~~~
---
title: "Crime Analysis"
author: "Pete, Dani, Ken"
date: "8/03/2018"
output:
  pdf_document: default
  html_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
~~~

### Introduction

This analysis serves to explore socioeconomic determinants of crime across counties of North Carolina and to build an OLS regression model that informs our political campaign containing the policies most likely to reduce crime rates among its constituency. Our dataset consists of various measurements and information that lends itself to building policy surrounding the following 3 categories: business and labor, policing, social programs. As such, we will focus our research around these broader topics. The areas of our research that highlight convincing levers with which to impact crime rates identify opportunities for policy interventions to reduce crime, and a campaign run on these policies will help our candidate secure crime troubled regions of his constituency. 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
* If crime rates are influenced by certain age groups or gender, should we introduce outreach programs in our campaign

Our analysis will first explore the data to ensure appropriate quality for analysis, perform univariate and multivariate analysis on key indicators, and finallly explore proposed regression models in building levels of complexity. Given our intention to highlight policy interventions to reduce crime, we will explore regressions of the variable crmrte on variables we believe key to predicting crime rates.

### Initial Data Loading and Cleaning

To ensure the quality and integrity of our analysis, we examine our data set for the following problematic criteria:

1. the existence of any problematic data types, to be adjusted
2. the completeness of data for each of our variables, checking for n/a

~~~
values and handling them appropriately
3. any duplicate entries which we will remove
4. any known input errors that violate the logic of the variables

```{r, message=FALSE, warning=FALSE, echo=FALSE}
# loading dependencies for analysis
library(car)
library(dplyr)
library(ggplot2)
library(cowplot)
library(PerformanceAnalytics)
library(stargazer)
library(lmtest)
library(sandwich)

#read in the original data
crime = read.csv("crime_v2.csv")
```

```{r, eval=FALSE}
#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.

```{r,message=FALSE,warning=FALSE}
crime = mutate(crime, prbconv = as.numeric(levels(crime$prbconv))
[crime$prbconv],
               pctmin80 = pctmin80/100)
```

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

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

Because only `r 1 - nrow(na.omit(crime))/nrow(crime)` of all data is
lost by dropping NAs, we will drop the NAs from our dataset instead of
imputing values.

```{r}
# removing all NA inputs
~~~

~~~~
crime = na.omit(crime)
```

We also note that `r 1 - nrow(distinct(crime))/nrow(crime) ` of our data
are duplicates which we'll drop as well.

```{r}
# removing all NA inputs
crime = distinct(crime)
```

Lastly, we note that there are `r nrow(crime %>% filter(crmrte > 1 |
prbarr > 1 | prbconv > 1 | prbpris > 1 | pctmin80 > 1 | pctymle > 1))`
rows with "input" errors because probabilities and percent of totals
can't be greater than 1. While some may argue these are ratios such as
arrests to offenses and it may be possible to have more than 1 arrest to
an offense, the intention of this variable is a proxy to probability
which should adhere to the laws of probability. Because there are few of
these seeming violations, we remove those entries as well.

```{r}
# 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

\begin{center}
Exploring Viability of Labor and Tax Policies to Reduce Crime
\end{center}

The dataset includes wage data for 9 industry groups in both the private
and public sectors. 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. Because of this motivation
for broader voter appeal, we have decided to aggregate wage data into
the following classes: private, public, blue collar, and white collar.

We use a sum aggregation rather than an average aggregation to minimize
concealment of potential outliers.


```{r}
# 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,wmfg)),
                                       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

```{r}
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 do not immediately raise concern, 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.

##### Univariate Analysis of Dependent Variable (crmrte):

From the previous histograms we noted skewness in our dependent
variable, crmrte. By implementing a log transformation on crmrte we now
have an approximately normal distribution in our regressand which allows
us to continue with linear modeling.

```{r}
cr_log = ggplot(crime,aes(log(crmrte))) + geom_histogram(color =
"black", fill = "blue", bins = 20)
plot_grid(cr_log)
```


##### 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.

~~~
```{r}
#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 its peer group 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 consist of property and sales tax. We don't have measurements on these factors, and so we cannot rule out the possibility that crime could be committed in high property valued areas such as retirement communities where weekly income 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.

