R squared value.

Discussing Statistical and Practical Significance

Across the four models, log(density) remains stasticically significant and has a relatively consistent coefficient, suggesting a 0.27-0.37 * x% increase in crime with every x% increase in density. When put in these terms, the density term is practically significant.

Table 1: Model Summaries

	Dependent variable:			
	log(crmrte)			
	Specification 1	Specification 2	Specification 3	Specification 4
	(1)	(2)	(3)	(4)
log(taxpc)	0.448 (0.243)	0.028 (0.178)	0.050 (0.194)	0.010 (0.186)
log(density)	0.378^{***} (0.063)	0.276^{***} (0.070)	0.269^{***} (0.075)	0.274^{***} (0.071)
pctymle	3.531^{**} (1.141)	0.265 (2.132)	1.065 (1.519)	0.328 (2.082)
log(polpc)		0.480^{***} (0.126)	0.537^{**} (0.186)	0.537^{***} (0.156)
prbarr	−0.850 (0.481)	-1.636^{***} (0.269)	-1.662^{***} (0.265)	−1.434 (2.250)
prbconv		-0.623^{***} (0.106)	-0.501^{***} (0.150)	-0.597^{***} (0.132)
prbpris		−0.425 (0.357)	−0.531 (0.399)	
avgsen			−0.015 (0.014)	0.014 (0.047)
arr_sent				−0.013 (0.153)
pris_sent				−0.048 (0.078)
arr_pris				0.033 (2.662)
pctmin80		0.013^{***} (0.002)	0.013^{***} (0.002)	0.013^{***} (0.002)
wcon			0.0003 (0.001)	
wtuc			0.0001 (0.001)	
wtrd			0.001 (0.001)	
wfir			−0.002 (0.001)	
wser			−0.0003 (0.002)	
wmfg			−0.0003 (0.0005)	
wfed			0.001 (0.001)	
wsta			−0.001 (0.001)	
wloc			0.001 (0.002)	
mix			−0.097 (0.575)	
Constant	-5.209^{***} (0.915)	0.103 (1.027)	0.468 (1.624)	0.383 (1.341)
AIC	79.2	-4.1	-0.9	0.1
BIC	94.1	20.8	51.4	32.5
R^2	0.578	0.849	0.877	0.852
Adjusted R^2	0.558	0.833	0.844	0.830

The percentage of young males is statistically significant at the 0.05 level in specification 1, and is not significant in the remaining models. This suggests that with the addition of variables between specifications 1 and 2, a true causal factor came into play and the effect of pctymle decreased. Thus, we don't take this to be a predictor of crime rate

2 of the 3 probability of punishment factors (arrest and conviction) are significant in our preferred specification, Specification 2. This suggest when the probability of arrest goes up by 1, the crime rate drops 160% (of course, the probility of arrest will likely fluctuate in smaller ammounts, but still the degree of change would still be practically significant at these levels.) The same applies for the probability of conviction (a unit increase in prbconv relates to a 62% drop in crime rate). Overall, it is likely that a higher chance of arrest and conviction deters would-be criminals from engaging in criminal activity.

The police per capita, included in specifications two and three is also statistically significant at the 0.005 level for the second model and at 0.01 in the third model. However, the police per capita variable is likely both a predictor and an outcome variable as well. Crime rate is likely affects the police per capita in area, but police presence may either incite or deter criminals. With all these in mind, we do not consider the police per capita to be practically significant.

Lastly, the percentage of minorities is also statistically significant. However, the coefficient on this term is very small (corresponding to a 1% increase in crime rate with every unit increase in pctmin80. We expect this coefficient to be practically smaller as the pctmin80 will likely fluctuate in values much less than 1). Thus, we do not feel that this coefficient is practically significant.

Joint Significance of Wage variables

```
waldtest(model2, model3, vcov= vcovHC)
```

```
## Wald test
##
## Model 1: log(crmrte) ~ log(taxpc) + log(density) + pctymle + log(polpc) +
##     prbarr + prbconv + prbpris + pctmin80
## Model 2: log(crmrte) ~ log(taxpc) + log(density) + pctymle + prbarr +
##     prbconv + prbpris + avgsen + log(polpc) + pctmin80 + wcon +
##     wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc + mix
##   Res.Df Df      F Pr(>F)
## 1     80
## 2     69 11 0.5993  0.823
```

We wanted to see if wages can be used as a proxy for employment. But we see that the wages are not jointly significant along with not being significant individually. So, this supports our argument of not including wages in our final model.

Comparing the effects of the different probabilities on crime rate

We want to examine if the different probability variables in the dataset affect the log(crime rate) in the same way, that is, are their coefficients the same in the population?

First we test the hypothesis that the probabibility of arrest and the probability of conviction affect the crime rate in the same way.

```
linearHypothesis(model2, c( "prbarr=prbconv"), vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr - prbconv = 0
```

```
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ log(taxpc) + log(density) + pctymle + log(polpc) +
##     prbarr + prbconv + prbpris + pctmin80
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     81
## 2     80  1 17.493 7.331e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of the F-test is statistically significant at the 0.001 level so we can reject the hypothesis that the two variables have the same coefficient. Therefore there is evidence that we need to keep both of these viarables in our model. We then compare the probability of arrest with the probabibility of conviction.

```
linearHypothesis(model2, c( "prbarr=prbpris"), vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr - prbpris = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ log(taxpc) + log(density) + pctymle + log(polpc) +
##     prbarr + prbconv + prbpris + pctmin80
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F   Pr(>F)
## 1     81
## 2     80  1 7.7716 0.006628 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again we reject this hypothesis with a type 1 error rate under 0.01. Finally we compare the probability of getting a prison sentence with the probability of conviction.

```
linearHypothesis(model3, c( "prbpris=prbconv"), vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## - prbconv  + prbpris = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ log(taxpc) + log(density) + pctymle + prbarr +
##     prbconv + prbpris + avgsen + log(polpc) + pctmin80 + wcon +
##     wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc + mix
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1     70
## 2     69  1 0.0048  0.945
```

In this final case we are unable to reject the null hypothesis that the two probabilities are equal. This is in line with what we would expect from the two probabilities, since often times conviction will result into a conviction so it is likely that the effect on crime rate is the same. However, since we noticed that the prbpris is not statistically significant in our regression table, we can attribute most of the effect to come from prbconv.

