Lab 3

David Hou, Scott Hungerfield, Irene Seo

March 20, 2018

Introduction

The purpose of this study is to provide information for a 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 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. Specifically, we find nine cases where prbconv is greater than 1 and one case where both are greater than one. We create a badprb flag which is set to 1 for the former nine cases and 2 for singular latter case. We also create a second data table, where all the questionable probabilities are removed.

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) %>%
    mutate(badprb = as.factor((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$urban) = c('Non-urban', 'Urban')
levels(t$badprb) = c('Normal', 'prbconv > 1', 'prbarr > 1 and prbconv > 1')

t2 = t %>%
    filter(prbarr < 1 & prbconv < 1)
```

Here is a summary of the data (with the questionable probabilities left in).

```
stargazer(data.frame(t), type = 'latex', nobs = FALSE, header = FALSE, float = FALSE)
```

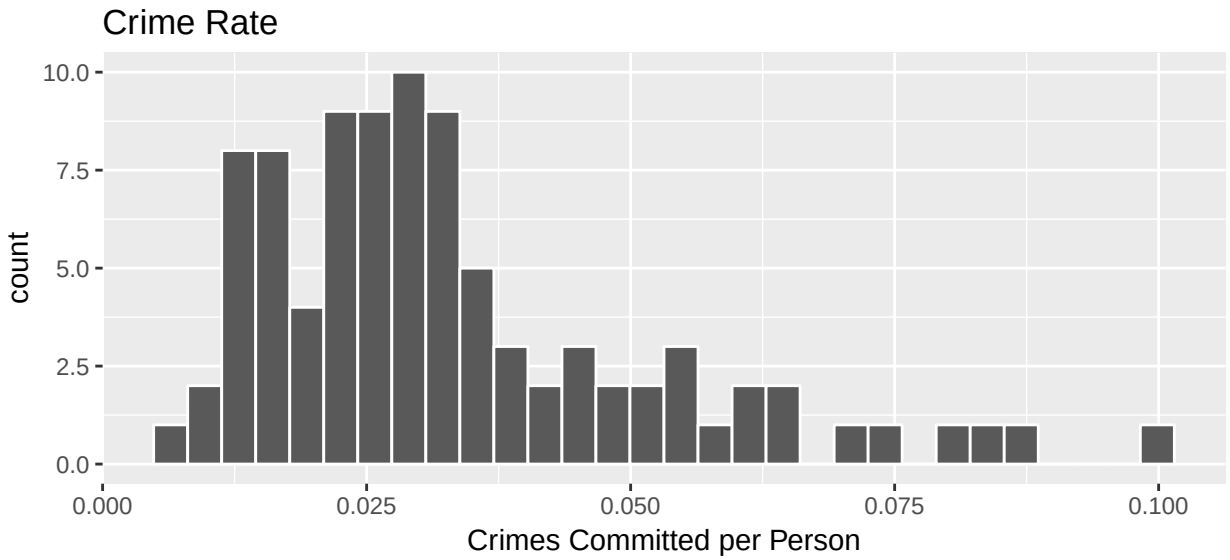
Statistic	Mean	St. Dev.	Min	Max
crmrte	0.033	0.019	0.006	0.099
prbarr	0.295	0.137	0.093	1.091
prbconv	0.551	0.352	0.068	2.121
prbpris	0.411	0.080	0.150	0.600
