Crime Analysis

Pete, Dani, Ken

7/21/2018

Intro

The intention of this analysis is to explore socioeconomic determinants of crime across counties of North Carolina and to build an OLS regression model that informs our political campaign of the policies most likely to improve crime rates among its constituency. Our dataset consists of various measurements and information that lend itself to policy surrounding the following 3 categories: business and labor, policing, social programs. As such, we will focus our research around these broader topics. Specifically, we'd like to understand:

- What, if any, tax policies could be levied to decrease crime rates
- If wages are a strong signal for crime rates, should we consider a minumum wage and or union related policy
- Will expanding police forces help improve crime rates?
- Should police training programs be implemented to improve crime rates?

Initial Data Loading and Cleaning

To ensure the quality and integrity of our analysis, we examine our data set for the following:
1. problematic data types 2. completeness of data for each of our variables 3. duplicate entries 4. known input errors such as probabilities and percentages of total greater than 1

```
#and take a glimpse of the variables and data types
g = glimpse(crime)
```

From our snapshot, we note that prbconv is a factor but should really be a double. Further, pctmin80 is a percent expressed as number instead of in decimal form. All other percentage variables are expressed in decimal form. We adjust our dataset to correct for this.

```
crime = mutate(crime, prbconv = as.numeric(levels(crime$prbconv))[crime$prbconv],
               pctmin80 = pctmin80/100)
```

Checking for the percentage of complete observations in each variable, we note that all observations are missing the same percent of data. We check to ensure that dropping NAs don't drastically change the number of observations from our dataset.

```
# how complete is the data for each of our variables
apply(is.na(crime),2,mean)
```

```
##     county       year     crmrte     prbarr    prbconv    prbpris
## 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567
##     avgsen      polpc    density      taxpc       west    central
## 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567
##      urban   pctmin80       wcon       wtuc       wtrd       wfir
## 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567
##       wser       wmfg       wfed       wsta       wloc        mix
## 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567 0.06185567
##    pctymle
## 0.06185567
```

Since only 0.0618557 of all data is lost by dropping NAs, we will drop the NAs from our dataset instead of imputing values.

```
# removing all NA inputs
crime = na.omit(crime)
```

We also note that 0.010989 of our data are duplicates which we'll drop as well.

```
# removing all NA inputs
crime = distinct(crime)
```

Lastly, we note that there are 10 rows with input errors since probabilities or percent of totals can't be greater than 1. We remove those entry errors.

```
# indexing input errors
indx = crime %>%
        filter(crmrte > 1 | prbarr > 1 | prbconv > 1 | prbpris > 1 |
                 pctmin80 > 1 | pctymle > 1) %>%
          select(county)

# removing all input errors
crime = filter(crime, !county %in% indx$county)
```

The Model Building Process

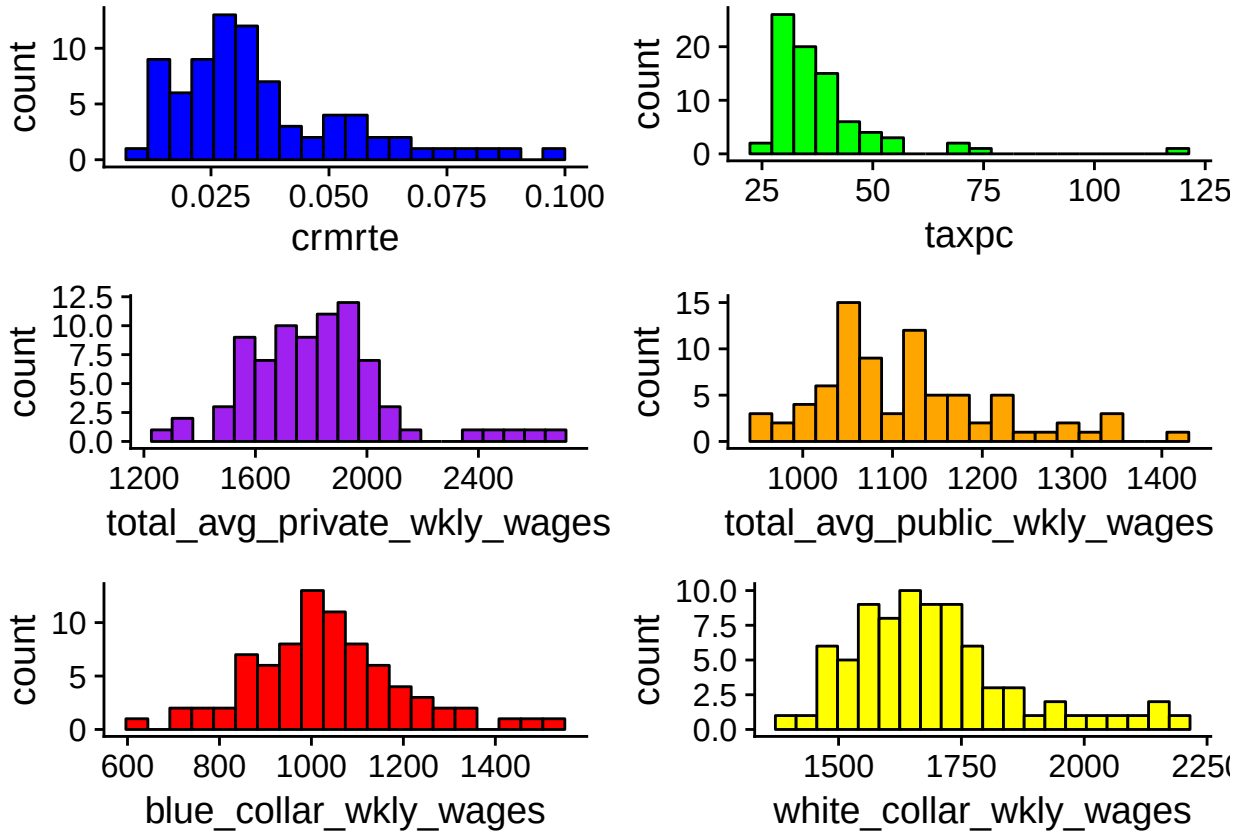
Exploring Viability of Labor and Tax Policies to Reduce Crime