Checking for CLM assumptions

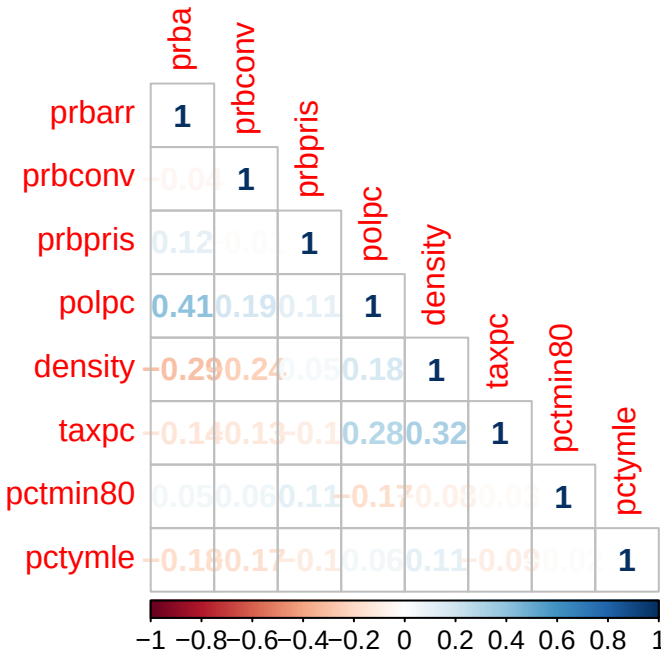
(1) Linear model assumption As we are not restricting the error term, we don't have to worry about the linear model assumption.

(2) Random Sampling We have 89 of the 100 counties in Carolina in our analysis sample population. As of 2016, we have 80 rural counties, we expect that number to be higher in 1980s, so we can assume that we have a random sample.

(3) Multicolinearlity

We know that there isn't perfect multi collinearity as R would throw an error saying it has encountered singularity. We have, as part of our data analysis, have identified few variables which are highly correlated and did not include them in the model.

```
corr_mod2 <- cor(crime[,c(4,5,6,8,9,10,14,25)])
corrplot(corr_mod2, method =c("number"), type="lower")
```



From the above correlation matrix that looks at the variables used in Specification 2, we gather that none of the untransformed explanatory variables have a correlation value of more than 0.41. We can assume that we have no multicolinearlity.

Another way to look at this is by looking at the Variance Inflation factors.

```
vif(model2)
```

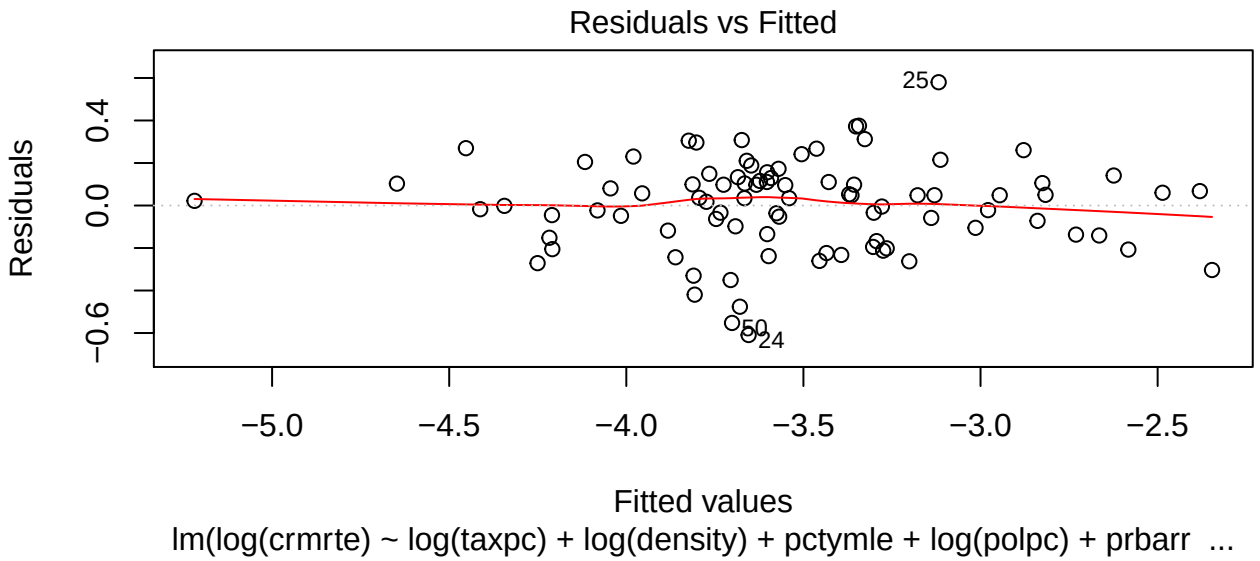
```
##   log(taxpc) log(density)      pctymle   log(polpc)       prbarr
##     1.440505     1.651743     1.227652     1.832683     1.572752
##      prbconv      prbpris     pctmin80
##     1.234763     1.067973     1.088247
```

Since none of the values are over 4, we are safe to say that there is no strong evidence of multicolinearity.

(4) Zero Conditional Mean

Let's start looking at the diagnostic plots to talk about the rest of the assumptions.

