

Lab 3

David Hou, Scott Hungerfield, Irene Seo

March 20, 2018

Introduction

The purpose of this study is to provide information for political campaign in North Carolina. Specifically, we want to determine what variables contribute to crime rate and help the campaign propose policy suggestions to local governments. To accomplish this, we were given crime data from several North Carolina counties along with other variables. We will run ordinary least square regressions to help determine which of these are the best predictors of crime.

Data Cleaning

First we need to clean the data. In the raw data, we notice that the last 6 rows are empty. The integer columns are probably more useful to us as factors. The prbconv is coded as a factor, so we turn it into a numeric.

We also notice that prbarr and prbconv have values that are greater than 1, which does not make much sense because they are probability variables. We assume that these values were coded incorrectly and filter those out.

As a minor change, we divide pctmin80 by 100, so that it matches the formatting of pctmle. Both variables are percentages and we've arbitrarily chosen to represent them as a number between 0 and 1 rather than 0 to 100.

```
raw = as_tibble(read.csv('crime_v2.csv'))
t = raw %>%
  filter(!is.na(county)) %>%
  mutate(prbconv = as.numeric(as.character(prbconv))) %>%
  mutate(pctmin80 = pctmin80 / 100) %>%
  mutate_if(is.integer, as.factor) %>%
  filter(prbarr < 1 & prbconv < 1)
levels(t$west) = c('East', 'West')
t$west = relevel(t$west, 'West') # Put West first so it appears on the left on facet plots
levels(t$central) = c('Outer', 'Central')
levels(t$urban) = c('Non-urban', 'Urban')
```

We also do not see an advantage to analyzing each wage individually by the industry. Thus, we create a new column that is the sum of all the wage columns.

```
t = t %>% mutate(wage = wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc)
```

Here is a summary of the data.

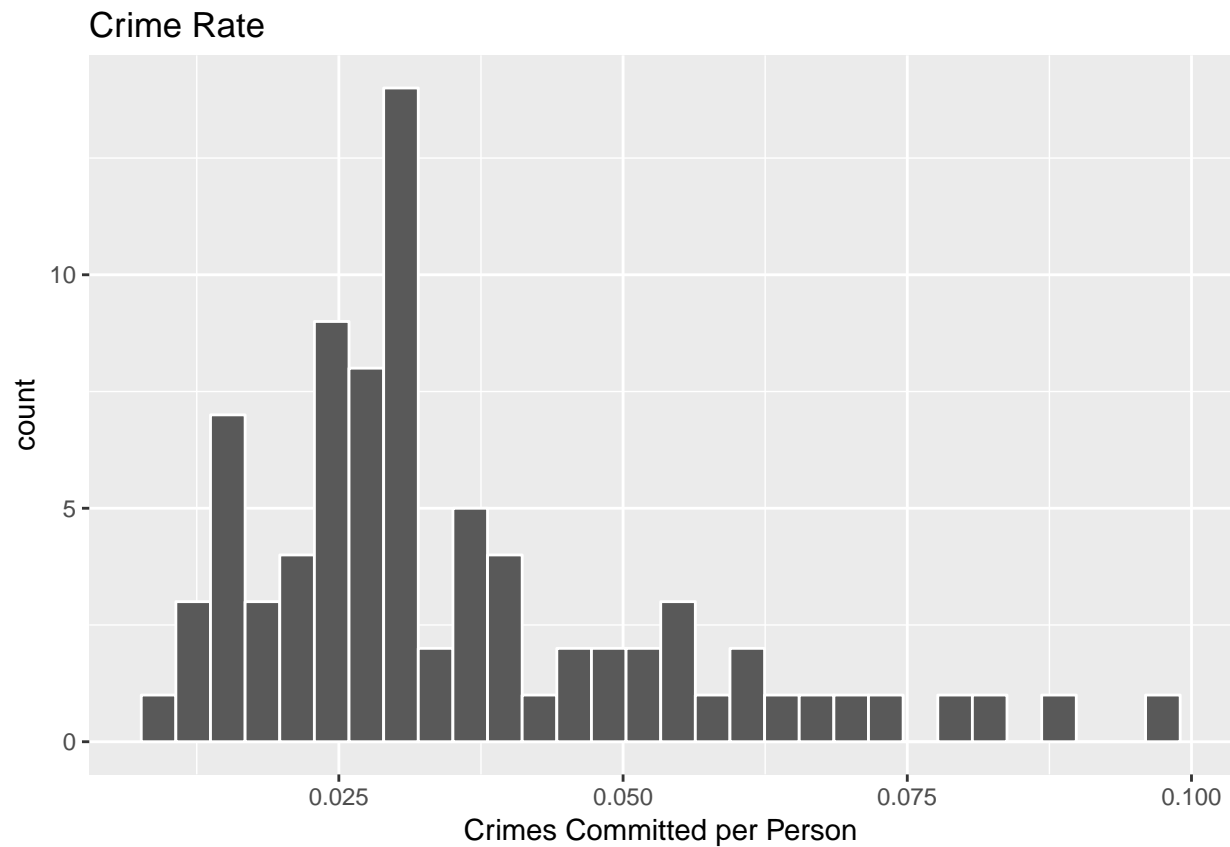
```
stargazer(data.frame(t), type = 'text')
```

```
##
## =====
## Statistic N      Mean      St. Dev.      Min      Max
## -----
## crmrte      81    0.035    0.019    0.011    0.099
```

```
## prbarr      81    0.297    0.109    0.093    0.689
## prbconv     81    0.448    0.172    0.068    0.973
## prbpris     81    0.412    0.078    0.150    0.600
## avgscen     81    9.362    2.372    5.450   17.410
## polpc       81    0.002    0.001    0.001    0.004
## density     81    1.508    1.580    0.00002   8.828
## taxpc       81   38.042   13.267   25.693  119.761
## pctmin80    81    0.258    0.168    0.015    0.619
## wcon        81  287.879   48.018  193.643  436.767
## wtuc        81  410.875   76.697  187.617  595.372
## wtrd        81  213.146   34.339  154.209  354.676
## wfir        81  322.574   50.684  234.522  509.466
## wser        81  255.201   44.775  133.043  391.308
## wmfgr       81  335.661   85.691  157.410  646.850
## wfed        81  445.202   61.039  326.100  597.950
## wsta        81  359.539   42.698  267.780  499.590
## wloc        81  312.081   28.345  239.170  388.090
## mix         81    0.136    0.082    0.051    0.465
## pctymle     81    0.085    0.024    0.064    0.249
## wage        81 2,942.159 331.384 2,338.455 3,975.223
## -----
```

