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 pctymle. 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.

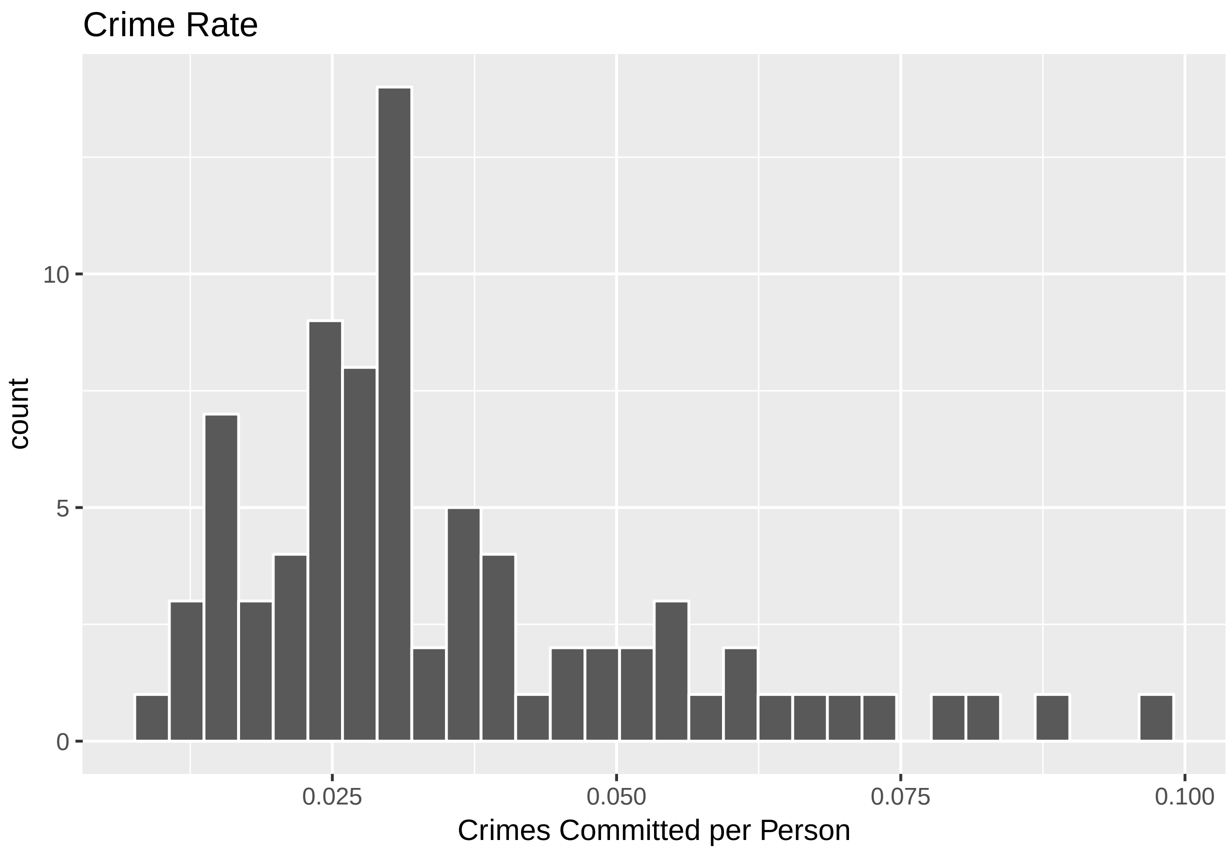
```
stargazer(data.frame(t), type = 'text')
```

```
##
## ======================================================
## 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
## avgsen    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
## wmfg      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')
```



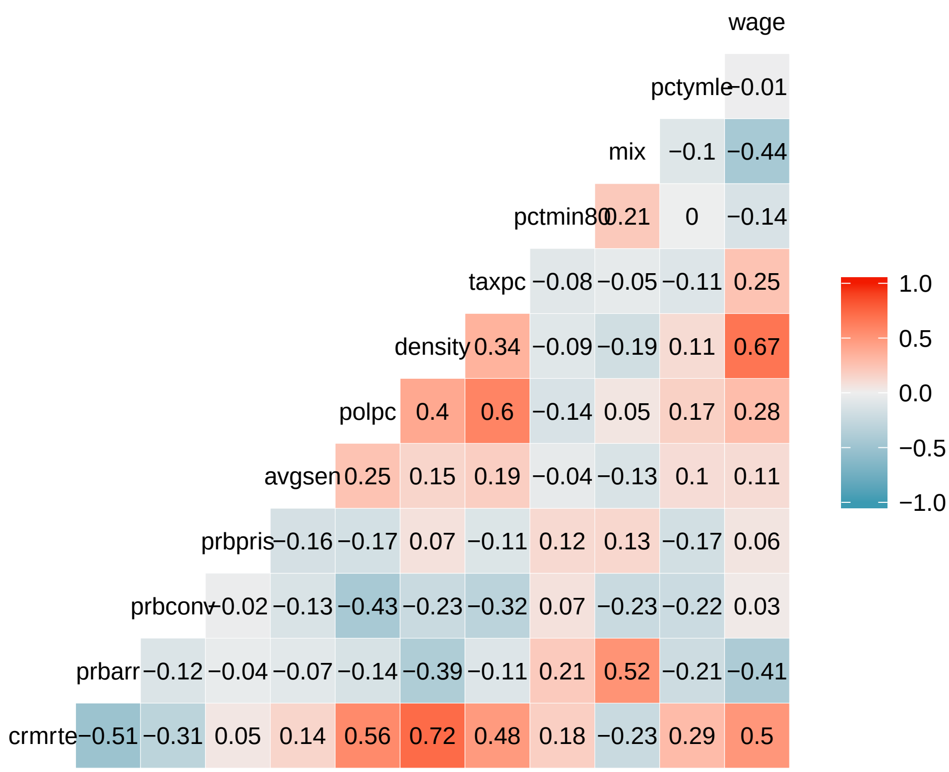

```
summary(t$crmrte)
```