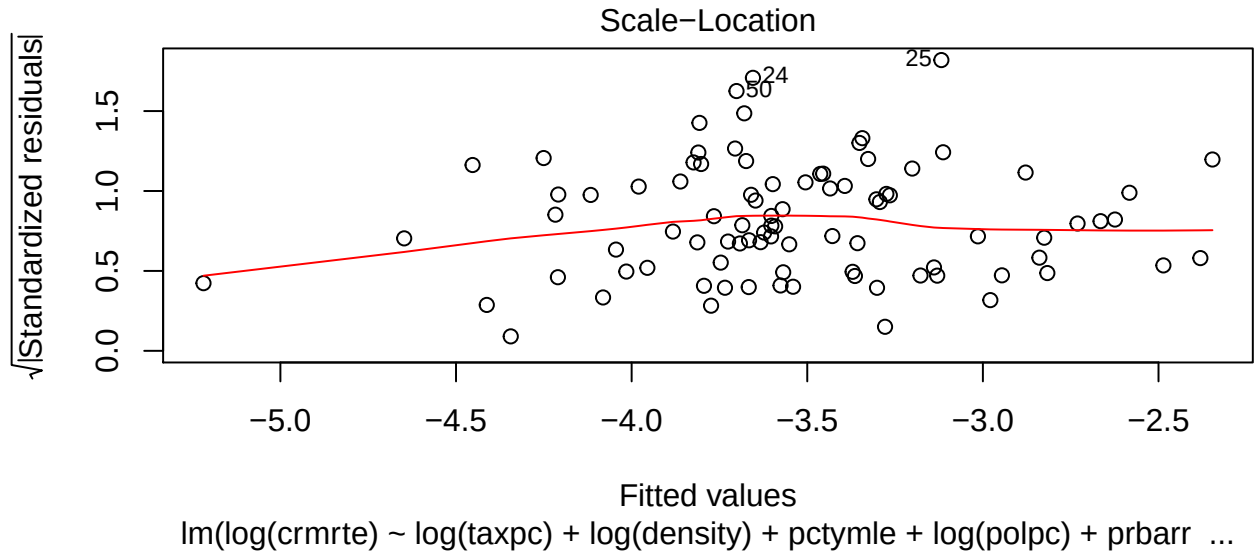
```
plot(model2, which = 1)
```



From the residual vs the fitted plot, we see that we have met the assumption of zero conditional mean, as the red spline curve is fairly flat across the whole graph. This suggests that our coefficients are both consistent and unbiased. Also, we can assume exogeniety. Still, we expect there to be a fair amount of omitted variables that were not captured in the dataset, though they may be strongly correlated with some of our model variables (and thus giving zero conditional means).

(5) Homoskedasticity

```
plot(model2, which = 3)
```



Here we examine the spread of the residuals from the residuals vs fitted plot as well as the straightness of the

mean values of the scale-location plot. From the residuals vs fitted plot, the variance of residuals seems to be highest in the center of the plot. Looking at the scale-location plot, there is a slight dip in the plot in the left region of the graph, however this could be attributed to the lower density of points at this region. As we do not have a strong case for homoskedasticity, we will continue to leverage the heteroskedastic-robust standard errors for our specificaitons, shown again below.

```
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   0.102601   1.026601  0.0999 0.9206404
## log(taxpc)    0.028099   0.177886  0.1580 0.8748842
## log(density)  0.276184   0.070063  3.9419 0.0001721 ***
## pctymle       0.265458   2.132412  0.1245 0.9012419
## log(polpc)    0.479807   0.125688  3.8175 0.0002646 ***
## prbarr       -1.635899   0.269359 -6.0733 3.999e-08 ***
## prbconv      -0.622569   0.106390 -5.8517 1.022e-07 ***
## prbpris      -0.424869   0.356722 -1.1910 0.2371614
## pctmin80      0.012920   0.001672  7.7270 2.766e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the robust errors, log(density), probability of arrest, probability of conviction, and percentage of minorities are statistically significant. Log(polpc) is also significant, but we take this with a grain of salt as this could both be a response and an explanatory variable.

The scale-location plot shows the heteroskedastic as we see that there are outliers within the data set. Another way for checking for heteroskedasticity is by looking at the variance. That can be checked as shown below

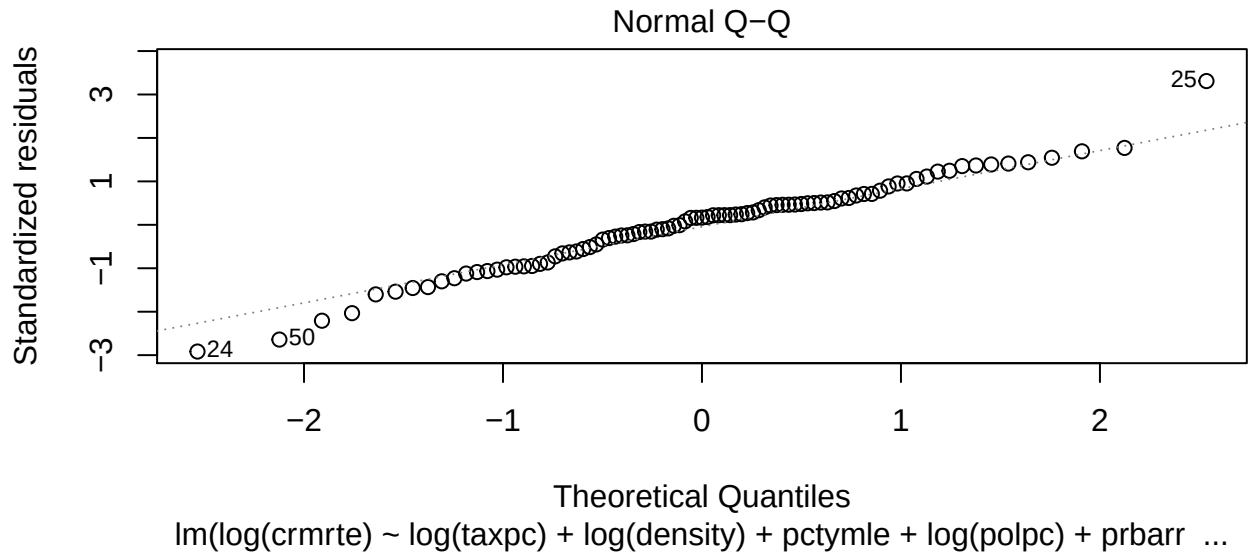
```
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.001397912, Df = 1, p = 0.97018
```

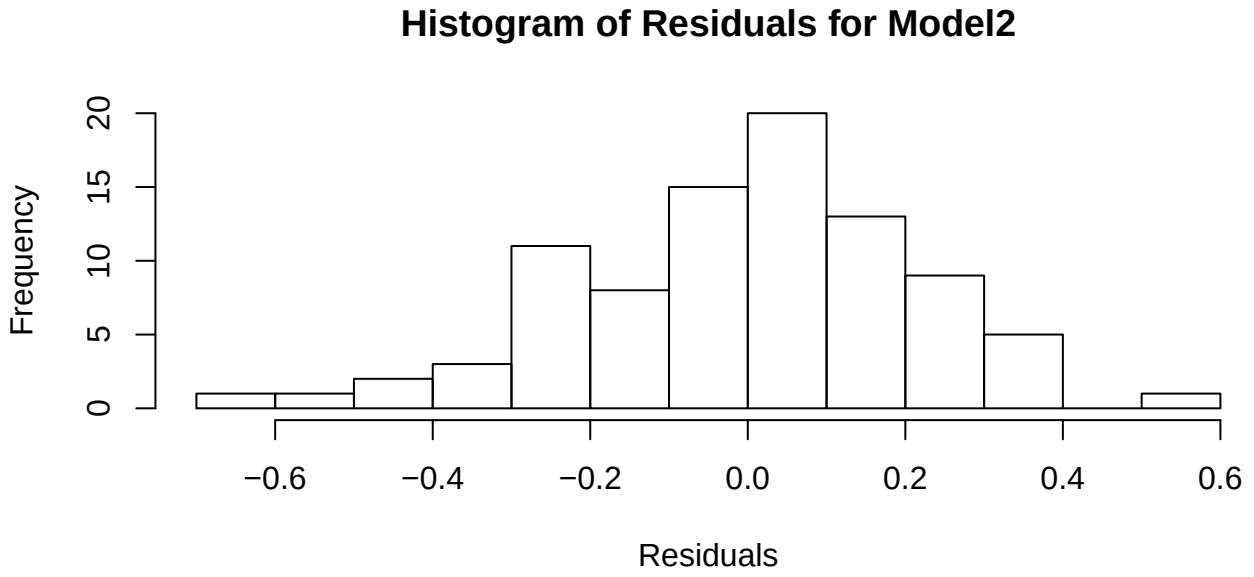
The above results indeed supports our inference that it is heteroskedastic as the p value is less than 0.05

(6) Normality of Errors