Examining Key Variables of Interest

```
qplot(t$crmrte, col = I('white')) +
  labs(title = 'Crime Rate', x = 'Crimes Committed per Person')
```



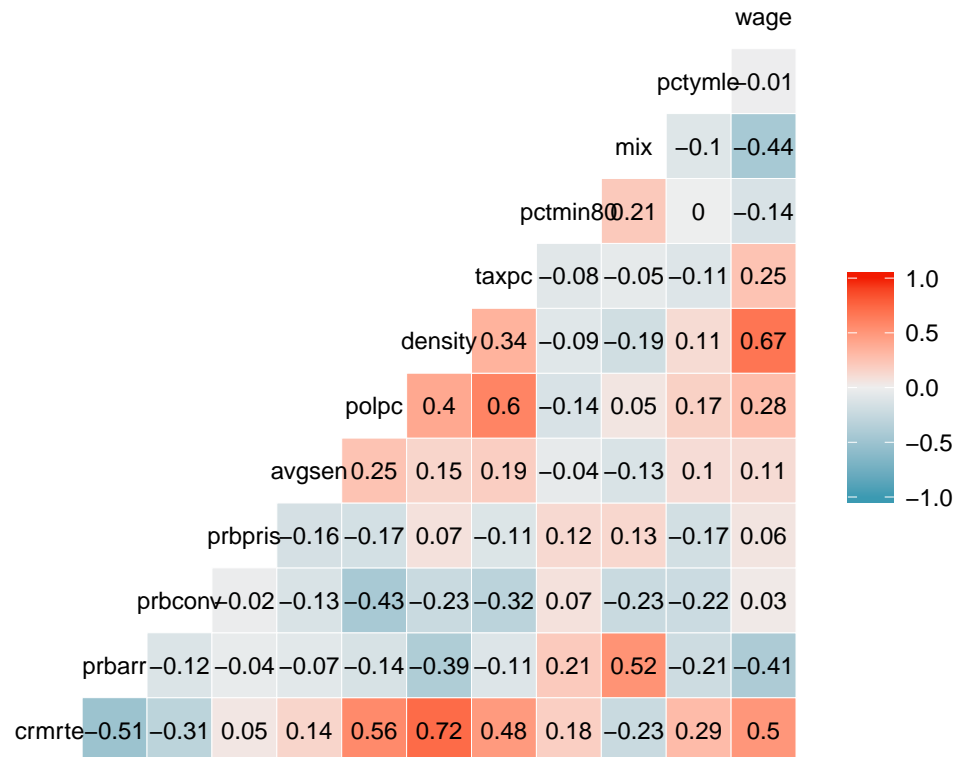
```
summary(t$crmrte)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02337 0.03043 0.03536 0.04374 0.09897
```

We see that the main variable of interest, crime rate, has some positive skew, but does not seem to have a very exotic distribution. To determine which variables are of interest to us when predicting crime rate, we look at the correlation matrix among the variables.