```
##     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.

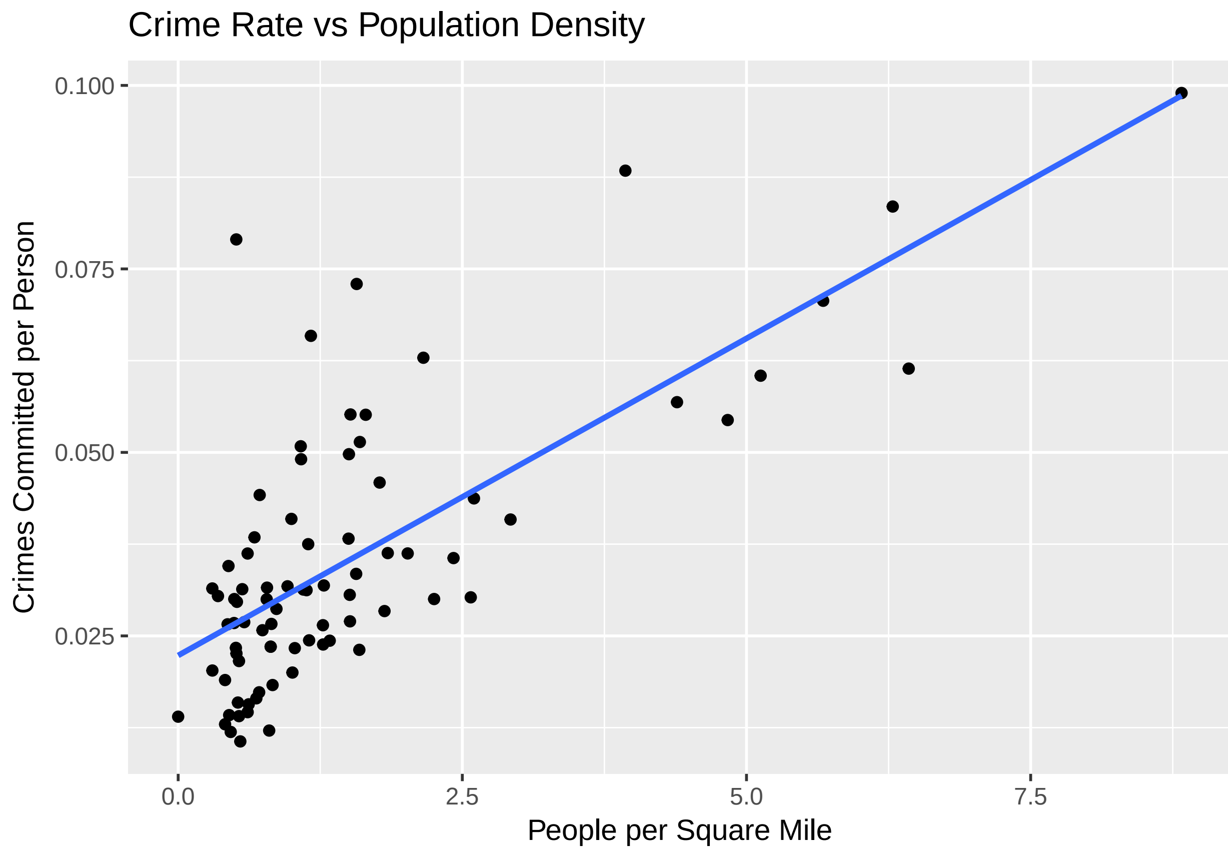
```
t2 = t %>% select(crmrte, prbarr, prbconv, prbpris, avgsen, polpc, density, taxpc, pctmin80, mix, pctyml
ggcorr(t2, label = TRUE, label_round = 2, label_size = 3, size = 3) + ggtitle('Correlation Matrix')
```

Correlation Matrix



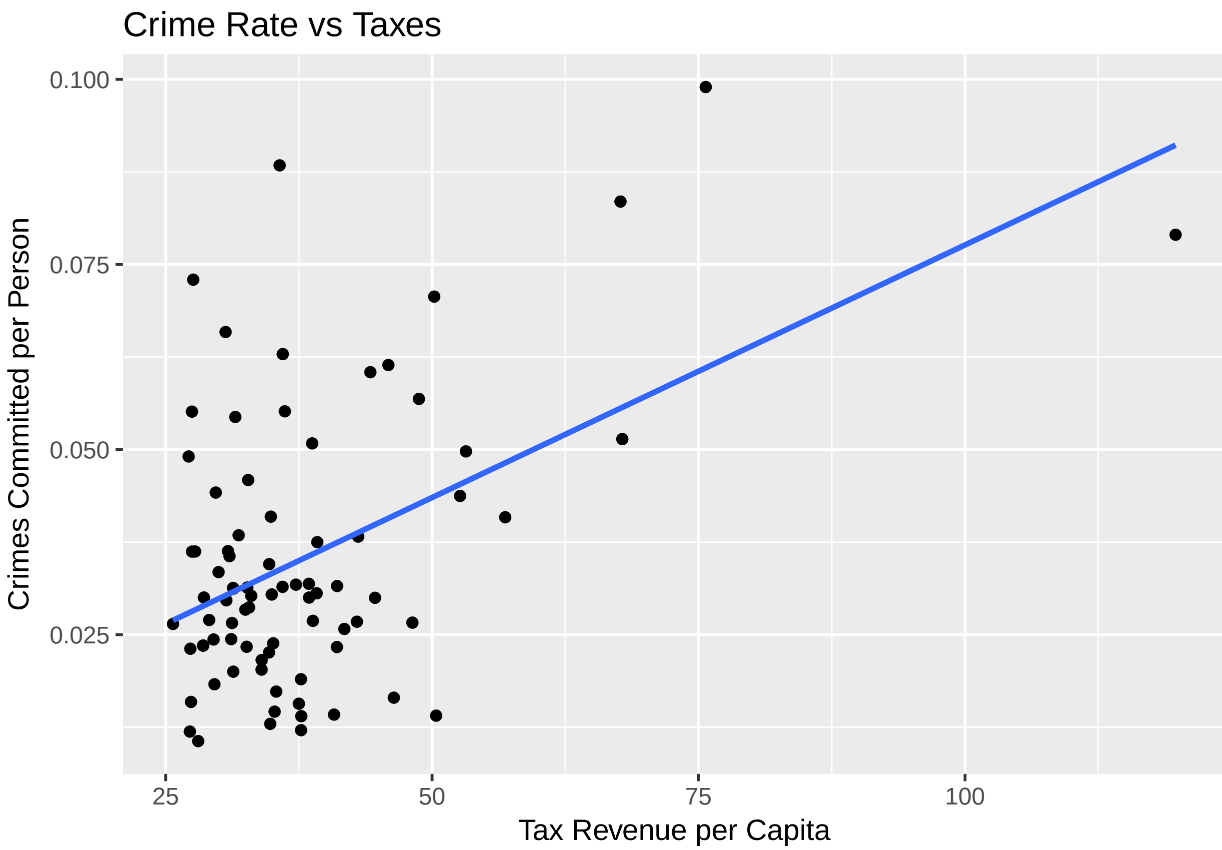
From the correlation matrix, we see that population density stands out as being highly correlated with crime rate (r = 0.72). This variable looks like a good candidate as a causal predictor for crime rate. One explanation could be that as more people move into an area, the increased number of interactions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

```
qplot(t$density, t$crmrte) +
    labs(title = 'Crime Rate vs Population Density', x = 'People per Square Mile', y = 'Crimes Committe
    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.