We're given wage data for 9 industry groups in both the private and public sector. Taking voter appeal into consideration, focusing labor policy on industry as a whole is likely to attract broader support than focusing policy on specific industries. So we've decided to aggregate wage data into the following classes: private, public, blue collar, and white collar.

```
# aggregating private, public wages, blue collar and white collar wages
crime = crime %>% rowwise() %>% mutate(total_avg_private_wkly_wages = sum(c(wcon,wtuc,wtrd,wfir,wser,wm
                                       total_avg_public_wkly_wages = sum(c(wfed,wsta,wloc)),
                                       blue_collar_wkly_wages = sum(c(wcon,wtuc,wmfg)),
                                       white_collar_wkly_wages = sum(c(wfir,wser,wfed,wsta,wloc)))

# dropping unneccessary columns
crime = crime %>% select(-c(year,wcon:wloc,west,central,urban))
```

We examine the distributions for each of our variables of interest

```
cr = ggplot(crime,aes(crmrte)) + geom_histogram(color = "black", fill = "blue", bins = 20)
tax = ggplot(crime,aes(taxpc)) + geom_histogram(color = "black", fill = "green", bins = 20)
pri_wage = ggplot(crime, aes(total_avg_private_wkly_wages)) +
  geom_histogram(color = "black", fill = "purple", bins = 20)
pub_wage = ggplot(crime, aes(total_avg_public_wkly_wages)) +
  geom_histogram(color = "black", fill = "orange", bins = 20)
blue_collar = ggplot(crime, aes(blue_collar_wkly_wages)) +
  geom_histogram(color = "black", fill = "red", bins = 20)
white_collar = ggplot(crime, aes(white_collar_wkly_wages)) +
  geom_histogram(color = "black", fill = "yellow", bins = 20)
plot_grid(cr,tax,pri_wage,pub_wage,blue_collar,white_collar, ncol = 2, nrow = 3)
```