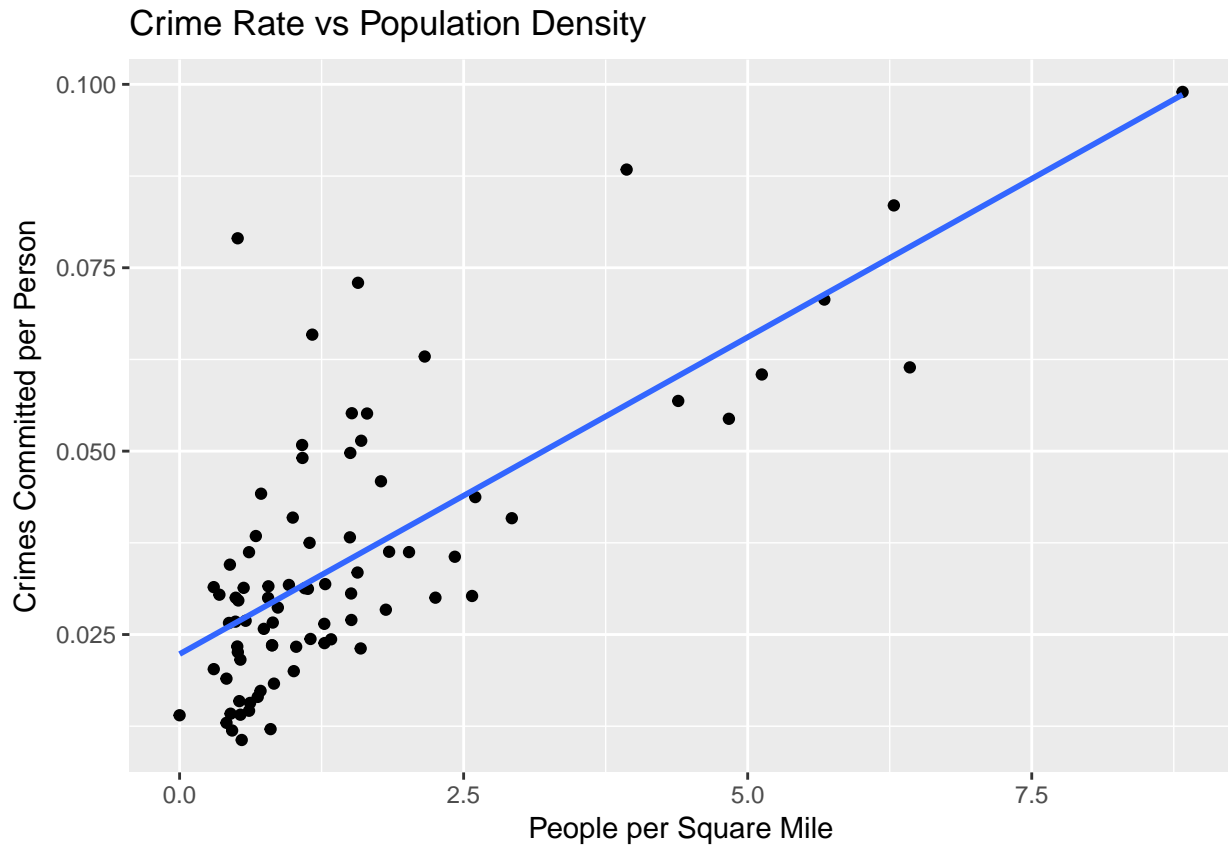
```
t2 = t %>% select(crmrte, prbarr, prbconv, prbpris, avgsen, polpc, density, taxpc, pctmin80, mix, pctym)
ggcorr(t2, label = TRUE, label_round = 2, label_size = 3, size = 3) + ggtitle('Correlation Matrix')
```