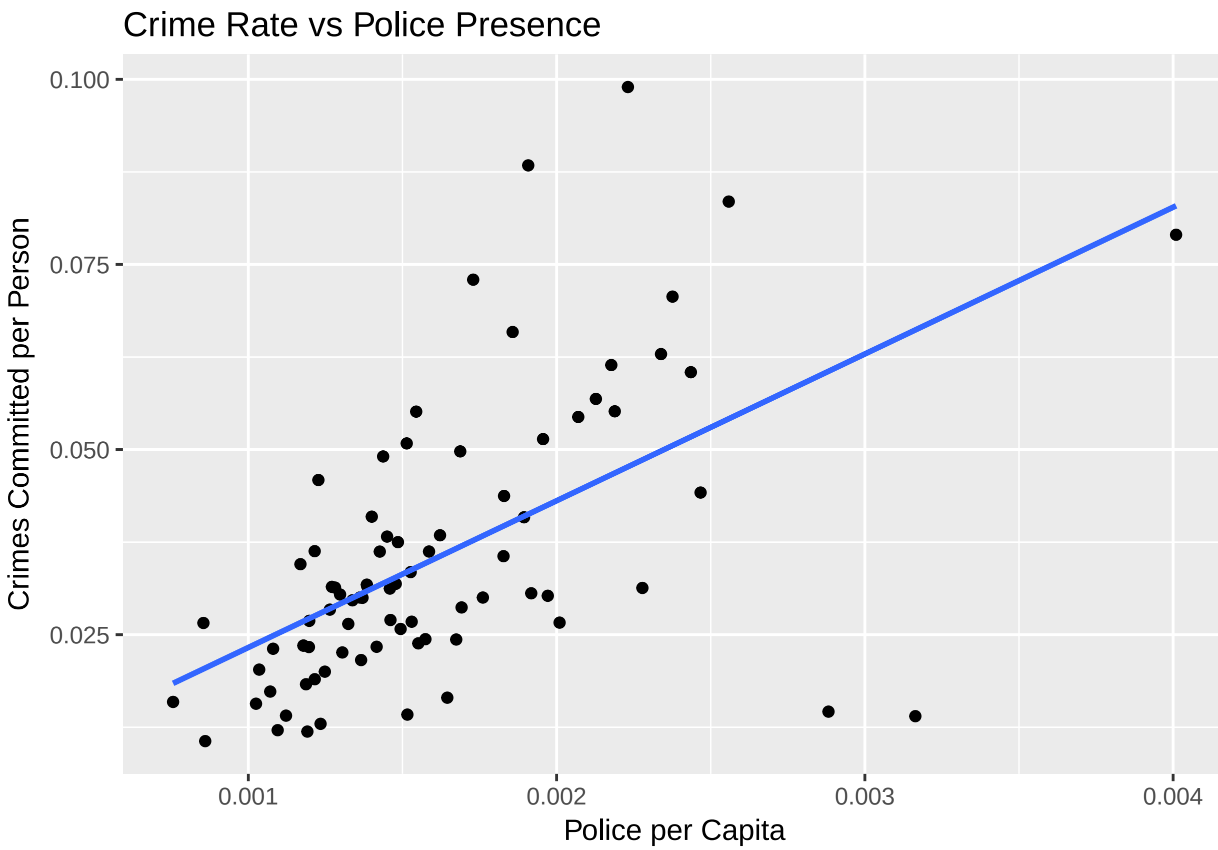
```
qplot(t$taxpc, t$crmrte) +
    labs(title = 'Crime Rate vs Taxes', x = 'Tax Revenue per Capita', y = 'Crimes Committed per Person')
    geom_smooth(method = 'lm', se = FALSE)
```




```r
qplot(t$wage, t$crmrte) +
    labs(title = 'Crime Rate vs Wages', x = 'Weekly Wages', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
```

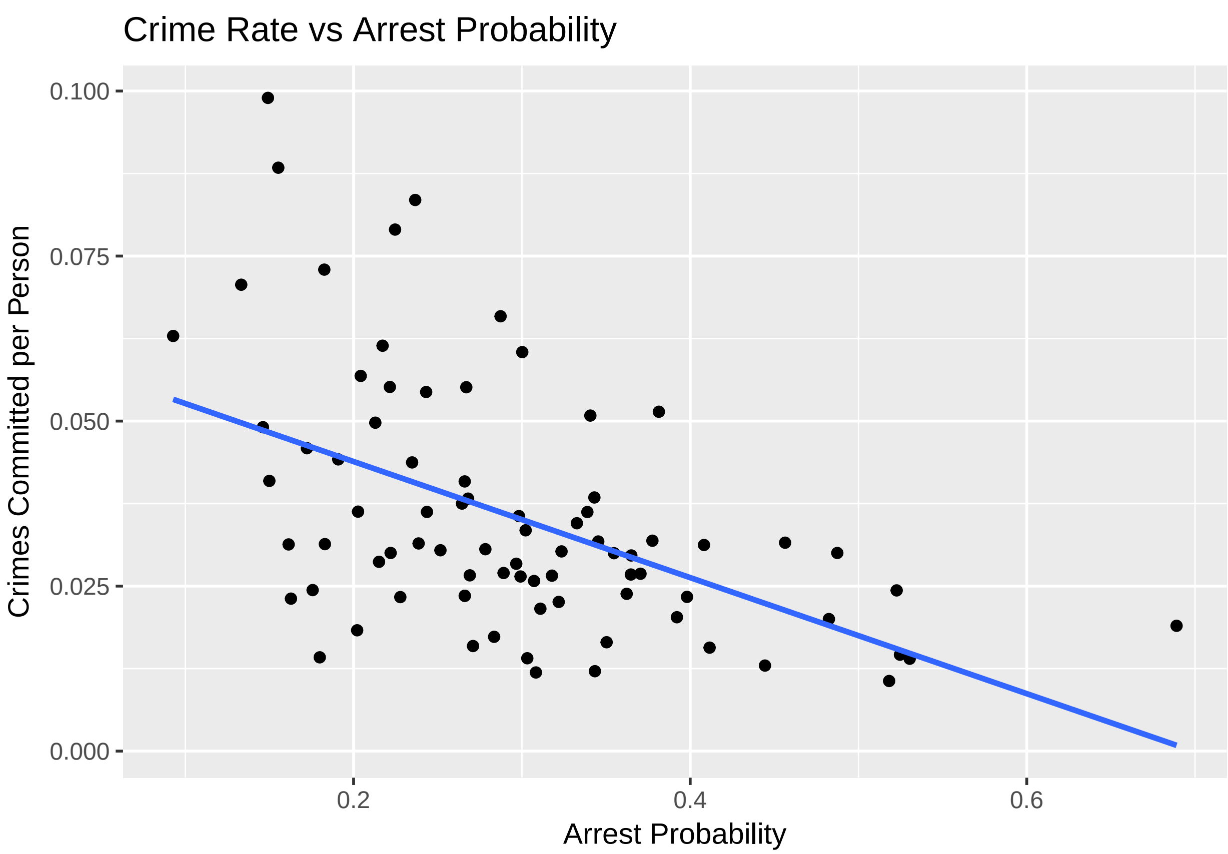
Interestingly, the relationship between police per capita and crime rate is positive and moderately large (r = 0.56). This means that either increasing police presence makes crime rate worse or that crime is causing an increase in police presence rather than vice versa. The latter explanation seems much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

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