avgsen	9.647	2.847	5.380	20.700
polpc	0.002	0.001	0.001	0.009
density	1.429	1.514	0.00002	8.828
taxpc	38.055	13.078	25.693	119.761
pctmin80	0.255	0.170	0.013	0.643
wcon	285.358	47.487	193.643	436.767
wtuc	411.668	77.266	187.617	613.226
wtrd	211.553	34.216	154.209	354.676
wfir	322.098	53.890	170.940	509.466
wser	275.564	206.251	133.043	2,177.068
wmfg	335.589	87.841	157.410	646.850
wfed	442.901	59.678	326.100	597.950
wsta	357.522	43.103	258.330	499.590
wloc	312.681	28.235	239.170	388.090
mix	0.129	0.081	0.020	0.465
pctymle	0.084	0.023	0.062	0.249

Examining Key Variables of Interest

Metric Variables

We start our analysis by first looking at the metric variables, i.e. all the variables less county, year, west, central, and urban. Crime rate is our most important variable as it is the output that we are trying to study.

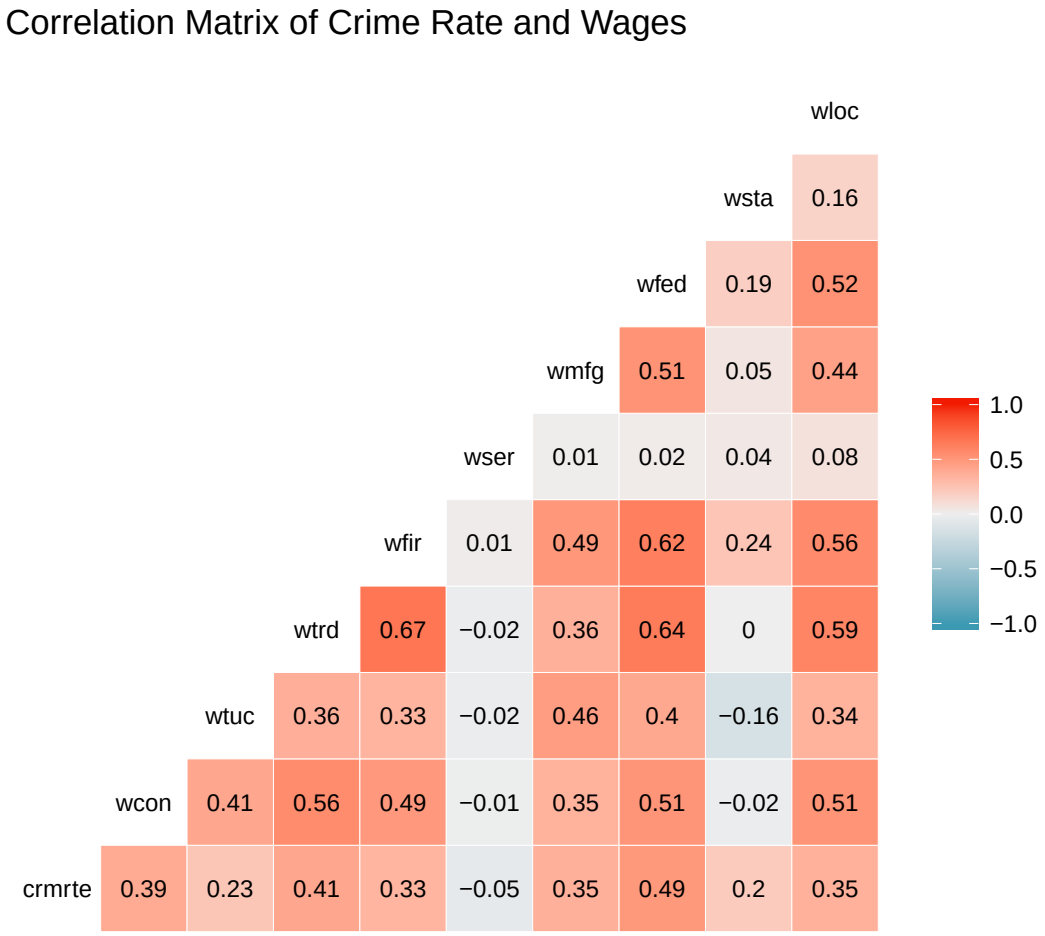
```
qplot(t$crmrte, col = I('white')) +
    labs(title = 'Crime Rate', x = 'Crimes Committed per Person')
```



We see that 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 matrices among the variables. Since a large portion of dataset deals with wage, let us first examine those variables by themselves. We first take a look at the correlation matrix among them and crime rate.