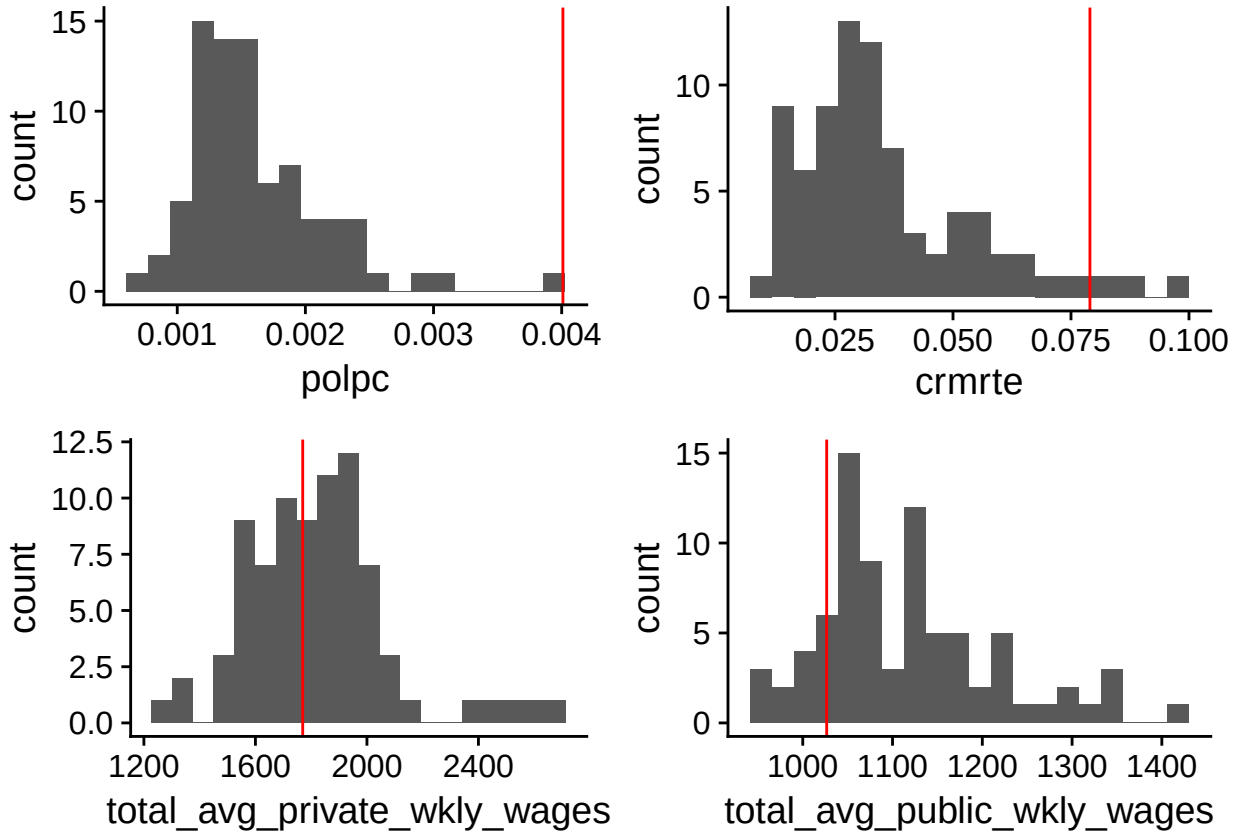


The distributions seem fairly normal, but our dependent variable could benefit from a log transformation. We also note an evident outlier in our tax revenue per capita variable and will investigate that further.

Outlier Observations:

- There appears to be an outlier in tax revenue per capita. We've identified some interesting points associated with this outlier and have indicated this outlier's data points with respect to the other data by a red vertical line

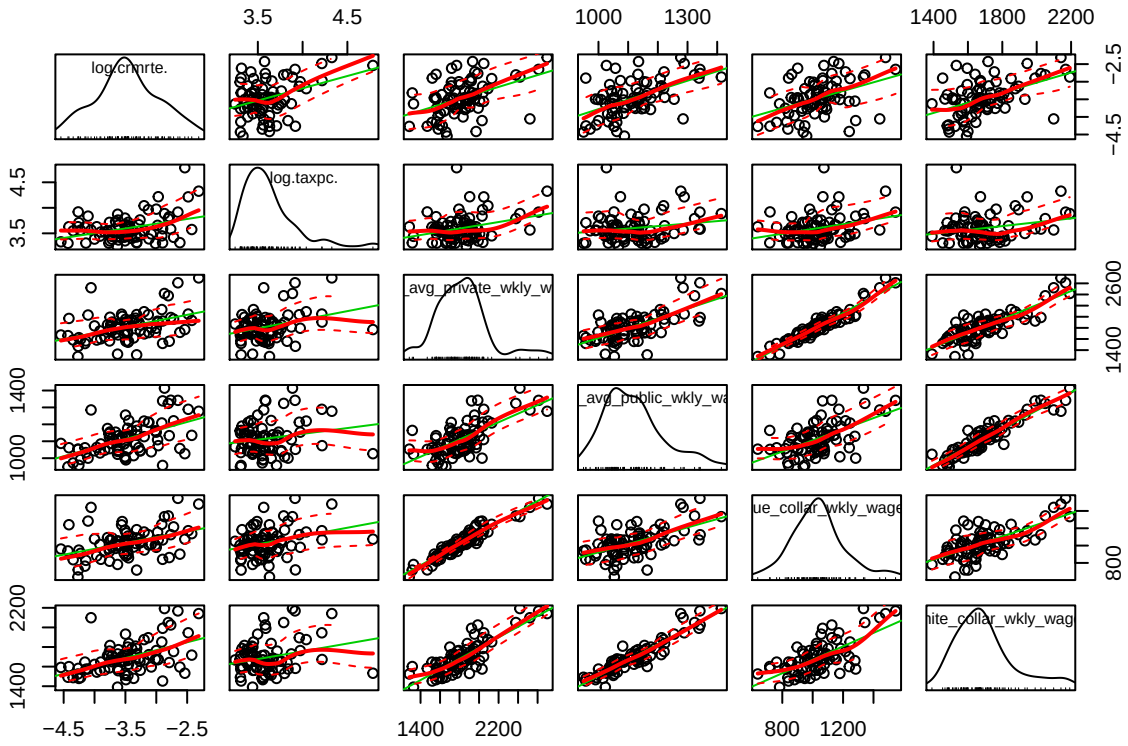
```
#taking a look at the outlier in tax revenue per capita
taxpc_outlier = crime[which(crime$taxpc>100),]

a = ggplot(crime,aes(polpc)) + geom_histogram(bins = 20) +
  geom_vline(xintercept = .00400962, color = "red")
b = ggplot(crime,aes(crmrte)) + geom_histogram(bins = 20) +
  geom_vline(xintercept =  .0790163, color = "red")
c = ggplot(crime,aes(total_avg_private_wkly_wages)) + geom_histogram(bins = 20) +
  geom_vline(xintercept =  1769.737, color = "red")
d = ggplot(crime,aes(total_avg_public_wkly_wages)) + geom_histogram(bins = 20) +
  geom_vline(xintercept =  1026.67, color = "red")
plot_grid(a,b,c,d,ncol = 2, nrow =2)
```



County 55 is substantially higher than all other counties and it also has the 4th highest crime rate despite having the highest police per capita. When investigating avg weekly wages in this county compared to the average wages of all counties, we note that county 55 has lower than avg wages. Low weekly wages and high tax revenue per capita isn't necessarily unexpected as major sources of local government tax revenue also consists of property and sales tax. We don't have measurements on these factors. So it's possible that crimes could be committed in high property valued areas like retirement communities where weekly incomes may be relatively low. This outlier may give an indication of possible omitted variable bias (property tax and median age demographic). It also hints that police per capita may not be a deterrent to crime.

