Lab 3 Reducing Crime

Team 5

11/16/2019

1.0 Introduction

In this report, we will be discussing the determinants of crime in North Carolina in order to generate policy suggestions that are applicable to local government to lower crime rate.

2.0 Data Cleansing

Data is drawn from data sources provided by the campaign.

```
raw_data <- read.csv('../data/raw/crime_v2.csv')
```

After sanity checks, we noticed a few anomalous values that are likely to be entering errors. We made the following major changes in order to proceed with a cleaned-up dataset.

1. Remove 6 NA values because they don't contain any values to the analysis.
2. Remove one of the duplicated values for County 193.
3. Top coding all the probability variable with the max value = 1, simply because any probability that's greater than 1 doesn't make sense.
4. For one particular variable `wser`, there's one value seems to be mis-entered by one digit, making it from 217.71 to 2177.07. We think it makes sense to convert it back to 217.71.

We've also renamed the variables for better understanding. For more details about how the data has been changed, see `./src/make_features.R`. The processed data is stored as `./data/processed/crime_v2.csv`.And, for supporting information that verifies the distributions of data, please see the EDA file in `./notebooks/EDA.md`.

```
crime_data <- read.csv('../data/processed/crime_v2.csv')
```

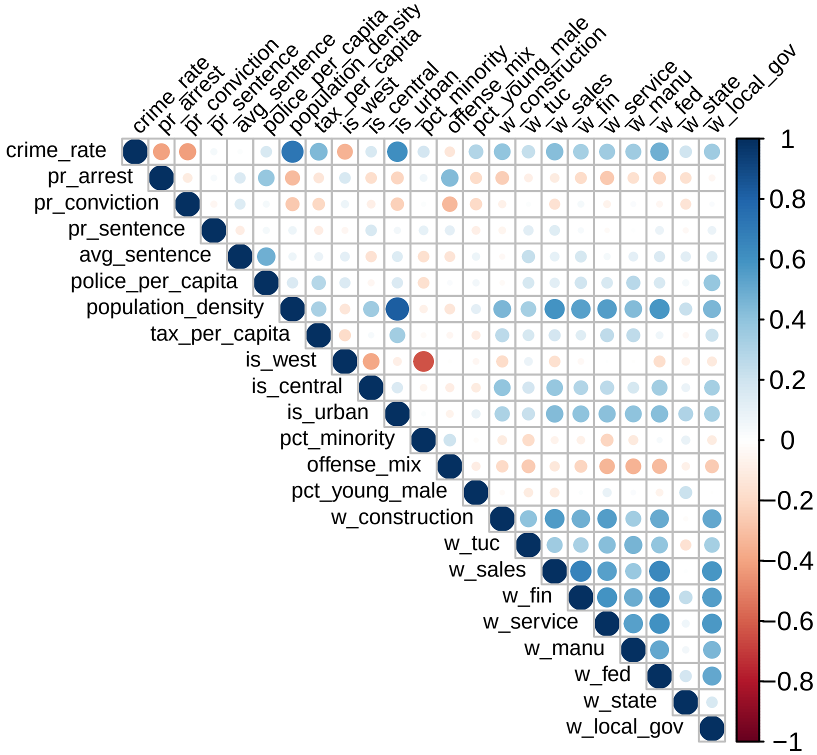
After cleansing, the data set now has 90 observations.

3.0 Correlation Plot

Based on the data at hand, we produced the correlation plot below to help visualize the relationship among all the variables -

```
crime_cols_cor <- cor(crime_data[c('crime_rate',
                                   'pr_arrest', 'pr_conviction', 'pr_sentence', 'avg_sentence',
                                   'police_per_capita','population_density', 'tax_per_capita',
                                   'is_west', 'is_central', 'is_urban',
                                   'pct_minority', 'offense_mix', 'pct_young_male',
                                   'w_construction', 'w_tuc', 'w_sales', 'w_fin',
                                   'w_service', 'w_manu', 'w_fed',
                                   'w_state', 'w_local_gov')])
corrplot(crime_cols_cor, type = "upper", tl.col = "black", tl.srt = 45, tl.cex = 0.7,
         mar = c(1, 1, 2, 1), title = 'Correlation between numeric variables in Crime Data')
```