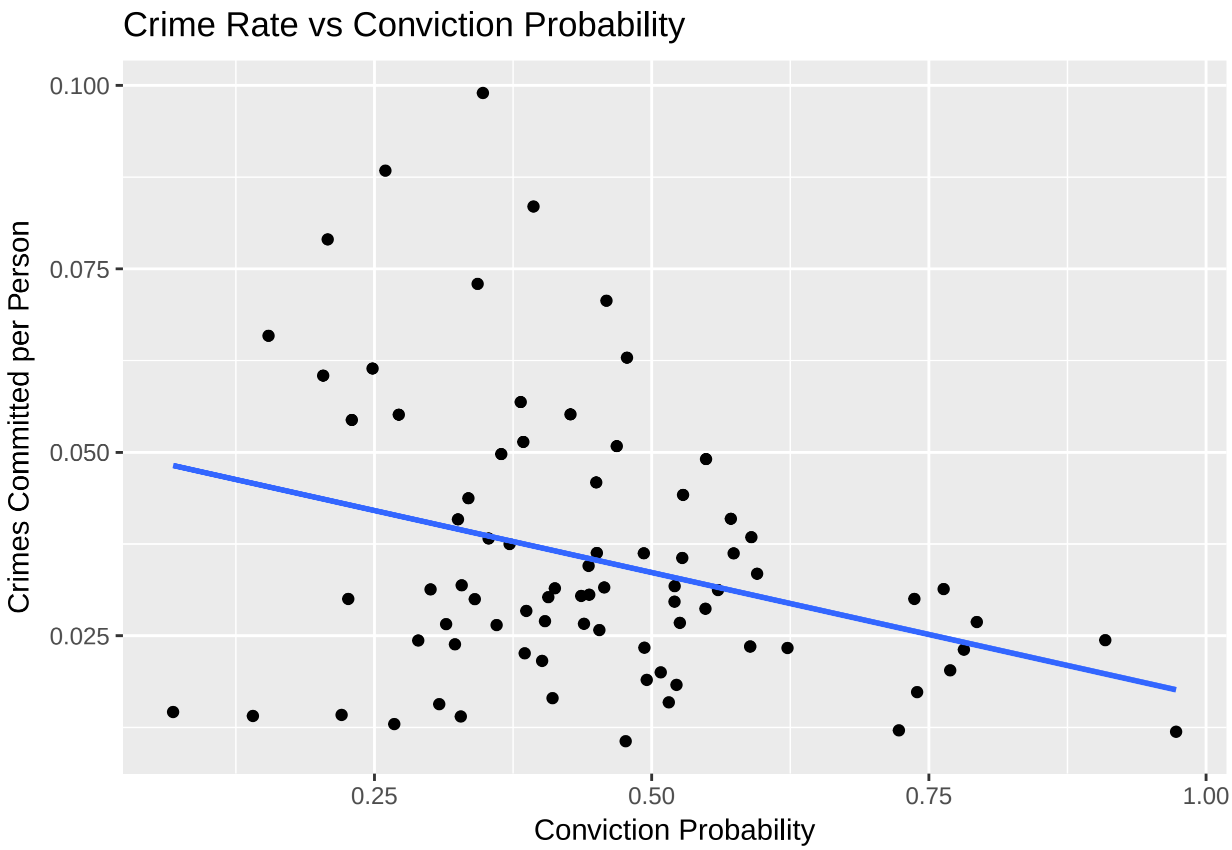


Of the three "certainty of punishment" variables, it looks like arrest probability has a moderate effect (r = -0.51) and conviction probability has a weak effect (r = -0.31), but probability of prison sentence has almost no effect (r = 0.05). It is important to note that these three probabilities seem uncorrelated with one another, so we will include multiple ones in our regression without fear of multicolinearity. 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 pe
    geom_smooth(method = 'lm', se = FALSE)
```




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



Model Building

First we include three varialbes with high correlation coefficients that we found above - density, probability of conviction, and probability of arrest.