Next, we examine associations between our labor and tax variables with our dependent variable, crime rate.

```
scatterplotMatrix( ~ log(crmrte) + log(taxpc) + total_avg_private_wkly_wages + total_avg_public_wkly_wa
                   blue_collar_wkly_wages + white_collar_wkly_wages , data = crime)
```



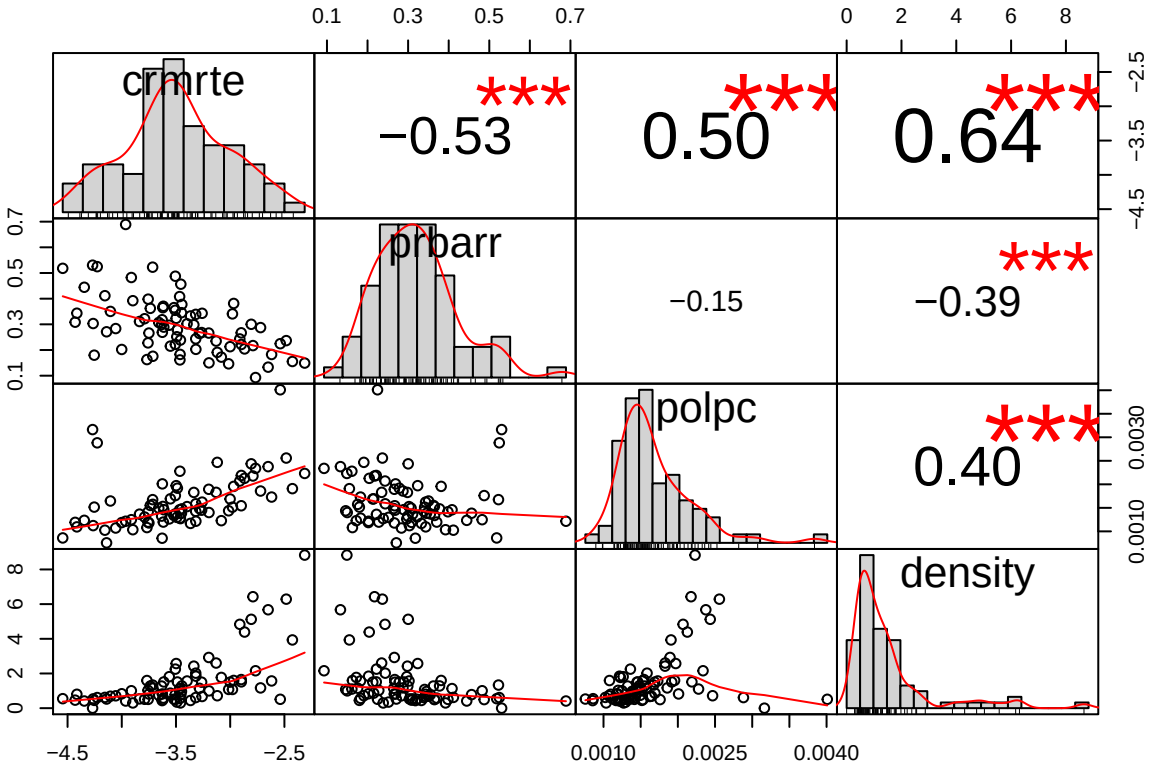
The intial relationships between crime rates, privates wages, and taxes don't show much promise. Without the outlier in tax revenue per capita, there isn't a linear relationship with crime rate. Further, it doesn't seem to make sense that increasing wages would increase crime rates. Not to mention that there isn't much of a polictical angle with that association. There must be some other variables that are influencing these trends. So we will drop wages and taxes as candidates for our model.

Exploring Viability of Policy Relating to Policing Efforts

We're given measurements around arrests, convictions and sentences. However, we believe the conviction and sentencing metrics such as probability of conviction, probability of prison sentence, and avg sentence days are more representative of the judicial system rather than police enforcement. Further, there is a dependency of those metrics on policing. One has to be arrested before they are convicted and sentenced. Since our focus is on policy pertaining to policing practices, we will exclude the conviction and sentencing data from our analysis. The policing variables we have are probability of arrest and police per capita. Both of the variables are influenced by the total number of people in a given county. Generally speaking the more people, the more arrest and crimes we'd expect. As a result, we factor in density to account for denominator sensitivity.

We first examine a correlation matrix between our explanatory variables of interest and our transformed dependent variable to help identify any strong linear associations

```
table1 = crime %>% mutate(crmrte = log(crmrte)) %>% select(crmrte,prbarr,polpc,density)
chart.Correlation(table1, histogram = TRUE, pch=19)
```



We observed that there is a high positive correlation between crmrte and density (0.63 with high significance). There is also strong positive correlation between crime rate and police per capita. This is a little counter-intuitive at first as one might expect crime rates to drop the more police in a given area all else equal. However, the dependencies could be reversed where there are more police in a given area simply because there is more crime. What's particularly interesting is that the probability of arrest doesn't change as police per capita increases. So perhaps there is a policing effacy issue where better police training is required. Lastly, we note that the probability of arrest has a convincing negative correlation with crime rates. Since we're interested in ultimately reducing crime, we will keep prbarr and density as our two explanatory variables for our base model.