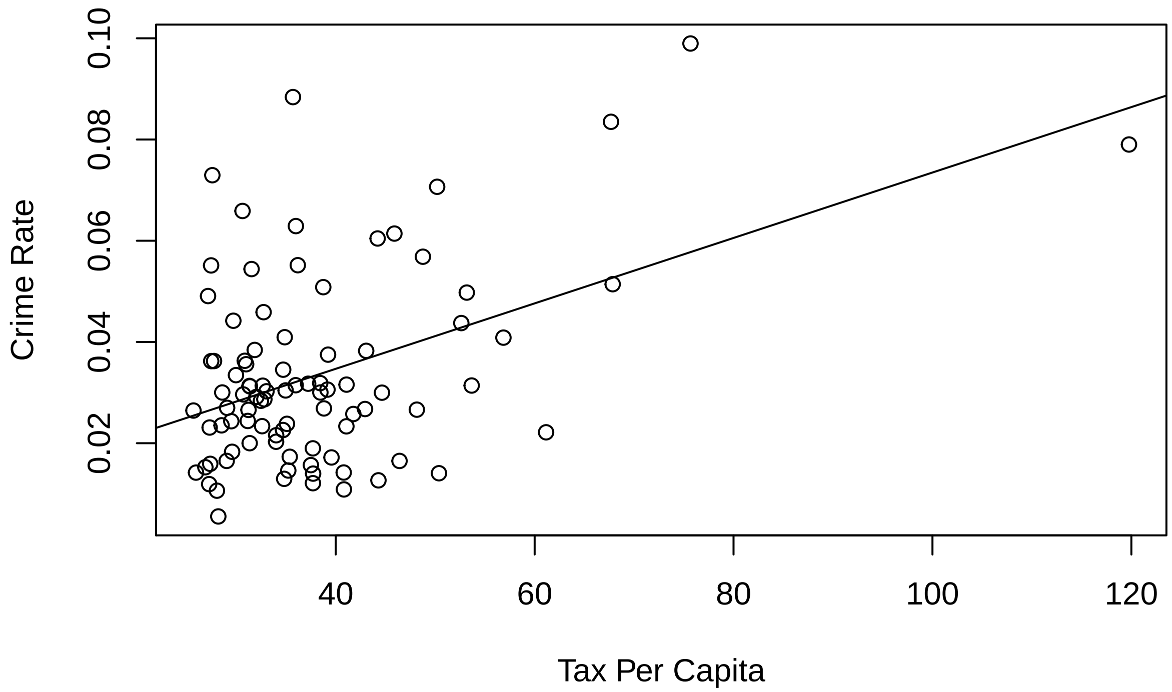
Correlation between numeric variables in Crime Data



We developed our model based on the correlation plot shown above. Many of the variables that show a correlation to crime rate, such as `pr_arrest`, `pr_conviction`, and `police_per_capita` can be construed as causally related variable to `crime_rate`. In other words, as a county experiences high crime rate, they are likely to enact tough on crime policies that result in a greater police presence and a greater probability of arrest and conviction.

We hypothesize that income inequality is a highly predictive metric of crime rate. As this variable is not available in our data set, we proxy its effect through the variable `tax_per_capita`. We believe this will be a good proxy because communities which pay more tax per individual have greater wealth. In the United States this wealth is generally concentrated amongst a small percentage of the overall population, accentuating income inequality. As these counties are all in the same state, our metrics will not be skewed by different state tax policies across our data.

```r
plot(crime_data$tax_per_capita, crime_data$crime_rate,
     xlab="Tax Per Capita", ylab="Crime Rate")
abline(lm(crime_data$crime_rate ~ crime_data$tax_per_capita))
```



The plot above tells us that there's a positive correlation between `Tax Per Capita` and `Crime Rate`, despite a few outliers. With this correlation scatter plot, we decide to use `Tax Per Capita` as our main explanatory variable to build the following models.

4.0 Model Building

Model 1: A model with only explanatory variables of key interest

Based on our hypothesis, we estimate a simple linear regression, explaining `crime_rate` with `tax_per_capita`.

```
model1 <- lm(crime_rate ~ tax_per_capita, data=crime_data)
stargazer(model1, title = 'Model 1: Base Model', type = 'text',
          omit.stat = c('ser', 'F'), header=FALSE, label='urban-year')
```

```
##
## Model 1: Base Model
## ===========================================
##                     Dependent variable:
##                 ---------------------------
##                         crime_rate
## -------------------------------------------
## tax_per_capita           0.001***
##                          (0.0001)
##
## Constant                   0.009
##                           (0.006)
##
## -------------------------------------------
## Observations                90
## R2                         0.201
## Adjusted R2                0.192
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

As you can see in Table 1, reported above, `tax_per_capita` has positive coefficient, meaning that if we increase tax_per_capita by $100, out model predicts the crime rate will increase by 0.065

Model 2: A model that includes key explanatory variables and important covariates

Now, we want to include some covariates that will increase the accuracy of the results without introducing substantial bias. A key covariate to include in our model is `population_density`, as it is highly correlated to our dependent variable and also correlated to our independent explanatory variable, `tax_per_capita`. Due to the the positive correlation is has to both variables, we predict that omitting it has exerted a positive bias and adding it to our model will bring our explanatory variable closer to 0.

```r
model2 <- lm(crime_rate ~ tax_per_capita + population_density,
             data=crime_data)
stargazer(model1, model2, title = 'Model 1 vs 2: Important covariates', type = 'text',
  omit.stat = c('ser', 'F'), header=FALSE, label='urban-year')
```

```
##
## Model 1 vs 2: Important covariates
## ===============================================
##                        Dependent variable:
##                    ----------------------------
##                             crime_rate
##                         (1)            (2)
## -----------------------------------------------
## tax_per_capita        0.001***      0.0003***
##                       (0.0001)      (0.0001)
##
## population_density                  0.008***
##                                      (0.001)
##
## Constant               0.009         0.009**
##                       (0.006)        (0.004)
##
## -----------------------------------------------
## Observations             90             90
## R2                     0.201          0.582
## Adjusted R2            0.192          0.573
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Population density indeed reflects a positive coefficient in our model, not surprising given the large positive correlation we witnessed to crime rate in our EDA.