We would like to leave out police per capita and mix variables, because they introduce

```
m1 = lm(t$crmrte ~ t$density)
m2 = lm(t$crmrte ~ t$density + t$prbarr + t$prbconv)
m3 = lm(t$crmrte ~ t$density + t$prbarr + t$prbconv + t$taxpc + t$pctmin80 + t$pctymle)
m4 = lm(t$crmrte ~ t$prbarr + t$prbconv + t$prbpris + t$avgsen + t$density + t$taxpc + t$pctmin80 + t$mi

stargazer(m1, m2, m3, m4, type = 'text')
```

```
##
## ================================================================================================
##                                                         Dependent variable:
##                     ----------------------------------------------------------------------------
##                                                              crmrte
##                               (1)                  (2)                  (3)                  (4
## ------------------------------------------------------------------------------------------------
## density                    0.009***             0.007***             0.006***             0.005
##                            (0.001)              (0.001)              (0.001)              (0.0
##
## prbarr                                         -0.055***            -0.059***            -0.05
##                                                 (0.014)              (0.011)              (0.0
```

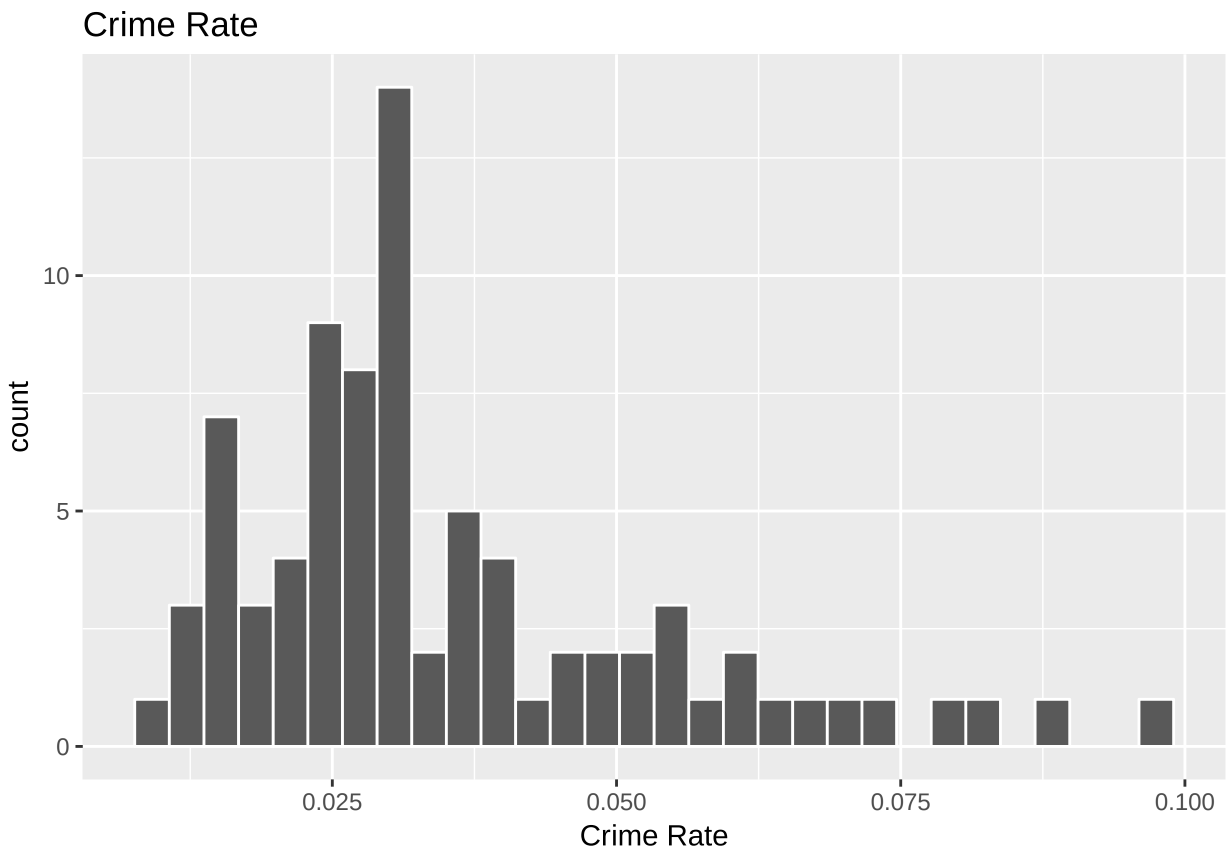
```
##
## prbconv                                        -0.024***              -0.015**               -0.0
##                                                (0.008)                (0.007)                (0.0
##
## prbpris                                                                                        0.0
##                                                                                               (0.0
##
## avgsen                                                                                        -0.0
##                                                                                              (0.00
##
## taxpc                                                                 0.0004***               0.000
##                                                                       (0.0001)               (0.00
##
## pctmin80                                                               