```
ggcorr(t %>% select(crmrte, wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc),
       label = TRUE, label_round = 2, label_size = 3, size = 3) +
    ggtitle('Correlation Matrix of Crime Rate and Wages')
```

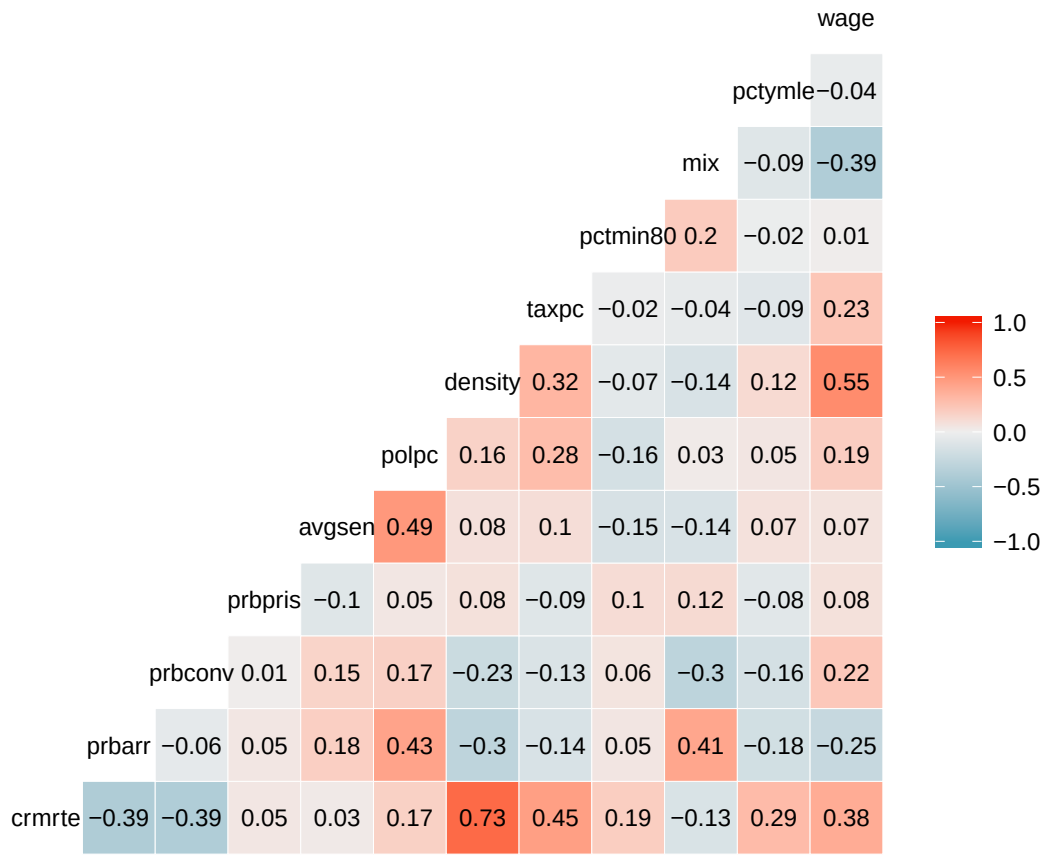


Surprisingly, we find that crime rate is actually positively correlated with all wages except from the service industry. This seems counter to common sentiment that crime is more prevalent in low income areas.

For ease of comparison with the other variables, we create a new one that is the sum of all the other wages. We will see later that this data transformation does not make a large difference in the regression analysis.