The coefficient of `tax_per_capita` in this model means that if we increase the tax_per_capita by $100, the crime rate can increase by 0.035. We expect this coefficient in model 2 to be less than the coefficient in model 1 because of the positive effect that omitting `population_density` posed on `tax_per_capita`.

The AIC between model 1 and model 2 below also indicates that model 2(AIC=-530.64) is a better fit than model 1(AIC=-474.3), as a lower AIC represents a better fit.

Model 3: Models that include other covariates

In our third model we add an additionaly covariate that reflected strong correlation to `crime_rate`- `pct_young_male`. We believe this variable deserves inclusion in our model due to the statistics showing that this demographic has a greater likelihood to commit crimes.

```
model3 <- lm(crime_rate ~ tax_per_capita + population_density + pct_young_male,
             data=crime_data)
stargazer(model1, model2, model3, title = 'Model 1 vs 2 vs 3: Improving accuracy',
          type = 'text', omit.stat = c('ser', 'F'), header=FALSE, label='urban-year')
```

```
##
## Model 1 vs 2 vs 3: Improving accuracy
## ================================================
##                         Dependent variable:
##                    -----------------------------
##                             crime_rate
##                       (1)       (2)       (3)
## ------------------------------------------------
## tax_per_capita     0.001***  0.0003*** 0.0004***
##                    (0.0001)  (0.0001)  (0.0001)
##
## population_density           0.008***  0.008***
##                               (0.001)   (0.001)
##
## pct_young_male                         0.197***
##                                         (0.053)
##
## Constant             0.009    0.009**   -0.009
##                     (0.006)   (0.004)   (0.006)
##
## ------------------------------------------------
## Observations          90        90        90
## R2                   0.201     0.582     0.640
## Adjusted R2          0.192     0.573     0.628
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

In this case, model 3 has the lowest AIC score at -542.12, while model 2 has -530.64 and model 1 has -474.3.

In model 3, the coefficients of `tax_per_capita` means that if we increase the tax per capita by $100, our model predicts a decrease in the crime rate by 0.04, while holding other covariates constant. Comparing this to our coefficient from model 2 of 0.035 reveals that the inclusion of the `pct_young_male` variable in our model has actually increased the effect of `tax_per_capita` on crime rate.

5.0 Omitted Variables

To determine the omitted variables we first researched factors that fundamentally cause people to commit crime. We believe people commit crime due to many reasons. Some of those reasons are due to income inequality, poverty rate, homelessness, hunger, unemployment, low education levels, alcohol, drug, opioid abuse, emotional wellness, broken families, unwanted pregnancies and hormones.

We then distilled the above factors into variables that are observable and measurable. * Income Inequality - Wage Distribution for Each County * Poverty Rate - % of Population living below Poverty Level * Homelessness - Number of Families without Home or Shelter * Hunger - Number of Families Having 1-2 Meals per Day * Unemployment - Unemployment Rate * Education - Number of People unable to Read and Write * Alcohol and Drug Abuse - Amount of Alcohol or Opioid Consumption * Emotional Wellness - Number of Unwanted Pregnancies * Broken Families - Number of Children in Orphanage * Hormones

We then looked at the data to see if we could use it as a proxy for our analysis.

- For Income Inequality we used Tax Per Capita
- For Hormones we use Percentage of Male Between 18-24

We then analyzed the following omitted variables -

- Education level - Negative to crime rate, positive to tax per capita; well-educated population tend to have higher income and commit less crime. Omitting education level creates a negative bias, and it could be a fairly large bias considering the importance of education.
- Alcohol and Drug abuse - Positive to crime rate, positive to percentage of young male; alcohol and drug abuse is more likely to happen among younger population, and it leads to more crimes. Omitting the alcohol and drug abuse level creates a positive bias.
- Unemployment Rate - Positive to crime rate, negative to tax per capita; the higher the unemployment rate, the lower average income becomes, and unemployment can lead to committing crimes in order to make a living. Omitting the unemployment rate creates a negative bias.
- Homelessness rate - positive to crime rate, negative to tax per capita; same as unemployment rate, homelessness rate has a two way effect towards high unemployment rate. Similarly, omitting the homelessnes rate pose a negative bias.
- Amount of local businesses/job opportunities - Negative to crime rate, positive to tax per capita; the more local businesses/job opportunities are, the lower unemployment rate will be, and hence lower crimer rate and higher tax per capita. Omitting the vibrantness of local business can create a negative bias.

6.0 Conclusions

Based on our analysis above, we think that closing the gap between income inequality can effectively decrease crime rate. However, the limitation in our model is that `tax_per_capita` is a proxy for income inequality, and we should think of a better proxy to model income inequality so that we can better understand the relationship here.