Base Model

We establish a base model with two key explanatory variables we identified from our exploratory analysis: prbarr and density

```
model1_data = crime %>% mutate(crmrte = log(crmrte)) %>% select(crmrte,prbarr,density)
model1 = lm(crmrte ~ .,data = model1_data )
coef(model1)
```

```
## (Intercept)      prbarr     density
##  -3.2607934  -1.5301997   0.1646164
```

Our base model accounts for 0.5063822 of the variation in log(crime rate). As it stands now, our model says that a one unit increase in the probability of arrest is associated with a 1.2 percent decrease in the crime rate holding density fixed. And a one unit increase in density is associated with a .19 percent increase in crime rates holding the probability of arrest fixed.

There are a handfule of assumptions that our model makes for the above conclusions. We cover those below.

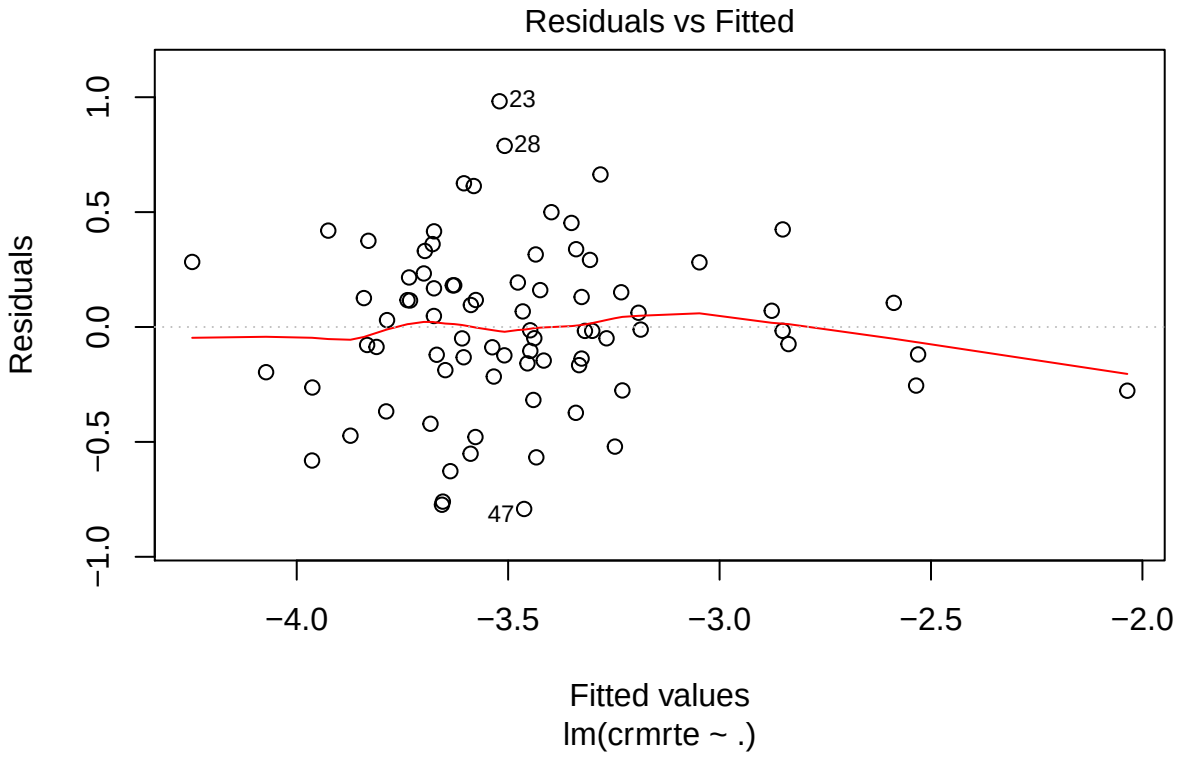
Addressing the 5 assumptions of the OLS model

A residuals vs fitted values plot is very effective at highlighting any violations of the assumptions of the OLS model which are key for the legitimacy of our policy recommendations. We will investigate how the residuals behave with the predicted values from our model, but first we summarize the main assumptions we're checking for.