0.036***               0.03
##                                                                        (0.006)               (0.0
##
## mix                                                                                           -0.0
##                                                                                               (0.0
##
## pctymle                                                                0.135***               0.14
##                                                                        (0.044)               (0.0
##
## wage                                                                                          0.00
##                                                                                              (0.00
##
## Constant                   0.022***               0.053***              0.015*                 0.0
##                            (0.002)                (0.007)               (0.009)               (0.0
##
## -----------------------------------------------------------------------------------------------------
## Observations                 81                     81                    81                    8
## R2                          0.525                  0.626                 0.801                 0.8
## Adjusted R2                 0.519                  0.612                 0.785                 0.7
## Residual Std. Error    0.013 (df = 79)        0.012 (df = 77)       0.009 (df = 74)        0.009 (
## F Statistic         87.187*** (df = 1; 79) 43.051*** (df = 3; 77) 49.713*** (df = 6; 74) 28.861*** (
## =====================================================================================================
## Note:                                                                                 *p<0.1; **p<0.0
```

Omitted Variables

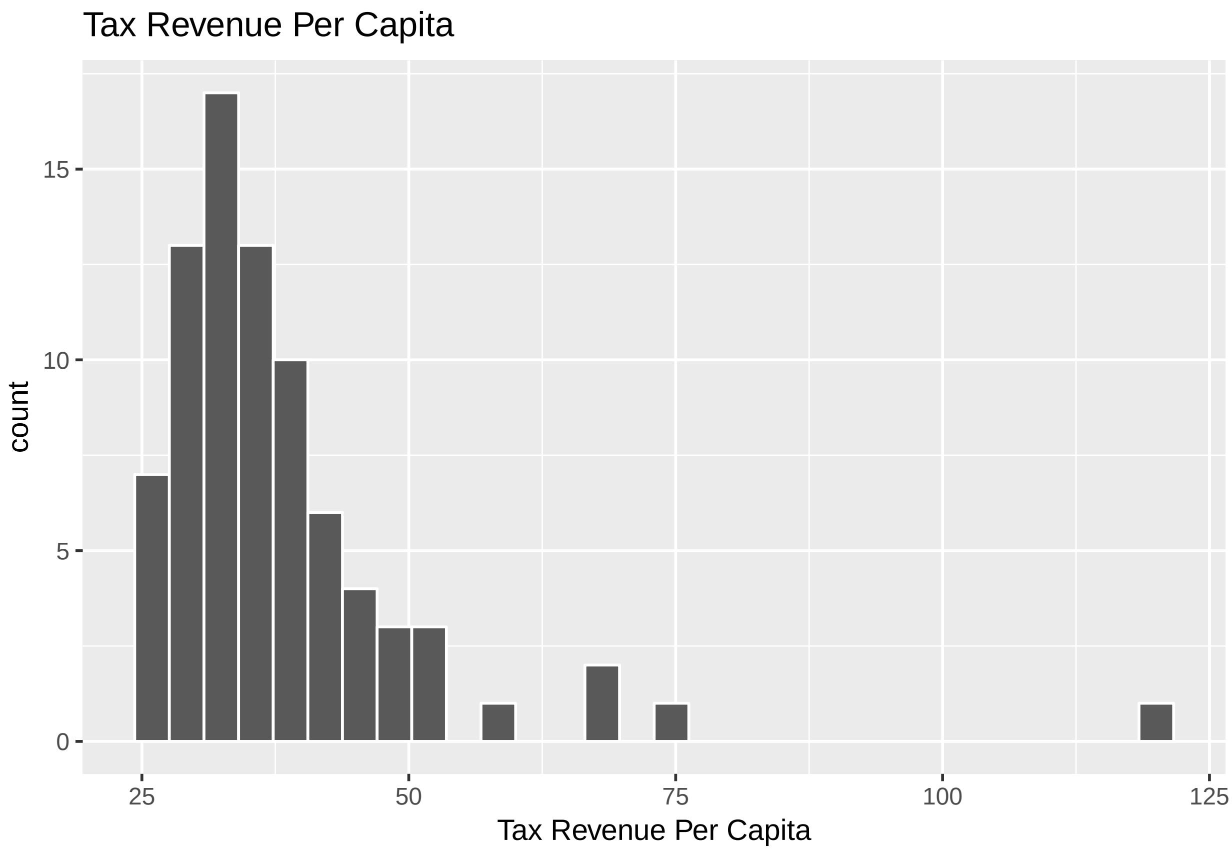
measured coefficient = true coefficient + omitted variable bias alpha1 = beta1 + beta2 delta1 y = beta0 + beta1x1 + … + betak xk + u omit xk xk = delta0 + delta1x1 +… + delta(k-1)x(k-1) y = (beta0 + betak delta0) + (beta1 + betak delta1)x1 + … + (beta(k-1) + betak delta(k-1))x(k-1) -.059 = beta1 + (-)(-) beta1 < -.059 Morality ~ (0)density (-)prbarr (-)prbconv (0)taxpc (0)pctmin80 (-)pctymle Education Climate

Here is some single variate EDA.