Correlation Matrix



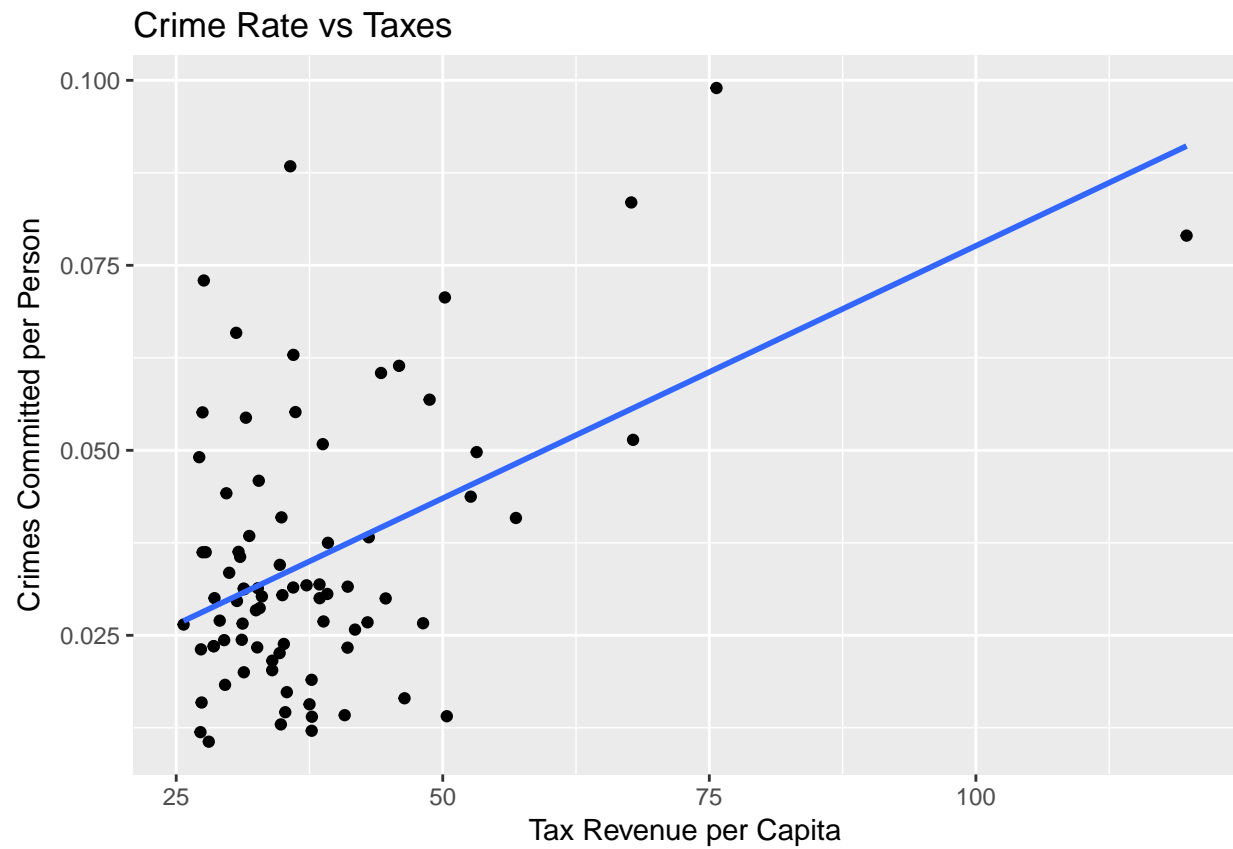
From the correlation matrix, we see that population density stands out as being highly correlated with crime rate ($r = 0.72$). This variable looks like a good candidate as a causal predictor for crime rate. One explanation could be that as more people move into an area, the increased number of interactions give opportunity for more crime.

```
qplot(t$density, t$crmrte) +
  labs(title = 'Crime Rate vs Population Density', x = 'People per Square Mile',
        y = 'Crimes Committed per Person') +
  geom_smooth(method = 'lm', se = FALSE)
```



The other two variables with moderately positive correlation are tax per capita ($r = 0.48$) and wages ($r = 0.5$). It is interesting to note that taxes and wages are not very correlated with themselves ($r = 0.25$). This finding is surprising, as one would expect that wages and taxes would go up very closely with each other. Also note that population density is weakly correlated with taxes ($r = 0.34$) and moderately correlated with wages ($r = 0.67$). We believe that taxes and wages are not directly causing higher crime rates but could be good indirect indicators.