A Review of the Assumptions in Our Model

1. Linearity in Parameters - The OLS model assumes a linear relationship between the coefficients of our explanatory variables and our predictor variable. If the population's relationship among these parameters is non-linear, then essentially any of the conclusions we draw from our analysis are highly skeptical and our prediction accuracy would be unreliable.
2. Random Sample - We need our data to be representative of the population we're trying to model against. Our OLS model assumes that the underlying data are independent and identically distributed. We're given our data was drawn from four organizations: FBI, North Carolina Department of Correction, Census Data, North Carolina Employment Security Commission. While we don't have insight into the reporting practices of the above organizations, we do know they are established, reputable organizations. So, we will assume their data is trust worthy. What initially seemed suspect is that we're given a sample of data from a selection of odd counties only. However, upon further research we noted that FIPS codes for NC counties only use odd numbers. With that said, we belive we can assume this sample to be random.
3. No Perfect Collinearity: We need to be cognizant of exact linear combinations between explanatory variables as these will skew the effect of individual explanatory variables, holding all other variables fixed, and convolute our policy recommendations.
4. Zero Conditional Mean - We want all other possible factors aside from our explanatory variables to be independent of our explanatory variables. This implies there are no lurking variables influencing our data and that we do not have omitted variable bias. If the expected value of our error term given the different explanatory variables is zero, then we have stronger faith in our model fitting the true population model.
5. Homoskedasticity - the standard errors, confidence intervals and hypothesis tests associated with the OLS model depend on a constant variance of error terms

```
# plotting residuals vs fitted values
plot(model1, which = 1)
```



We can see that the residuals seem evenly distributed on on both sides of zero and that there are no major fanning or curve effects. What is suspect is the cluttering around specific predicted values of our dependent variable, (log(crmrte)), and the non-uniform distribution of points which may be indicative of omitted variable bias.

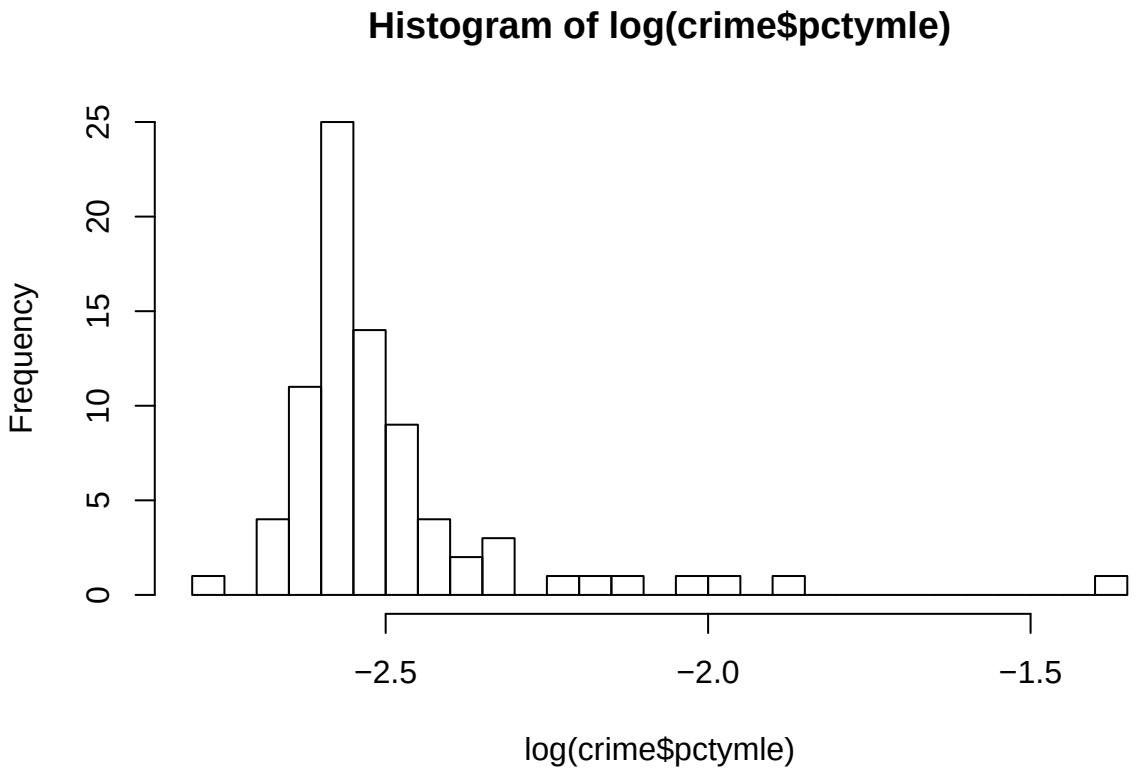
A note on possible omiited variable bias

Our density variable represents the entire population per square mile in a given county. However we do not include any information on whether the majority of crimes were committed by any particular persons or groups in the given population which could also be a lurking factor in influencing crime. Since there is one variable in our dataset pctymle (percent of young males), we explore adding that as another key variable in our model (Model II).

Second Model

We now add another variable, pctymle, in our regression model and observe that transforming, pctymle, into log shows a clear normal distribution.

```
#plotting variable transformation and outlier influence
hist(log(crime$pctymle),breaks = 30)
```