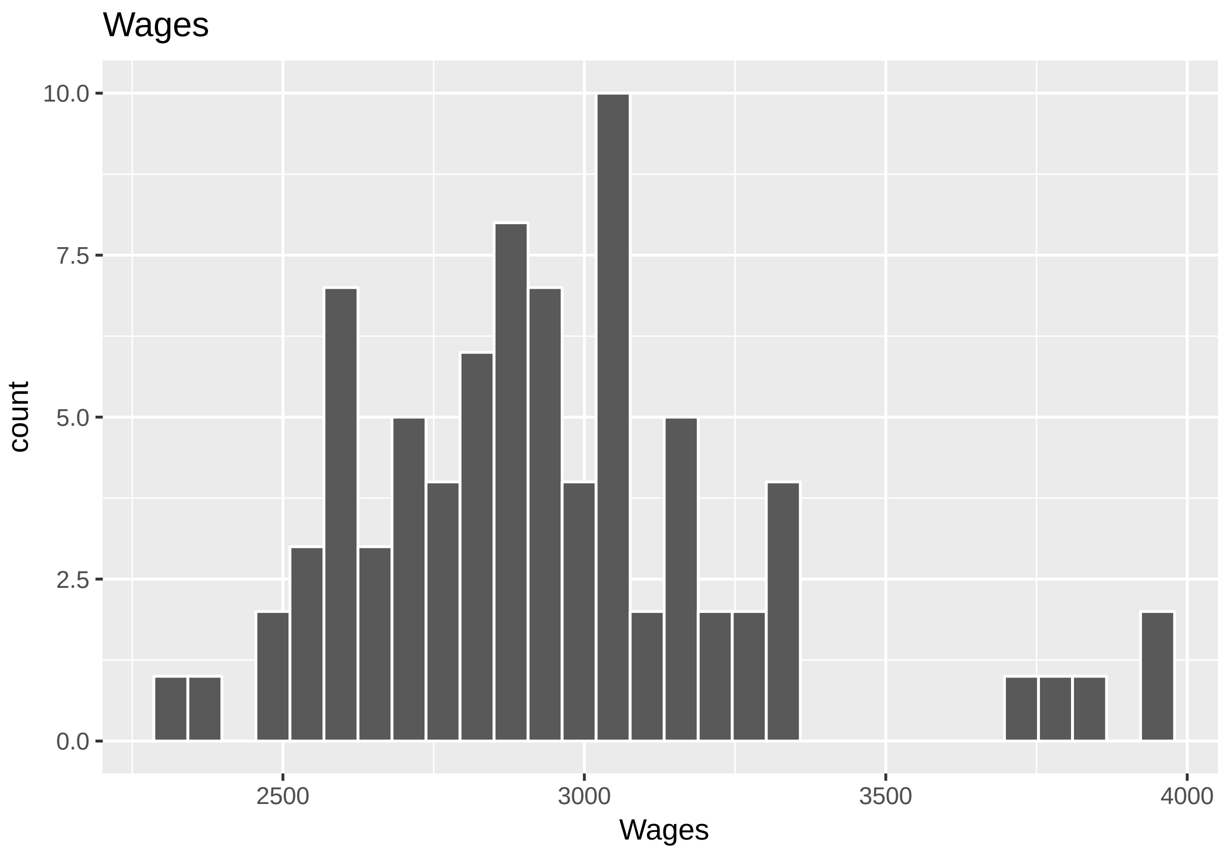
```
qplot(t$crmrte, geom = 'histogram', col = I('white'), main = 'Crime Rate', xlab = 'Crime Rate')
```




```
qplot(t$taxpc, geom = 'histogram', col = I('white'), main = 'Tax Revenue Per Capita', xlab = 'Tax Revenu
```




```
qplot(t$wage, geom = 'histogram', col = I('white'), main = 'Wages', xlab = 'Wages')
```



Wages
10.0
7.5
5.0
2.5
0.0
count
2500
3000
3500
4000
Wages