~~~
```{r}
scatterplotMatrix( ~ log(crmrte) + log(taxpc) +
total_avg_private_wkly_wages + total_avg_public_wkly_wages +
                     blue_collar_wkly_wages + white_collar_wkly_wages ,
data = crime)
```
~~~

The intial relationships between crime rates, privates wages, and taxes don't lead to a strong conclusion. Without the outlier in tax revenue per capita, there no longer a linear relationship with crime rate. Further, it doesn't seem to make sense that increasing wages would increase crime rates, on top of the fact 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.

\begin{center}
Exploring Viability of Policy Relating to Policing Efforts
\end{center}

The dataset includes 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 efficacy of the judicial system rather than police enforcement. Further, there is a dependency of those metrics on policing. One has to be arrested before one is subsequently convicted and sentenced. Because 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, areas with higher populations higher rates of arrest and crime. 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

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

We observe 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 with higher police presence in a given area, all else being 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.

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

\begin{center}
Interpretation of Base Model
\end{center}

Our base model accounts for `r summary(model1)$r.squared` of the variation in log(crime rate). As it stands now, our model demonstrates that a one unit increase in the probability of arrest is associated with a 1.2 percent decrease in the crime rate, holding population 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 handful of assumptions that our model makes for the above conclusions. We cover those below.

\begin{center}
Addressing the 6 assumptions of the CLM model
\end{center}

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

6. Normality of Errors - the population error is normally distributed

~~~
```{r}
# plotting residuals vs fitted values
plot(model1, which = 1)

```
~~~

We can see that the expected value, or average, of the residuals approximates zero up until greater fitted values where there's a downward curve which may be indicative of omitted variable bias. What is also suspect is the cluttering around specific predicted values of our dependent variable, (log(crmrte)), and the non-uniform distribution of points which will likely impact the statistical significance of our results

We can standardize the residuals and observe a scale-location plot of the fitted values for better insight into potential heteroskedasticity (ie trends in the residuals)

~~~
```{r}
# plotting residuals vs fitted values
plot(model1, which = 3)
```
~~~

It becomes clear from the scale-location plot that we have heteroskedasticity in our variables. As non-constant variance of residuals is a common problem in practice, we can accomodated for this by using robust standar erros in our model.

```{r}
#using robust standard errors
coeftest(model1, vcov = vcovHC)
```

\begin{center}
Interpretation of Base Model with Robust Standard Errors
\end{center}

While our interpretation of the slope coefficients haven't changed, we are more confident that our results are not due to variations in random sampling. The probability that we’d see these coefficient values given that the true slope relationship is zero, all else equal, is very tiny: .0005 and 4.361e-09 for prbarr and density respectively. Further, we can see that our t-values are beyond 3 in any direction which gives us high confidence in the statistical significance of our model. Still our current model only accounts for ~50% of the variation in our predicted values so with respect to model fit, we need improvement. Given the heteroskedasticity in our variables and best practices, we will continue to use robust standard error models when evaluating our coefficients

#### 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. Because there is one variable in our dataset pctymle (percent of young males), we consider 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(pctymle) shows a clear normal distribution. We chose percent male to explore whether or not there is practical evidence that developing outreach programs for young males could be a viable means of reducing crime and thus resonate better with constituents.

```{r}
#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.

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

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

Since the observation falls within a cook's distance less than 1/2, it does not have enough influence to change the regression results.

```
\begin{center}
Interpretation of Second Model (with Robust Standard Errors)
\end{center}
```

~~~
```{r}
coeftest(model2, vcov = vcovHC)
```
~~~

Again the p-values and t-values give us high confidence in the statistical significance of model 2. What's interesting is that our first two variables' coefficients haven't changed much even after partialling out our new variable,log(pctymle). This appears to imply that there isn't much covariance between log(pctymle) and our original variables. The coefficient of 0.436 for pctymle indicates that for every percent increase in pctymle there is a .43% increase in crime rate, holding density and probability of arrest fixed. While this is stastically significant, it doesn't seem practically signficant since a small percent increase of a percent (crime rate) is very small.

Further, our second model has an AIC score of `r AIC(model2)` which is lower than our first model's AIC score of `r AIC(model1)` so we will keep both the variable and outlier in our second model.

Our second model accounts for `r summary(model2)$r.squared` of the variation in log(crime rate) which captures slightly more variation than the first model. However this leaves ~47% of the variability unaccounted for which isn't a good fit.

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

~~~
```{r}
plot(model2,which = 1)
```
~~~

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

### Third Model

Finally, we include all possible covariates of our dataset into a 3rd model as a test against our variable selections in models 1 and 2.

```{r}
#computing model 3
model3_data = crime %>% mutate(crmrte = log(crmrte),pctymle =
log(pctymle))
model3 = lm(crmrte ~ .,data = model3_data )
coeftest(model3, vcov = vcovHC)
summary(model3)$r.squared

```

Our 3rd model accounts for `r summary(model3)$r.squared` 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 `r AIC(model3)` 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.