```
plot(model2, which = 2)
```

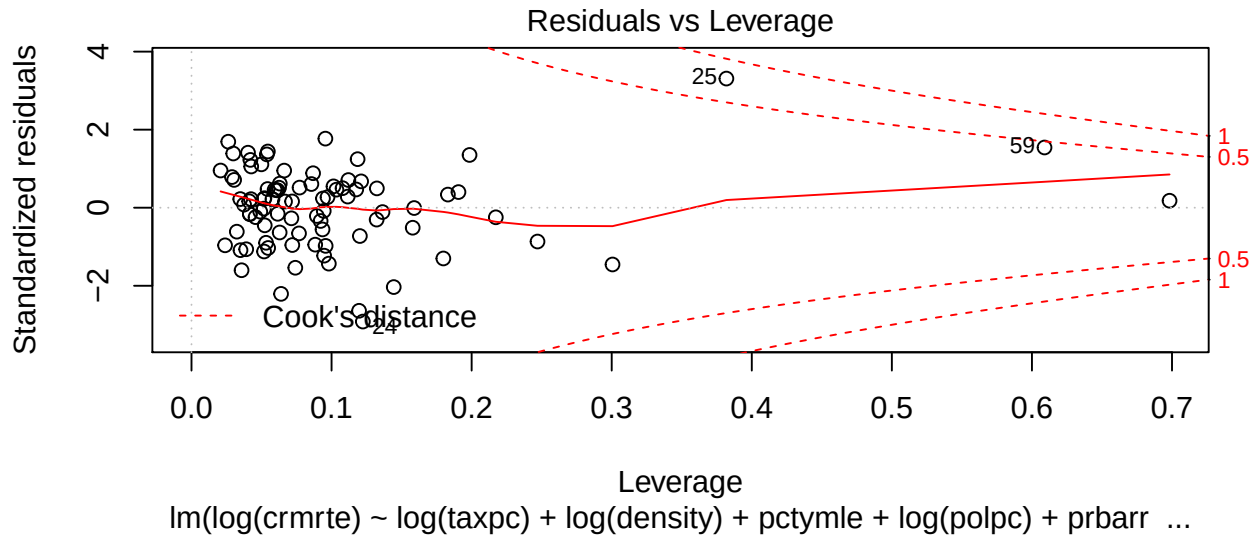



```r
hist(model2$residuals, main = "Histogram of Residuals for Model2", xlab = "Residuals", breaks =15)
```



When we look at the Q-Q plot, we notice that except for a few points toward the left hand side of the graph, most of the data lies close to the dotted line, ensuring we have a normal error distribution. Additionally, by plotting the residuals above, we see that this distribution looks fairly normal.