```
t = t %>% mutate(wage = wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc)
t2 = t2 %>% mutate(wage = wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc)

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



From the correlation matrix, we see that population density stands out as being highly correlated with crime rate (r = 0.73). This variable looks like a good candidate as a 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. In addition, more people in an area probably increases the chance that crime will actually be seen.

The other two variables with moderately positive correlation are tax per capita (r = 0.45) and total wages (r = 0.38). Note that population density is weakly correlated with taxes (r = 0.32) and moderately correlated with wages (r = 0.55). We believe that taxes and wages are not directly causing higher crime rates but are rising along with crime rate because they are rising along with density. Also, it is interesting that taxes and wages are not very correlated with themselves (r = 0.23). This finding is surprising, as one would expect that wages and taxes would go up very closely with each other. Along with the questionable probability numbers, we are left to question the integrity of this dataset. At the minimum, we desire some extra explanation as to how the data were taken.

An important finding is that the relationship between police per capita and crime rate is positive (r = 0.17). 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. Thus, we will not regress crime rate on police per capita, as the direction of causality is questionable.

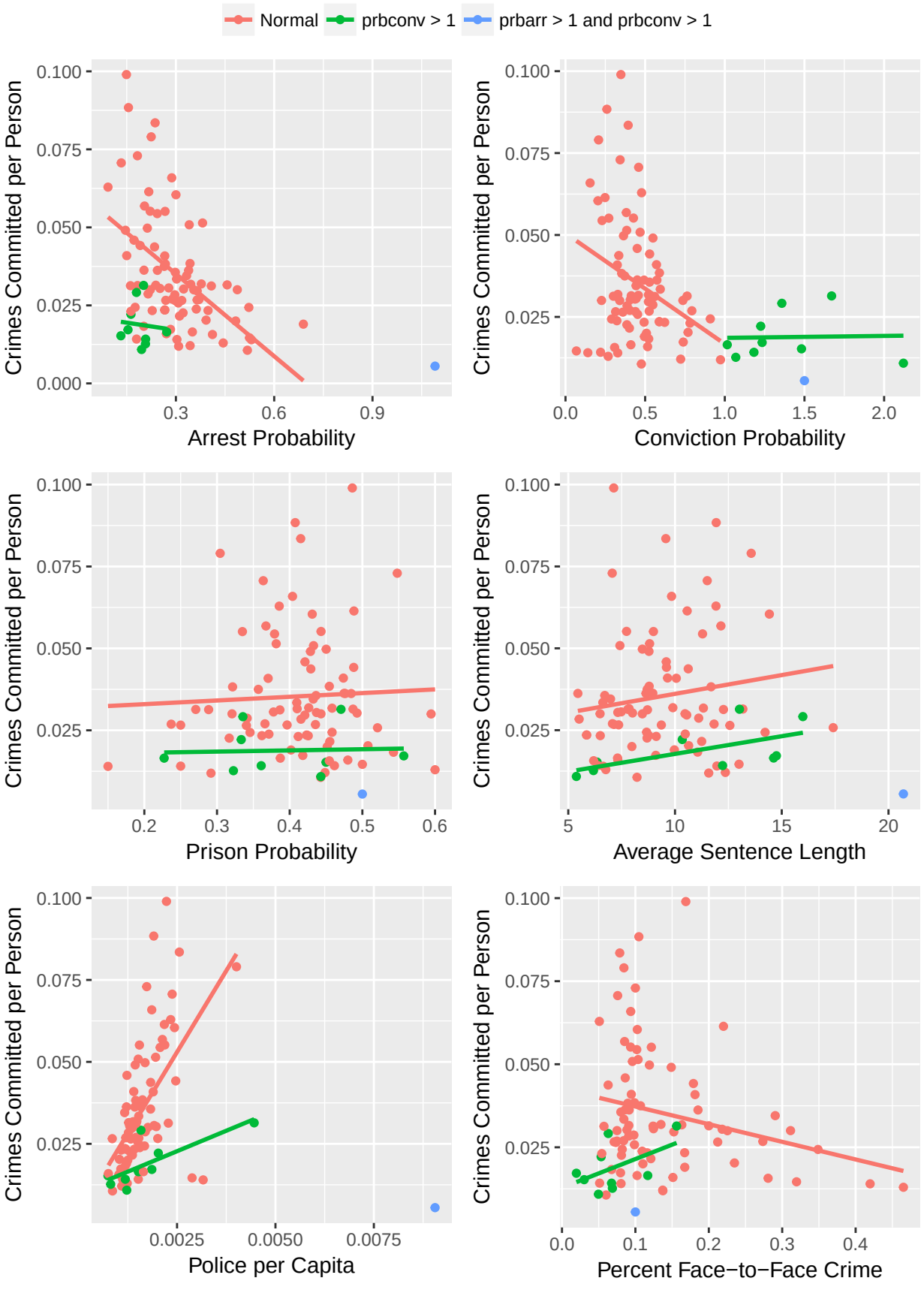
Of the three "certainty of punishment" variables, it looks like arrest probability has a moderate effect (r = -0.39) and conviction probability has a weak effect (r = -0.39), but probability of prison sentence has almost no effect (r = 0.05). It is important to note that these three probabilities seem uncorrelated with one another, so we can 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 = 0.03).

Finally, the two demographic variables seem to have relatively weak correlations with crime rate. However, their directions are at least in line with historic sentiment (young male minorities are commonly associated with crime).