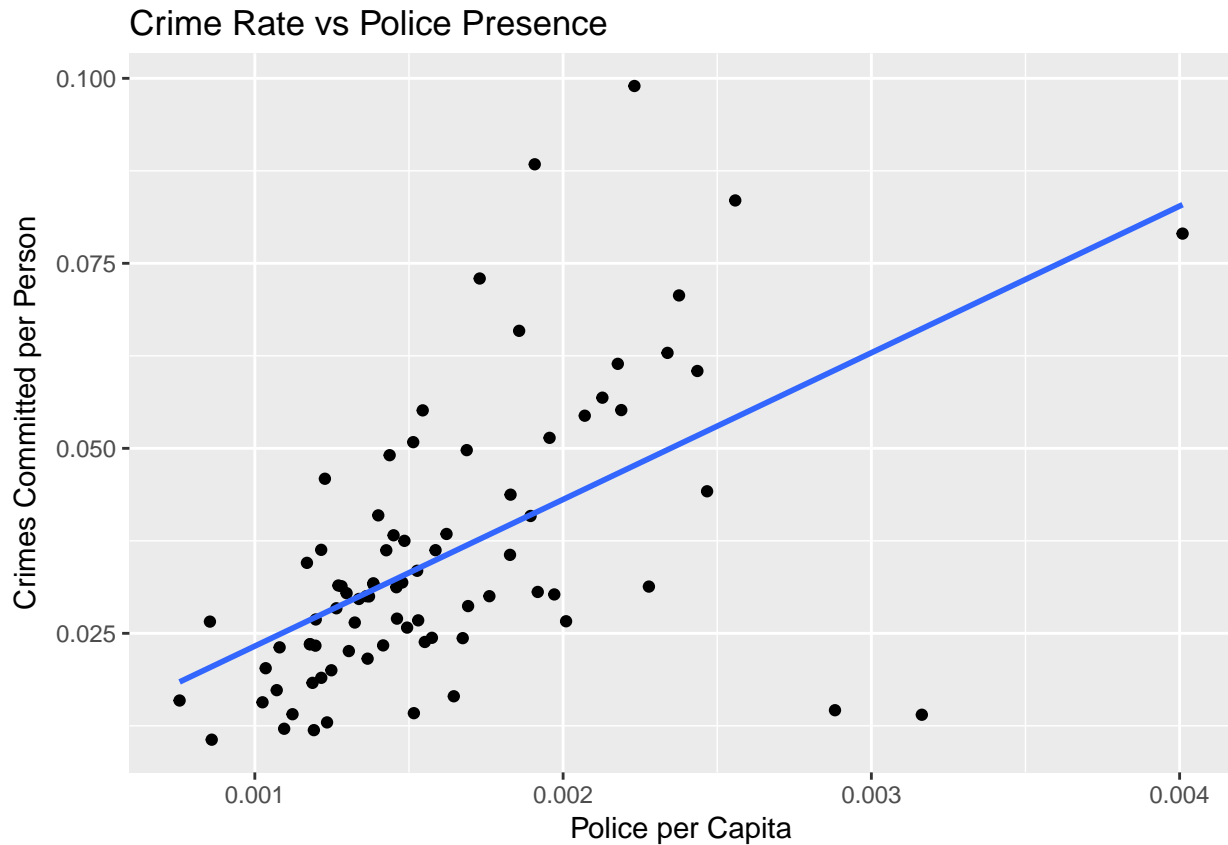
```
qplot(t$taxpc, t$crmrte) +
  labs(title = 'Crime Rate vs Taxes', x = 'Tax Revenue per Capita', y = 'Crimes Committed per Person') +
  geom_smooth(method = 'lm', se = FALSE)
```





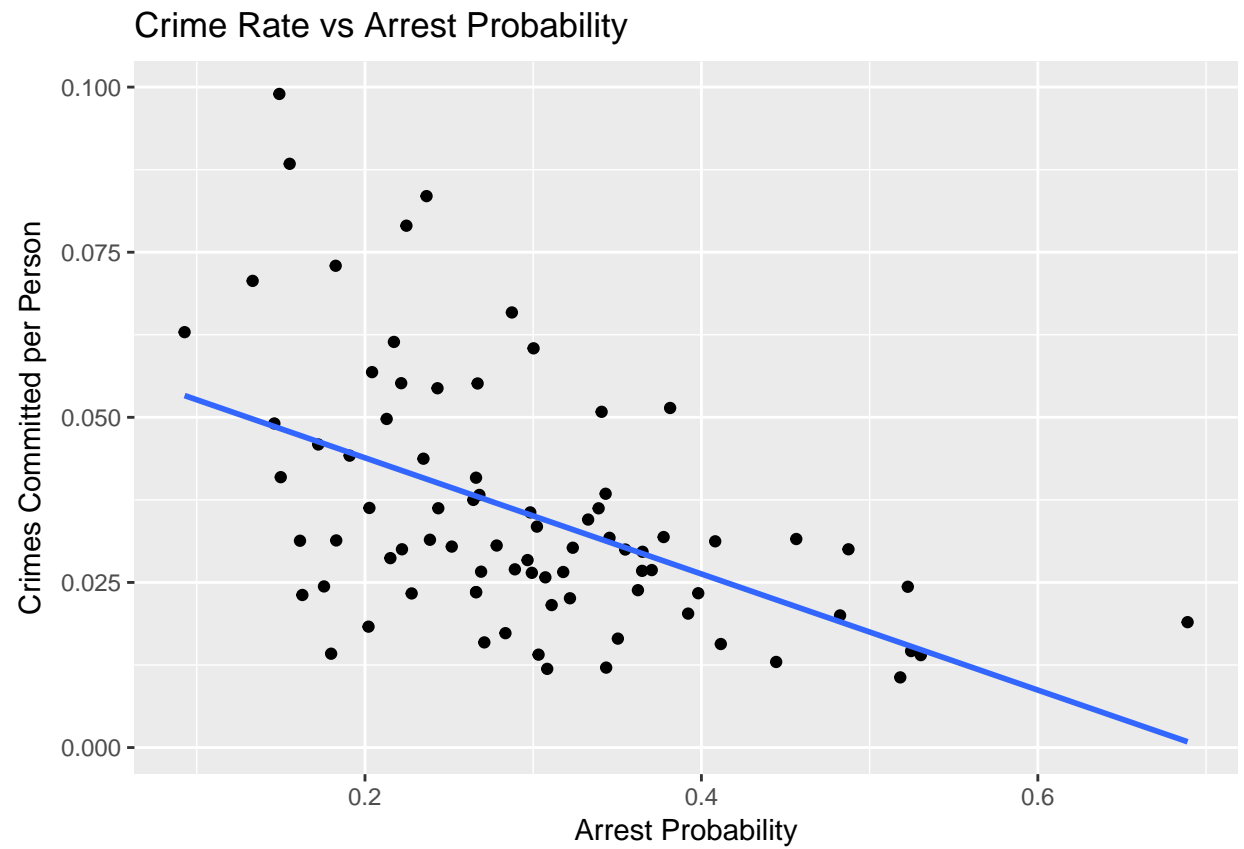
Interestingly, the relationship between police per capita and crime rate is positive and moderately large ($r = 0.56$). This means that either increasing police presence makes crime rate worse or that crime is causing an increase in police presence rather than vice versa. The latter explanation seems much more logical.

```
qplot(t$polpc, t$crmrtc) +  
  labs(title = 'Crime Rate vs Police Presence', x = 'Police per Capita',  
        y = 'Crimes Committed per Person') +  
  geom_smooth(method = 'lm', se = FALSE)
```