```r
plot(model2, which = 5)
```



Also, we see observation 25 falls outside the Cook's distance line of 0.5 but within a Cook's distiance of 1. So, this could be an influencial outlier that we should be looking at. The main challenge with this observation is that tax per capita is way more than what we see with all the other observations.

Another way to look at the outliers

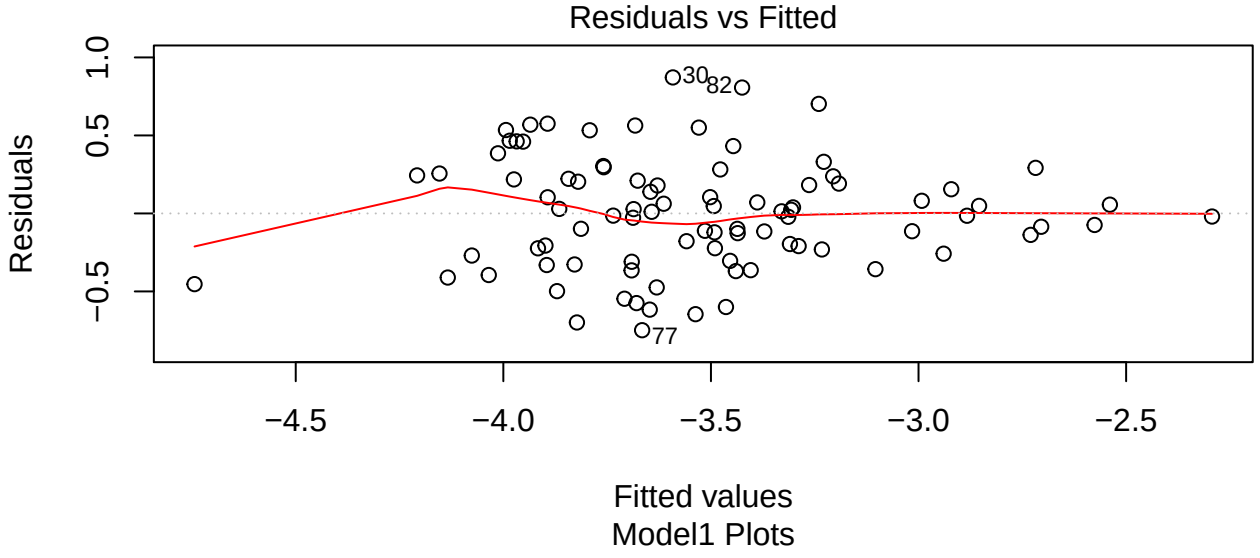
```
outlierTest(model2)
```