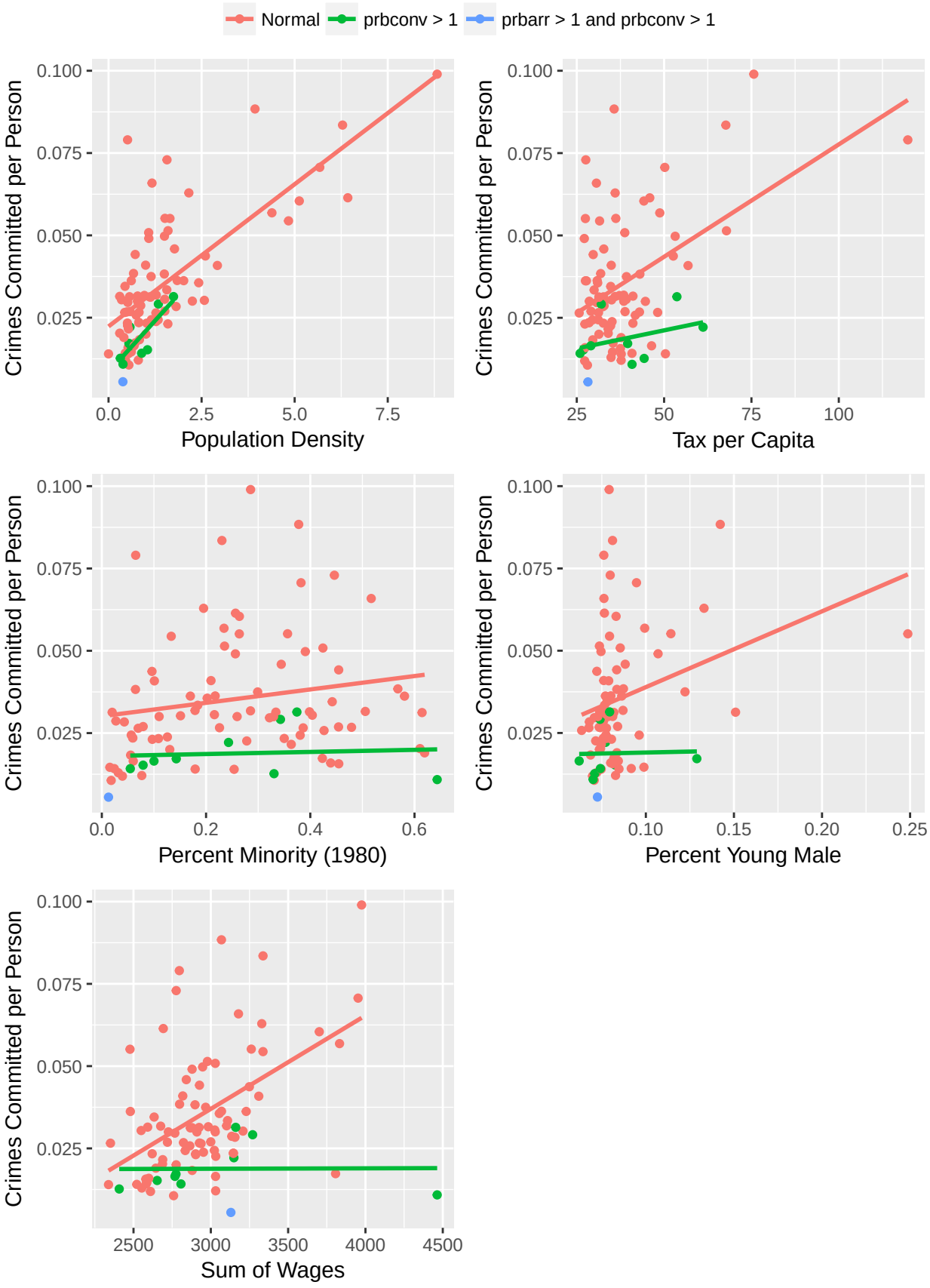
Below are the bivariate scatter plots between crime rate and each of the input variables. The linear regression lines are shown on the plots for convenience but are not meant to be rigorous models at this point. We have also divided the variables up into two rough groups: one dealing directly dealing with law enforcement and one dealing with socioeconomic/demographic factors. Just for a better layout, we will only display the code used to generate the first set of plots. The second set of plots were generated completely analgously.

Law Enforcement Variables

```
p1 = qplot(t$prbarr, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Arrest Probability', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p2 = qplot(t$prbconv, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Conviction Probability', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p3 = qplot(t$prbpris, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Prison Probability', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p4 = qplot(t$avgsen, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Average Sentence Length', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p5 = qplot(t$polpc, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Police per Capita', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p6 = qplot(t$mix, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Percent Face-to-Face Crime', y = 'Crimes Committed per Person')
    geom_smooth(method = 'lm', se = FALSE)
ggarrange(p1, p2, p3, p4, p5, p6, nrow = 3, ncol = 2, common.legend = T)
```