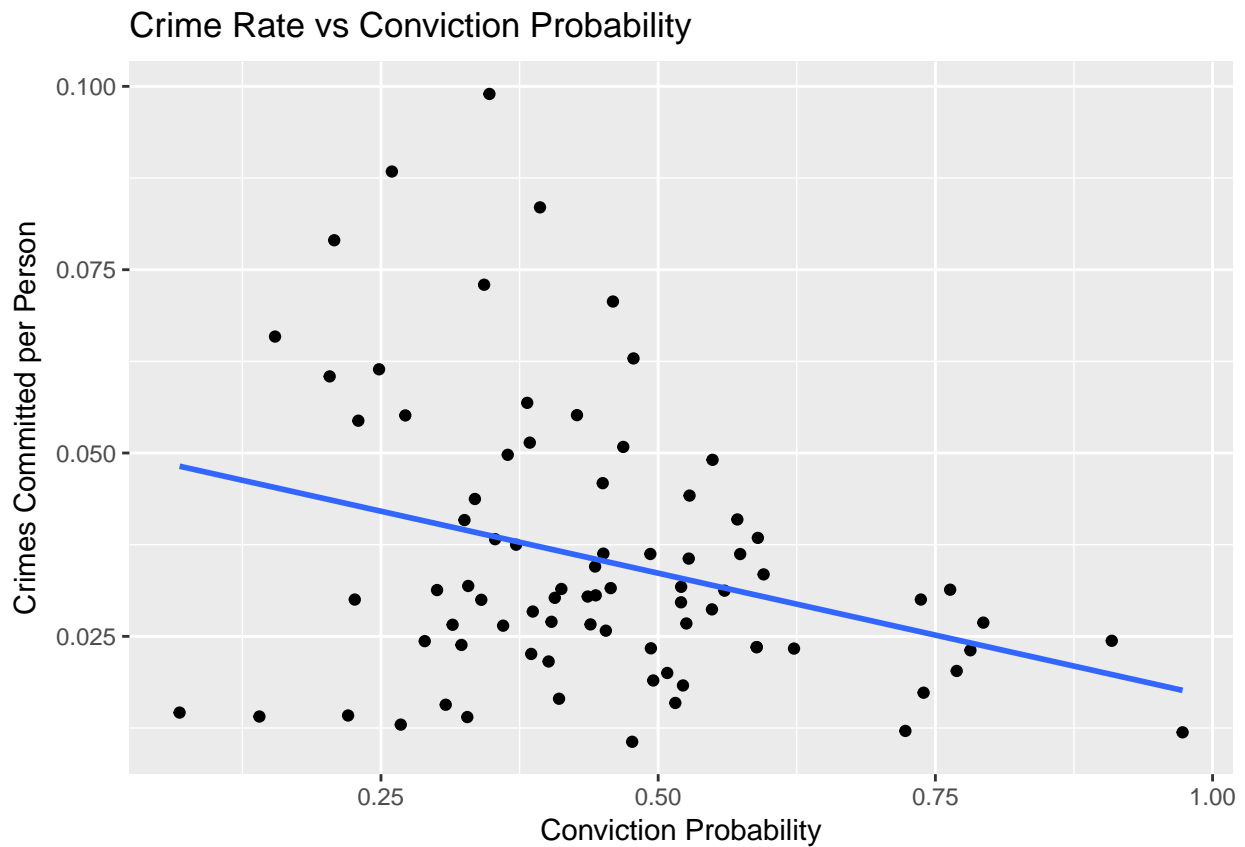


Of the three “certainty of punishment” variables, it looks like arrest probability has a moderate effect ($r = -0.51$) and conviction probability has a weak effect ($r = -0.31$), but probability of prison sentence has almost no effect ($r = 0.05$). It is important to note that these three probabilities seem uncorrelated with one another, so we will include multiple ones in our regression without fear of multicollinearity. The “severity of punishment” variable, average prison sentence length, does not seem to be correlated with crime rate ($r = 0.14$).

```
qplot(t$prbarr, t$crmrte) +
  labs(title = 'Crime Rate vs Arrest Probability', x = 'Arrest Probability',
        y = 'Crimes Committed per Person') +
  geom_smooth(method = 'lm', se = FALSE)
```