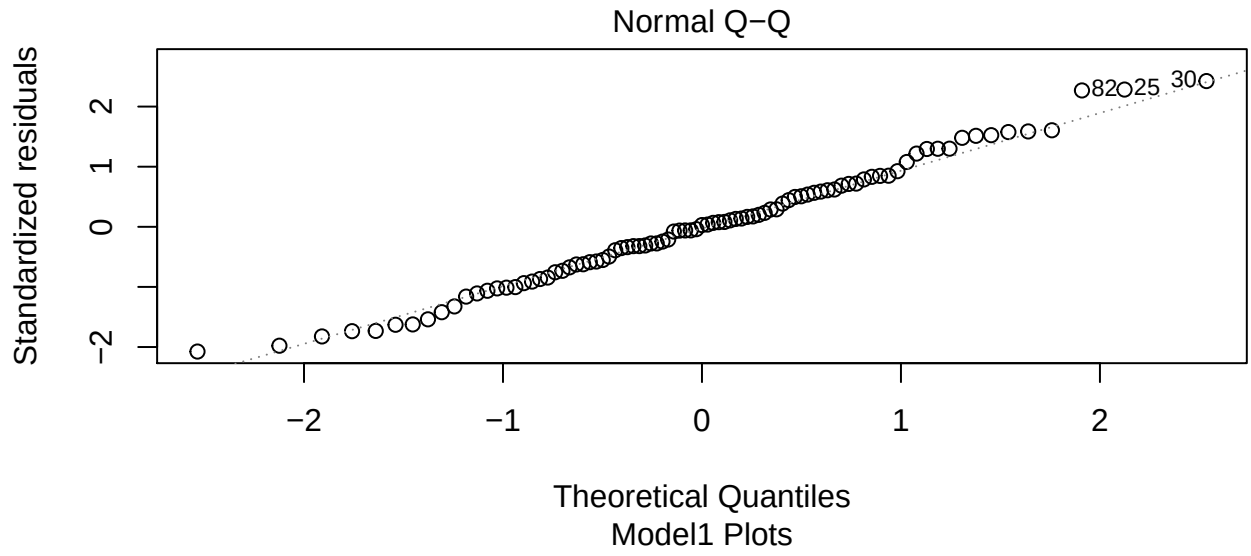
```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 25 3.539776         0.00067464     0.060043
```

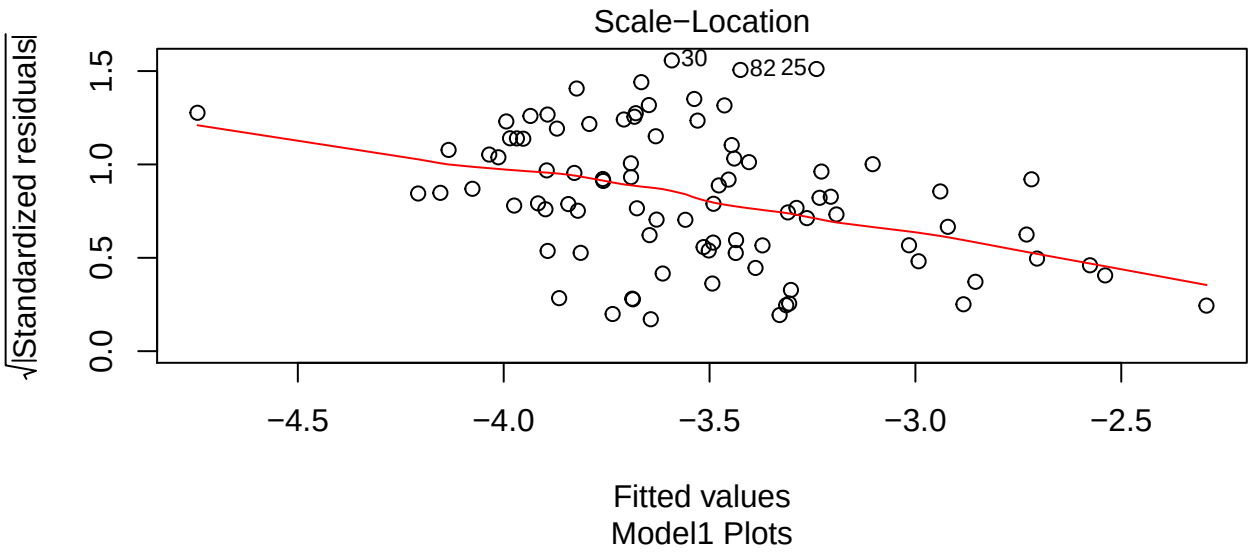
Similar to what we saw in the plot, we see that observation 25 is the most extreme value, but the P value is pretty colse to 0.05, but the earlier plot shows this is an outlier.

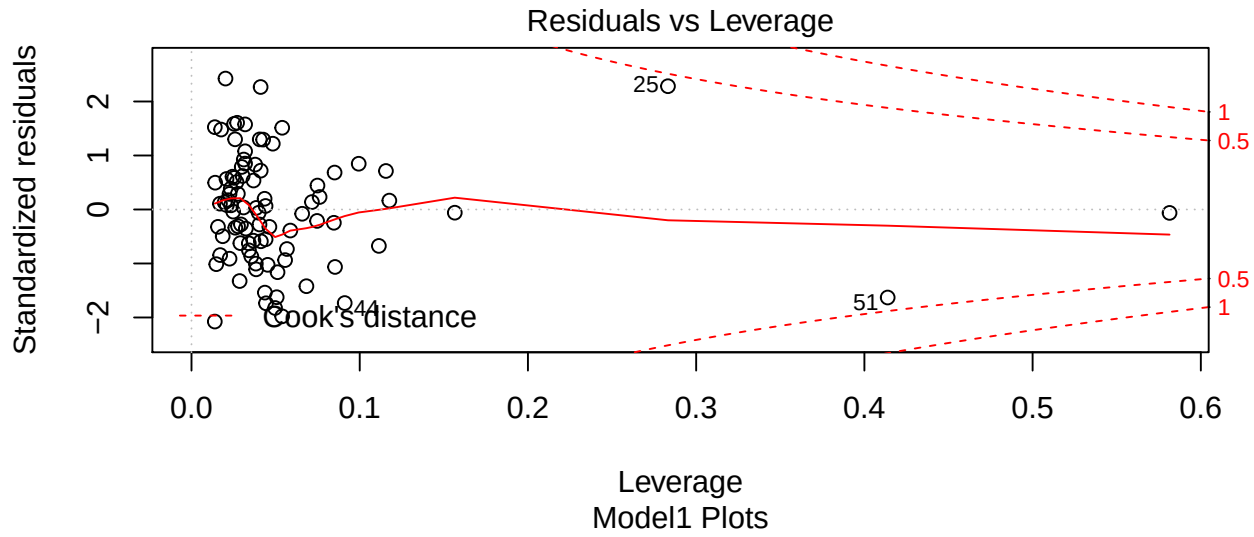
We can look at the diagnostic plots for other models as well:

```
plot(model1, sub.caption  = "Model1 Plots", cex.caption = 1)
```



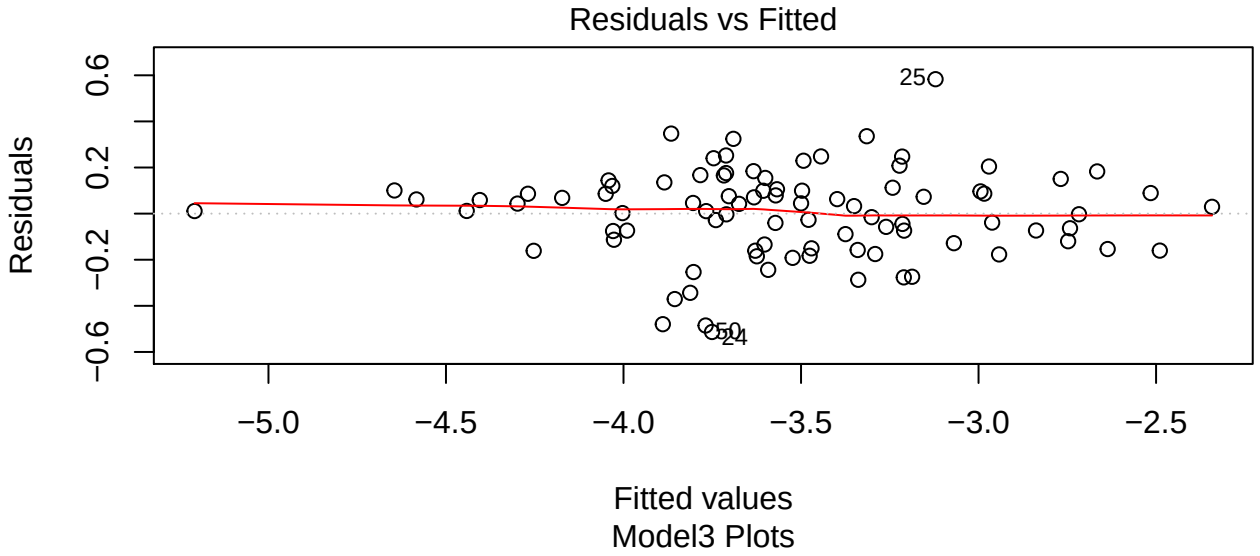

Normal Q−Q
Standardized residuals
Theoretical Quantiles
Model1 Plots
82 25 30
−2
−1
0
1
2

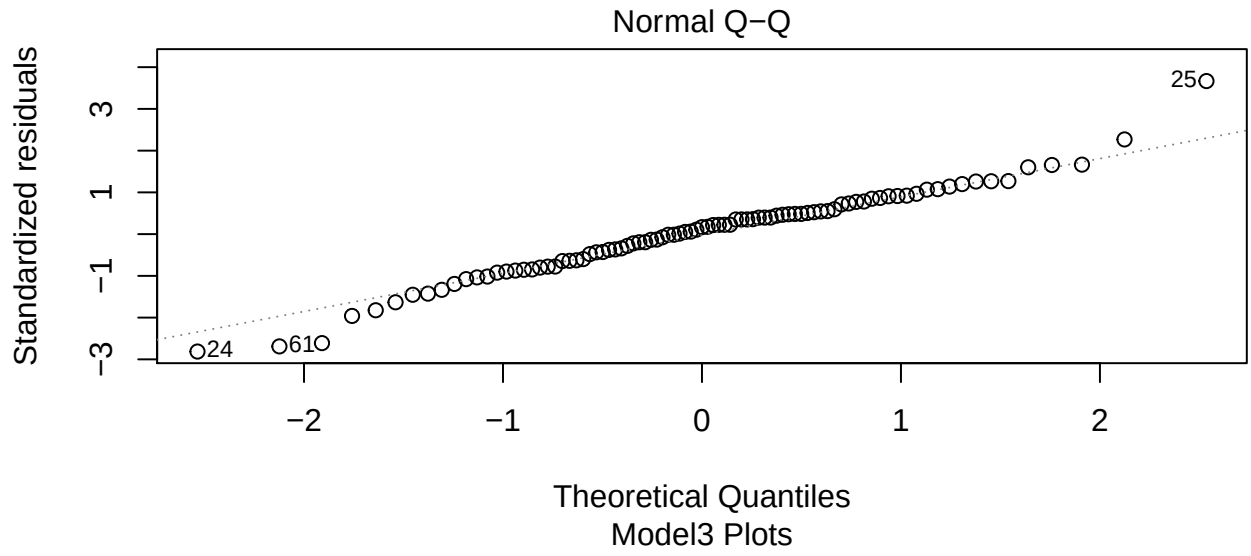

Scale−Location
√|Standardized residuals|
Fitted values
Model1 Plots
30
82 25
0.0
0.5
1.0
1.5
−4.5
−4.0
−3.5
−3.0
−2.5

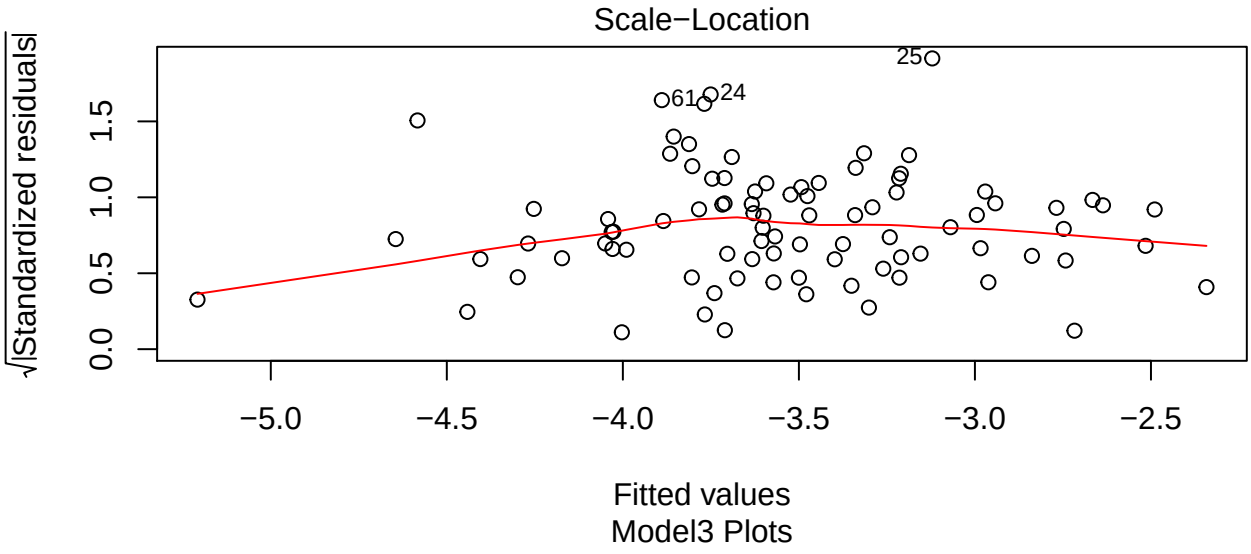


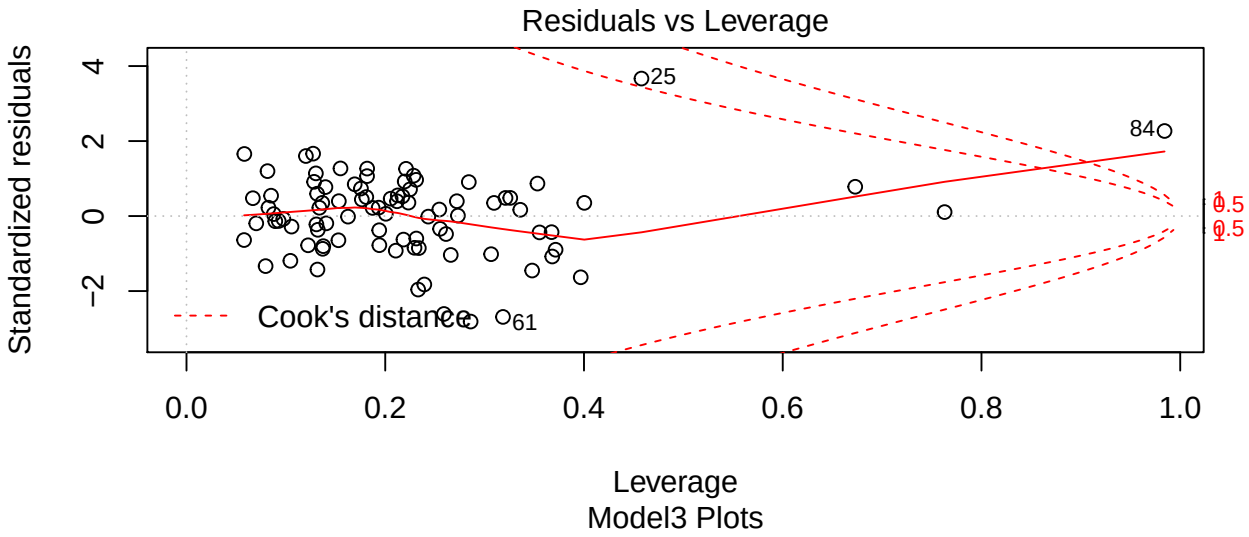
The residuals vs fitted plot for Model1 is less flat than the one for model2, aligning with our expectations that model2 is the more suitable model.