Normal
prbconv > 1
prbarr > 1 and prbconv > 1
Crimes Committed per Person
0.100
0.075
0.050
0.025
0.000
0.3
0.6
0.9
Arrest Probability
0.0
0.5
1.0
1.5
2.0
Conviction Probability
0.2
0.3
0.4
0.5
0.6
Prison Probability
5
10
15
20
Average Sentence Length
0.0025
0.0050
0.0075
Police per Capita
0.0
0.1
0.2
0.3
0.4
Percent Face−to−Face Crime

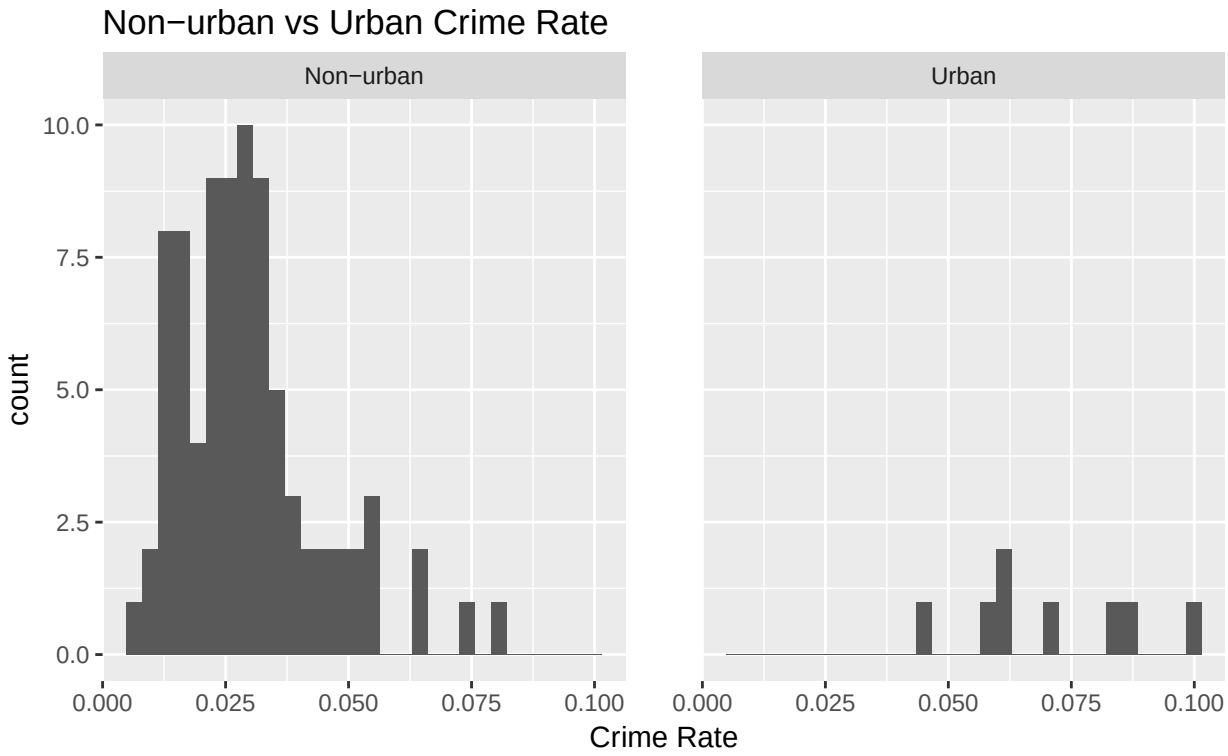
Socioeconomic/Demographic Variables



Dummy Variables

Next, we examine the effect of the three dummy indicators. First we see if there is a difference in crime rate between non-urban and urban counties. As an aside, we are assuming "non-urban" refers to a combination of suburb-dominated and rural areas. We are not sure if this geographical assumption is actually true in North Carolina.