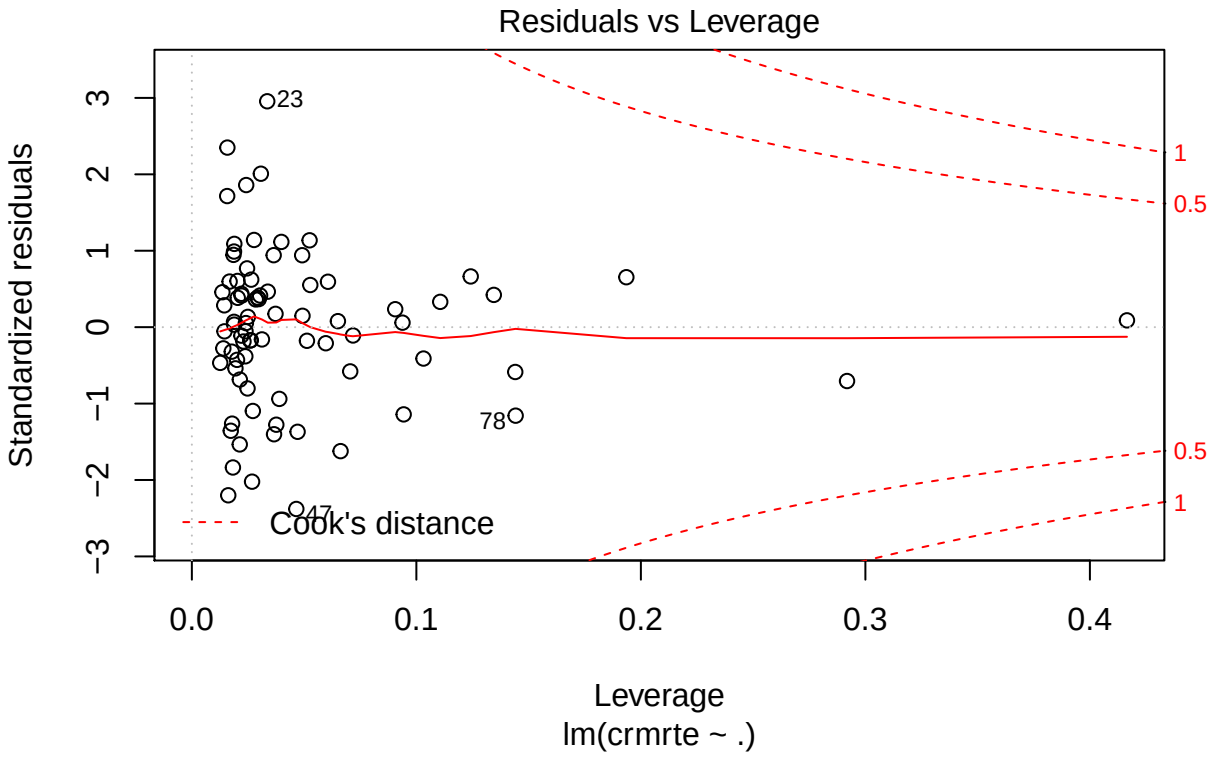


We also take note of the apparent outlier in the transformed pctymle variable and will check on this outlier's influece of the regression and whether or not regression results would be altered if we excluded it. We can do this via glancing at the residuals vs leverage plot.

```
#computing model 2
model2_data = crime %>% mutate(crmrte = log(crmrte),pctymle = log(pctymle)) %>%
  select(crmrte,prbarr,density,pctymle)
model2 = lm(crmrte ~ .,data = model2_data )

coef(model2)
```

```
## (Intercept)      prbarr     density     pctymle
##  -2.2183037  -1.3592315   0.1613698   0.4364196
```

```
# residuals vs leverage plot
plot(model2,which = 5)
```