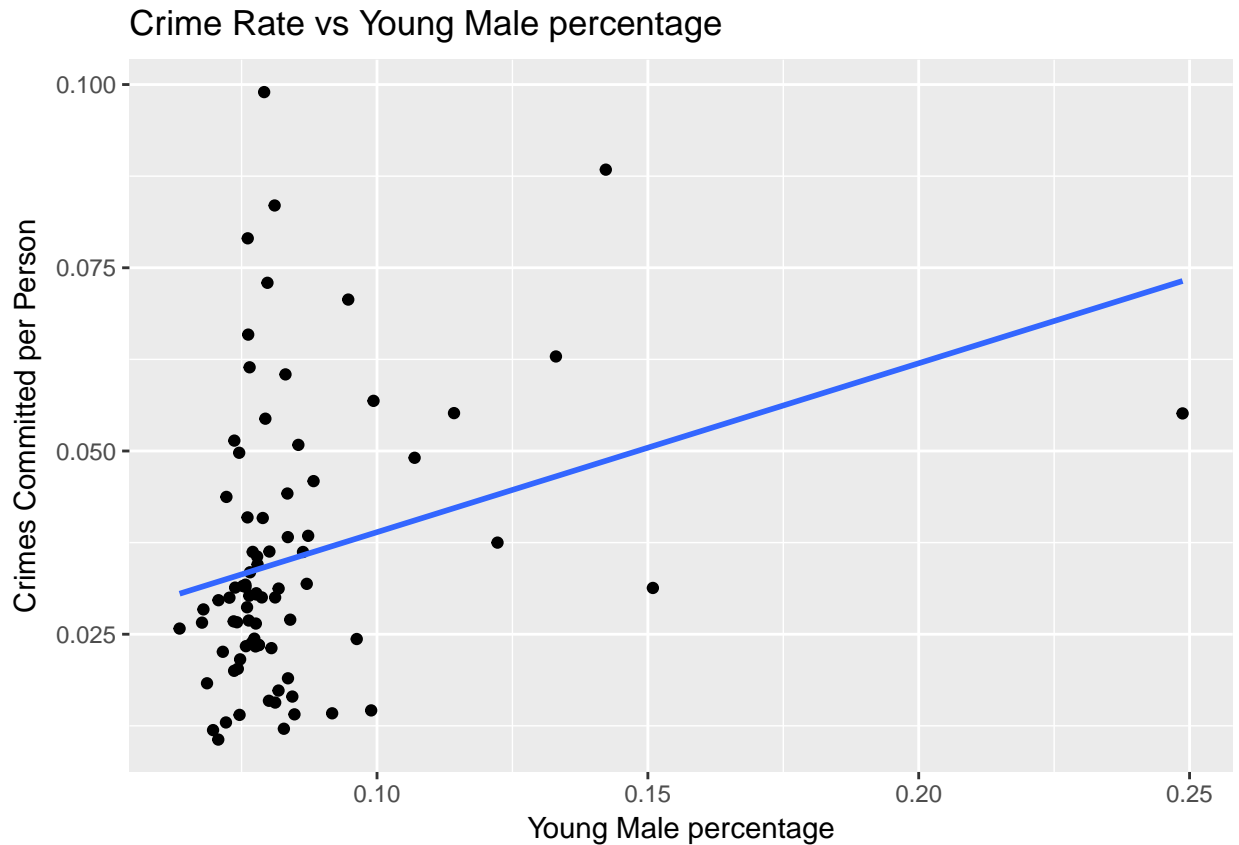



```
qplot(t$prbconv, t$crmrate) +  
  labs(title = 'Crime Rate vs Conviction Probability', x = 'Conviction Probability', y = 'Crimes Comm.  
  geom_smooth(method = 'lm', se = FALSE)
```



We also see that percentage of young male is slightly correlated with crime rate.

```
qplot(t$pctymle, t$crmrte) +
  labs(title = 'Crime Rate vs Young Male percentage', x = 'Young Male percentage', y = 'Crimes Committed per Person') +
  geom_smooth(method = 'lm', se = FALSE)
```