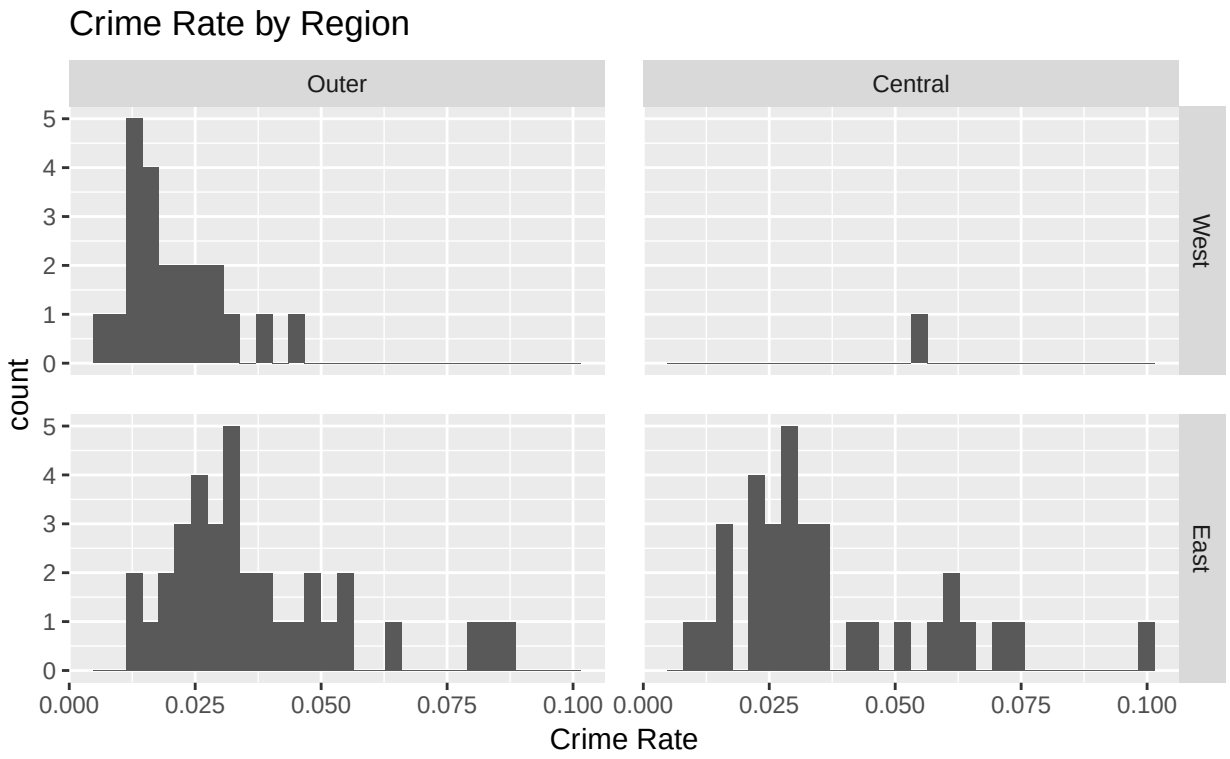
```
ggplot(t, aes(crmrte)) +
    geom_histogram() +
    facet_grid(. ~ urban) +
    theme(panel.spacing = unit(2, "lines")) +
    labs(title = 'Non-urban vs Urban Crime Rate', x = 'Crime Rate')
```



We see that there are only 8 counties coded as urban, which is probably too few to make any sweeping inferences. We will only mention in passing that the crime rate in urban counties does look higher than that in non-urban counties.

Next we examine the differences in geographic region.

```
ggplot(t, aes(crmrte)) +
    geom_histogram() +
    facet_grid(west ~ central) +
    theme(panel.spacing = unit(1, "lines")) +
    labs(title = 'Crime Rate by Region', x = 'Crime Rate')
```



Again we notice a sparcity in data; this time there are only 23 western counties, with a mere single county in the western central area. However, we do see a relatively even division between central and outer counties, so we will run a t-test to see if there is any difference in crime rate between the two.

```
t.test(t[t$central == 'Outer', ]$crmrte,
       t[t$central == 'Central', ]$crmrte)
```

```
##
##  Welch Two Sample t-test
##
## data:  t[t$central == "Outer", ]$crmrte and t[t$central == "Central", ]$crmrte
## t = -1.5802, df = 63.769, p-value = 0.119
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01485057  0.00173351
## sample estimates:
##  mean of x  mean of y
## 0.03094979 0.03750832
```

With a p-value of 0.16, we fail to reject the null hypothesis that there is difference in crime rate between central and outer counties.

Model Building

We will now proceed to build several ordinary least squares (OLS) regression models of crime rate. We will be reporting heteroskedasticity robust standard errors.