Lastly, because of the sheer number and variety of variables included in model 3, it is much more difficult to interpret and conveying meaninful policy recommendations based off it is nearly impossible.

### The Regression Table

```{r, results='asis'}
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
```

### Omitted Variables and Directional Impacts

Omitted variable bias presents a real challenge in the analysis included in this paper, as crime rates are determined by many compounding factors, few of which are truly independent of each other and many of which are difficult to quantify across a relevant peer group. Because we have identified that the second model is the most useful for the purposes of this analysis, the omitted variables exploration will focus on those included in that model.

Model 2 evaluates how the probability of arrest, density of population, and percent of young males in the population are associated with changes in crime rate. One of the challenges with this data set is that the measurements of these variables exclude a lot of detail that could signficantly strengthen the analysis, namely with the pctymle variable. This measures a very specific demographic of the population, presumably based on the assumption that young men are the highest contributors to crime rate. Of course, this assumption may not be entirely representative of the reality, and there are likely other characteristics of a population that affect crime rates in a region such as employment rates.

We believe that there is a negative correlation between employment rate and prbarr as well as employment rate and the percent of males associated with crimes. We also believe it makes sense that there is a small positive correlation between employment rate and desnity since jobs are concentrated in areas. Since our coefficent of prbarr is negative and the correlation between the two is negative, we'd have positive bias in our prbarr (controlling for partialling out effects of the other two variables). We'd also have positive bias in our density variable since there is a positive correlation between density and employment rates. With that said, our model may have positive bias. In other words, the OLS estimates of our regression equation are on average too large.

The second variable that likely represents a notable omitted variable bias is the probability of arrest. It is unclear what specific factors motivate a high probability of arrest. For example, a particularly efficient police department may in fact result in a higher probability of arrest. Additionally, this variable doesn't indicate the seriousness of the crime, aggregating the probabilities of arrest for lighter and much more serious crimes smooths out any micro trends that exist within this variable.

Finally, we acknowledge that our outcome variable, crime rate, also smooths out likely useful nuances in the data. There are different types of crimes and degrees of criminality that would be useful to analyze, particularly as the policy approaches that could be informed by such analysis would vary based on precisely what type of crime was being considered.

### Verifying the Classical Linear Model Assumptions

In order for our proposed model to be usable, we must be able to test and satisfy the following assumptions. Because we propose using Model 2, our validation of these assumptions are tested using that model.

1. Linear Population Model

This assumption requires that the depent variable, $crmrte$, be a linear combination of the explanatory variables and the error term. Because we have not placed any constraint on the error term, we can assume that this criteria is satisfied.

2. Random Sampling

Without having additional background information on how the supplied data was collectd, we cannot know for certain that the sample was in fact collected randomly. However, we interpret the methodology of the data collection to be sound.

3. No Perfect Colinearity

By successfully performing our regression analysis in r, we can be sure this assumption is not violated.

4. Zero Conditional Mean

To test if this assumption is satisfied, we can evaluate the results of the residual versus fitted diagnostic plots.

```{r}
plot(model2, which= 1)
```

The concave down parabolic shape of the residuals indicates that there is a deviation from the zero conditoinal mean. However, our data set has a sufficiently large sample that the bias is not significant enough to invalidate the model. This can be seen visually by the fact that, while it is in fact parabolic, it is only slightly so.

5. Homoskedasticity

The previous residuals versus fitted plot does not immediately allert us to heteroskedasticity. However, we can use the scale-location plot to verify that this assumption has not been violated, as evident in the near flatness of the line.

```{r}
plot(model2, which=3)
```

6. Normality of Errors
Finally, we use the normal q-q plot to test for the normality of errors.

```{r}
plot(model2, which=2)
```

This visual indicates minimal skew, and with n= 97 observations in our data set, we can assume that our estimators will have a normal sampling distribution.

### Conclusion

The analysis in this study explored a variety of data that may be contributing factors to increased crime rates in North Carolina. While the analysis is limited based on the available data, we were nonetheless able to determine that the probability of arrest, the density of the area and the percent of the population that are young males can all be observed to correlate with crime rate.

Based on these observations, we recommend a campaign based on policies that address these populations. By deploying crime reducing resources through a targeted approach, in densely populated communities and particularly where police forces have a greater difficulty with their probability of arrest, the politician in question can instate programs design to outreach young men in the area.

However, informing the campaign strategy based on this analysis alone would overlook significant contributors to crime rates as well as a more sophisticated qualitative understanding of the distinct needs of the very differing counties in North Carolina.