Model Building

First we include key variables, one from each category of geographical variables, punishment variables, and economic variables, that have high correlation with crime rates - density, probability of arrest, and tax per capita. From economic variables category, we choose tax revenue per capita ($r = 0.48$) instead of wage ($r = 0.5$), because we suspect that high wage is not directly correlated with high crime. On the other hand, higher tax relating to higher crime seems to make logical sense, assuming that our crime data includes financial crimes as well from “mix” variable that shows non-face-to-face crimes.

```
m1 = lm(t$crmte ~ t$density)
m2 = lm(t$crmte ~ t$prbarr)
m3 = lm(t$crmte ~ t$taxpc)
m4 = lm(t$crmte ~ t$density + t$prbarr + t$taxpc)
stargazer(m1, m2, m3, m4, type = 'text')
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               crmte
##                               (1)          (2)          (3)          (4)
## -----
## density                      0.009***
##                               (0.001)
##
## prbarr                       -0.088***
##                               -0.04
```

```
## (0.017) (0.0
##
## taxpc 0.001*** 0.000
## (0.0001) (0.0
##
## Constant 0.022*** 0.061*** 0.009* 0.02
## (0.002) (0.005) (0.006) (0.0
##
## -----
## Observations 81 81 81 8
## R2 0.525 0.256 0.231 0.6
## Adjusted R2 0.519 0.247 0.221 0.6
## Residual Std. Error 0.013 (df = 79) 0.016 (df = 79) 0.017 (df = 79) 0.011 (d
## F Statistic 87.187*** (df = 1; 79) 27.209*** (df = 1; 79) 23.674*** (df = 1; 79) 47.496*** (
## =====
## Note: *p<0.1; **p<0.05
```

As probability of arrest increases by 4.7%, as tax revenue per capita decreases by XX dollars (???), and as population is less dense by ???, crime rate falls by 1%. As the standard of arresting is more strict, crime rates are expected to be lower. This model can explain 63.6% of the change in crime rates.

Finding the best fit model

Next we add more variables with relatively higher correlation with crime rates to make our model more precise. However, we would like to leave out variables that absorb the causal effects, such as police per capita variable.

```
m5 = lm(t$crmte ~ t$density + t$prbarr + t$prbconv + t$pctymle + t$taxpc + t$pctmin80)
m6 = lm(t$crmte ~ t$prbarr + t$prbconv + t$prbpris + t$avgse + t$density + t$taxpc + t$pctmin80 + t$m
stargazer(m5, m6, type = 'text')
```

```
##
## =====
## Dependent variable:
## -----
## crmrte
## (1) (2)
## -----
## density 0.006*** 0.005***
## (0.001) (0.001)
##
## prbarr -0.059*** -0.054***
## (0.011) (0.012)
##
## prbconv -0.015** -0.017**
## (0.007) (0.007)
##
## prbpris 0.005
## (0.014)
##
## avgse -0.0002
## (0.0004)
##
## pctymle 0.135*** 0.140***
```

```

##                                (0.044)                (0.047)
##
## wage                          0.00000
##                                (0.00000)
##
## taxp                          0.0004***             0.0004***
##                                (0.0001)             (0.0001)
##
## pctmin80                      0.036***             0.037***
##                                (0.006)             (0.006)
##
## mix                          -0.012
##                                (0.016)
##
## Constant                      0.015*              0.011
##                                (0.009)             (0.018)
##
## -----
## Observations                  81                  81
## R2                           0.801                0.805
## Adjusted R2                   0.785                0.777
## Residual Std. Error    0.009 (df = 74)          0.009 (df = 70)
## F Statistic            49.713*** (df = 6; 74)  28.861*** (df = 10; 70)
## =====
## Note:                        *p<0.1; **p<0.05; ***p<0.01

```

With the highest R^2 at 0.785, our best fit model includes the following variables - density, probability of arrest, probability of conviction, percentage of young male, percentage of minority, and tax revenue. Crime rates can be best predicted when taking all of these variables into consideration.

Crime rates can be decreased by 1% with following measures: 1. Increase police presence in areas with higher population density 2. Make arresting and convicting standards more strict 3. Make more programs targeted towards keeping young male minorities out of crime 4. Decrease tax rate

Omitted Variables

measured coefficient = true coefficient + omitted variable bias
 $\alpha_1 = \beta_1 + \beta_2 \Delta_1 y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$
 $\text{omit } x_k \quad x_k = \Delta_0 + \Delta_1 x_1 + \dots + \Delta_{(k-1)} x_{(k-1)}$
 $y = (\beta_0 + \beta_1 \Delta_0) + (\beta_1 + \beta_2 \Delta_1) x_1 + \dots + (\beta_{(k-1)} + \beta_k \Delta_{(k-1)}) x_{(k-1)}$
 $-0.059 = \beta_1 + (-)(-)$
 $\beta_1 < -0.059$ Morality ~ (0)density (-)prbarr (-)prbconv (0)taxpc (0)pctmin80 (-)pctymle Education Climate