```
plot(model3, sub.caption  = "Model3 Plots", cex.caption = 1)
```




Normal Q−Q
Standardized residuals
−3
−1
1
3
25
24
61
−2
−1
0
1
2
Theoretical Quantiles
Model3 Plots


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
25
61
24
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
Fitted values
Model3 Plots



In model 3, to contrast, the residual vs fitted plot is again flat like it is for Model 2. We see similar results for the normality of errors and heteroskedasticity. However, our outlier plot shows an additional point 84 that is now an outlier. This could be due to overfitting the model by using a kitchen-sink approach when developing model 3.

The Omitted Variables Discussion

There are several omittied variables that would be valuable in conducting this analysis:

1. Severity of crime. Crimes can vary from being petty (jaywalking or parking in a no parking zone) to severe crimes that do warrant arrest, conviction and prison sentences (kidnapping, thefts, sexual violence). Having a parameter that indicates the severity of the crime would help differentiate the varying levels of crime and focus analysis on reducing the likelihood of harsher crimes. The crime severity would be positively correlated with the crime rate and the probability of conviction but negatively correlated with the probability of arrest and the average sentence. This may lead to a negative coefficient because of the higher magnititude of the coefficient for the probability of arrest.

2. Income gap. There are several variables that point to the affluence of a region, but we are interested in seeing the percentage of upper/middle class individuals compared to percentage of lower class. We predict that the difference in these percents would be a better indicator of crime rate. Currently, we only have the wage within each sector (it is unclear whether this wage is a median or a mean or some other aggregated measure). There also could be omitted sectors, and we don't know the relative proportion of individuals in each sector. The size of the income gap may be positively correlated with the crime rate as well as the tax revenue and wage variables for high paying sectors like service while being negatively correlated with the wage variables for low paying sectors like manufacturing. As such the size of the income gap is likely to be have a positive coefficient.

3. Police bias. Bias among police officers in certain areas may contribute to the crime rate because of spurious arrests and convictions. This may be difficult to measure directly. We would expect police bias to be positively correlated with the probability of arrests and to a lesser extent the probability of convictions. It would also be positively correlated with the crime rate leading to the coefficient being positive. This implies we have positive bias on the probability of arrests and convictions.

4. Crime rate in neighboring counties. Proximity to other areas where crime is high may have an influence on the crime rate in a particular county due to spillovers of activity. This variable may be correlated with other variables like the probability of convictions and probability of arrests as well as the outcome variable, crime rate. We would as a result expect the coefficient of police bias to be positive.

5. Size of the economy. The size of the economy for each county may be a factor. Explanations could be made for crime rate to be higher or lower in a given county depending on other counties. It would be intersting to see how the crime rate varies with the size of the economy (measured by GDP or similar measure). This would likely be positively correlated with the density and tax per capita variables as well as the wage variables and the crime rate variable. The sign of the coefficient for this variable would be expected to be positive and skewed existing income-related variables upward.
6. Unemployment rate. We have the wage level within each sector, but we don't have the unemployment rate within each county. An unemployed person has a higher propensity to commit a crime than someone who is working. So a higher unemployment rate in a county would increase the crime rate in the county. We expect that this has a positive bias on the coefficients with a positive slope.
7. Family Composition. Having a variable which defines the degree of cohesiveness or divorces will play an important role in the crime rate. We predict that people with a less than healthy childhood have a higher chance of commiting crime than a person who had a normal childhood. The higher the family composition, the lower the crime rate which would imply that we will have a negative coefficient. We expect there to be no correlation between the family composition and the other variables currently in the model. Therefore, it is likely absorbed by the error terms.
8. Poverty Level. In our current model taxpc acts as a proxy for the poverty level, but the challenge with this is that people within or close to the poverty level do not contribute to taxes and there might be outliers with higher income that can skew the data substantially.
9. Repeat Crimes - It would be helpful to understand what percentage of the crimes are repeat crimes, this can help us to understand the importance of judicial system and see if there can be any policy decisions that can be made to reduce crime rate. Repeat crimes variable will probably have a positive bias on the coefficient of crime rate. We don't believe that there is a strong correlation between repeat crimes and the other explanatory variables, so we believe this effect is absorbed by the error term.

Conclusion

Comments on the Modeling Process

Between Specifications 1, 2 and 3, we observed an increase in the adjusted R squared value, implying that the robustness of the model did indeed increase with the addition of more variables. Looking between the second and third specification, all of the variables in the second specification more or less retained their coefficient value, suggesting that the variables in specification 2 are indeed the ones we should be focusing on, however the AIC and BIC scores for the third specification were better than that for the second specification which is an indication that a reasonable amount of the variation is explained by the wage variables that were left out of the second model specification. However, by the Wald test, the third model was not a significantly better specification than the second specification.

There are a number of omitted variables, discussed above, that are suspected to have an impact on the crime rate, or that we suspect are highly correlated with the variables available in the dataset.

Additionally there are confounding factors we must consider. We noticed that crime rate is going up due to police presence, which may seem counterintuitive. However, more police may lead to more reports of crime, so there are a large number of unreported crimes that are potentially being missed here. This leads us to question whether the crmrte variable is really a true representation of how safe a neighborhood is, as, in some areas most crimes can go unreported either due to a lack of trust in law enforcement, or simply because victims do not want to spend the energy to report a crime. We must also consider the possibility that there is a higher police presence in some variables because there is a high crime rate. This would actually suggest that polpc is an effect of crmrate than the other way round. If this is true, then we would actually advise removing the polpc variable from our model specifications.

Also, there are a variety of economic variables that could be highly correlated with another economic parameter that is causal to crmrate. Specifically, we would suggest exploring income gaps and unemployment data in counties, as these have a clearer causal mechanism to crime rate.

Implications for the Political Campaign

In terms of actionable steps for the political campaign, we recommend looking at the outputs of the regression model with a more critical eye. From first glance, it looks like reducing the number of police, arresting people for crimes more often, and forcing people out of counties would be the solution. These are neither advised nor are they ethical in some cases. Instead, we recommend looking at why these explanatory variables are related to crime rate.

Some actionable steps we can recommend are communicating the judicial penalties for severe crimes (as fear of punishment does seem to have an impact on crime rate) and programs to improve the relationship between civilians and police forces. Setting up a rehabilitation programs rather than prison time for less certain crimes might help as stricter prison sentence has no statistical significance on the crime rate. Results from a future analysis that looks at income gaps explicitly would also inform which groups of individuals require a wage increase or better employement opportunities (if any). Additionally, to normalize for density, we recommend that future studies look at crime rate per capita.