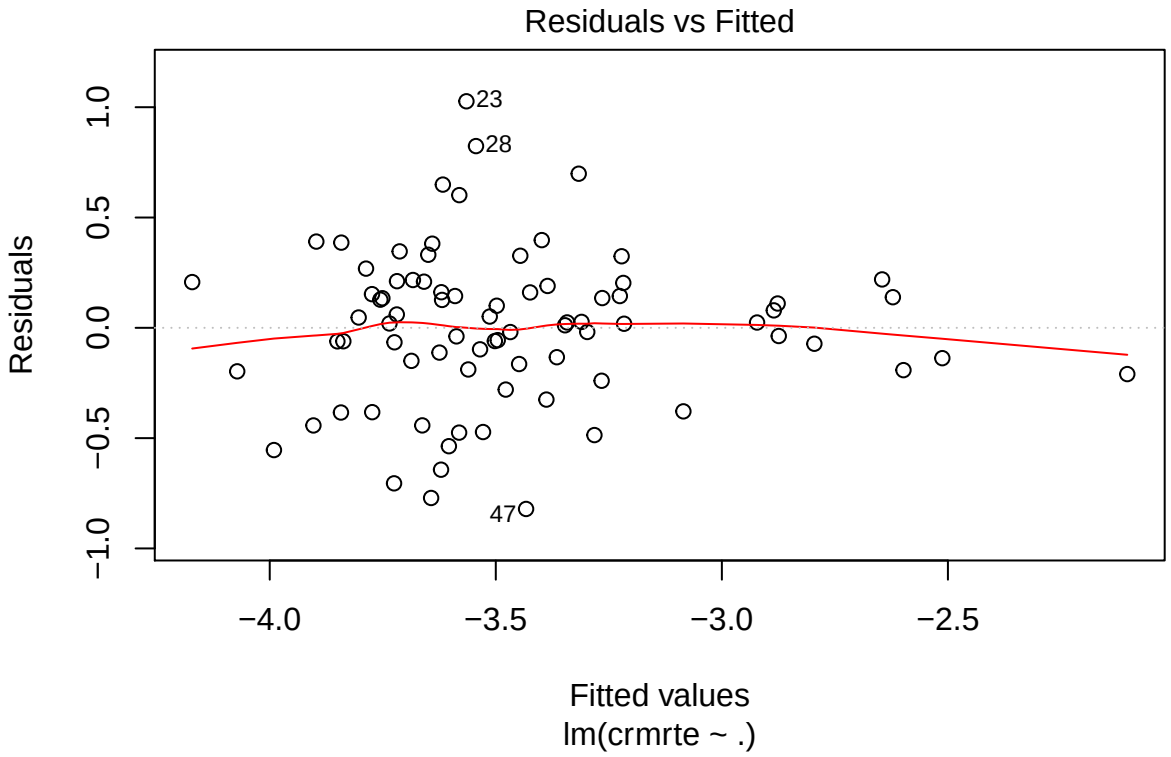


Since the observation falls within a cook's distance less than 1/2, it doesn't have enough influence to change the regression results. Further, our second model has an AIC score of 66.5020919 which is lower than our first model's AIC score of 69.0961185 so we'll keep both the variable and outlier in our second model.

Our second model accounts for 0.5339299 of the variation in log(crime rate) which captures slightly more variation than the first model. As it stands now, our model says that a one percent increase in the percenate of males is associated with a .48 percent increase in the crime rate holding density and probability of arrest fixed. We also note that adding the pctymle variable didn't change the relationship between the dependent variable and other independent variables.

We again check for possible violations of the classical linear model assumptions stated earlier with a residuals vs fitted values plot.

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



This looks very similar to the plot resulting from model 1 which again may be evidence of additional ommitted variable bias.

Third Model

We will now include all possible covariates of our dataset into a 3rd model as a test against our variable selection in model 1 and model 2

```
#computing model 3
model3_data = crime %>% mutate(crmrte = log(crmrte),pctymle = log(pctymle))
model3 = lm(crmrte ~ .,data = model3_data )
coef(model3)
```

```
##                   (Intercept)                        county
##                 -2.781929e+00                  5.557028e-05
##                        prbarr                        prbconv
##                 -1.628919e+00                 -2.639331e-01
##                        prbpris                        avgsen
##                  3.210747e-02                 -2.161382e-02
##                         polpc                       density
##                  2.829552e+02                  1.087185e-01
##                         taxpc                       pctmin80
##                  3.120358e-03                  1.181193e+00
##                            mix                       pctymle
##                 -9.614738e-01                  3.760884e-01
## total_avg_private_wkly_wages  total_avg_public_wkly_wages
##                  2.729798e-03                  4.878160e-03
##       blue_collar_wkly_wages      white_collar_wkly_wages
##                 -2.321272e-03                 -4.665806e-03
```

```r
summary(model3)$r.squared
```

```
## [1] 0.8061629
```

Our 3rd model accounts for 0.8061629 of the variability in crime rates, but we'd expect it to capture more variability because it contains all the data. We also note that it has a lower AIC score at 20.316651 than our prior two models, but this doesn't necessarily mean this is the best model fit for the true population model. We also point out that many of the variables have high p-values so we would fail to reject the null hypothesis that the coefficient is equal to zero (ie. no effect). In other words many of these coefficients with p-values greater than .2 like prbpris, avgsen, mix, etc could very well have no effect on crime rate.

```r
summary(model3)$coefficients[,4]
```

```
##                   (Intercept)                        county
##                  2.073288e-05                  9.151717e-01
##                        prbarr                        prbconv
##                  1.141191e-05                  2.099119e-01
##                        prbpris                        avgsen
##                  9.352013e-01                  1.015611e-01
##                         polpc                       density
##                  1.176799e-03                  2.530716e-04
##                         taxpc                      pctmin80
##                  3.061414e-01                  5.906225e-08
##                           mix                       pctymle
##                  4.167125e-02                  3.611393e-02
## total_avg_private_wkly_wages  total_avg_public_wkly_wages
##                  3.617203e-02                  6.371685e-03
##       blue_collar_wkly_wages      white_collar_wkly_wages
##                  8.089775e-02                  5.825023e-03
```

Lastly, model 3 is extremely difficult to interpret and conveying meaninful policy recommendations based off it is nearly impossible.

The Regression Table

```r
stargazer(model1, model2, model3, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting Crime Rates in NC",
          keep.stat = c("rsq", "n"),
          omit.table.layout = "n") # Omit more output related to errors
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 22, 2018 - 16:14:51

The Omitted Variables Discussion

Conclusion

Table 1: Linear Models Predicting Crime Rates in NC

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
county			0.0001
prbarr	−1.530	−1.359	−1.629
prbconv			−0.264
prbpris			0.032
avgsen			−0.022
polpc			282.955
density	0.165	0.161	0.109
taxpc			0.003
pctmin80			1.181
mix			−0.961
pctymle		0.436	0.376
total_avg_private_wkly_wages			0.003
total_avg_public_wkly_wages			0.005
blue_collar_wkly_wages			−0.002
white_collar_wkly_wages			−0.005
Constant	−3.261	−2.218	−2.782
Observations	80	80	80
R^2	0.506	0.534	0.806