Wage Transformation

First, we examine whether combining the wages was a prudent choice.

```
# function for getting heteroskedasticity robust standard errors
seHC = function(...) {
    lapply(list(...), function(x) sqrt(diag(vcovHC(x))))
}
```

```
m1_wage = lm(t$crmrte ~ t$wfed)
m2_wage = lm(t$crmrte ~ t$wcon + t$wtuc + t$wtrd + t$wfir + t$wser + t$wmfg + t$wfed + t$wsta + t$wloc)
m3_wage = lm(t$crmrte ~ t$wage)

stargazer(m1_wage, m2_wage, m3_wage, type = 'latex',
          omit.stat = c('f', 'n'),
          se = seHC(m1_wage, m2_wage, m3_wage),
          star.cutoffs = c(0.05, 0.01, 0.001),
          dep.var.labels = c('Crime Rate'),
          header = FALSE,
          float = FALSE,
          title = 'Crime Rate Regressed on Wage Variables',
          covariate.labels = c('Construction', 'Trans, Util, Commun', 'Whlesle, Retail, Trade',
                               'Fin, Ins, Real Est', 'Service', 'Manuacturing', 'Federal', 'State',
                               'Local', 'Total Sum')
)
```

	Dependent variable:		
	Crime Rate		
	(1)	(2)	(3)
Construction		0.0001	
		(0.0001)	
Trans, Util, Commun		−0.00000	
		(0.00002)	
Whlesle, Retail, Trade		0.0001	
		(0.0001)	
Fin, Ins, Real Est		−0.0001	
		(0.00005)	
Service		−0.00001	
		(0.00001)	
Manuacturing		0.00004	
		(0.00003)	
Federal	0.0002***	0.0001	
	(0.00004)	(0.0001)	
State		0.0001*	
		(0.00004)	
Local		0.00000	
		(0.0001)	
Total Sum			0.00002
			(0.00001)
Constant	−0.034	−0.067*	−0.024
	(0.018)	(0.027)	(0.030)
R^2	0.236	0.322	0.141
Adjusted R^2	0.228	0.247	0.132
Residual Std. Error	0.017 (df = 89)	0.016 (df = 81)	0.018 (df = 89)
Note:		*p<0.05; **p<0.01; ***p<0.001	

We see from the above regression table that including each individual wage variable in the regression only provides a small improvement in adjusted R^2 from including just the federal wages. It also causes all the coefficients to lose signifance. When we combine all the wages into a sum, we see that the adjusted R^2 improves more and we end up with a single highly-significant coefficient. Thus, the total wage variable is a parsimonious way to model the wage effect.

Now we will proceed to build models with all the other variables. Note that we will not regress on police per capita, as we think that it absorbs some of the causal effect.

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

```
stargazer(m1, m2, m3, type = 'latex',
          omit.stat = c('f', 'n'),
          se = seHC(m1, m2, m3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          header = FALSE,
          float = FALSE,
          dep.var.labels = c('Crime Rate'),
          title = 'Crime Rate Regressed on Other Variables',
          covariate.labels = c('Population Density', 'Arrest Probability', 'Conviction Probability',
                               'Prison Probability', 'Average Prison Sentence', 'Tax per Capita',
                               'Percent Minority', 'Offense Mix', 'Percent Young Male',
                               'Sum of Wages')
)
```

	Dependent variable:		
	Crime Rate		
	(1)	(2)	(3)
Population Density	0.008***	0.007***	0.006***
	(0.001)	(0.002)	(0.002)
Arrest Probability		−0.026	−0.024
		(0.032)	(0.031)
Conviction Probability	−0.012***	−0.012	−0.015*
	(0.004)	(0.008)	(0.006)
Prison Probability			0.010
			(0.017)
Average Prison Sentence			0.0004
			(0.001)
Tax per Capita		0.0004	0.0003
		(0.0003)	(0.0003)
Percent Minority		0.028**	0.031***
		(0.010)	(0.009)
Offense Mix			−0.022
			(0.019)
Percent Young Male		0.147*	0.139**
		(0.064)	(0.053)
Sum of Wages			0.00000
			(0.00001)
Constant	0.028***	0.005	−0.0003
	(0.003)	(0.018)	(0.025)
R^2	0.583	0.766	0.778
Adjusted R^2	0.573	0.749	0.750
Residual Std. Error	0.012 (df = 88)	0.009 (df = 84)	0.009 (df = 80)
Note:			*p<0.05; **p<0.01; ***p<0.001

For model 1, we included only the explanatory variables of key interest. In this case we picked density and conviction probability because they were both relatively correlated with crime rate, but not correlated with each other. We see from the regression table that density has a highly significant coefficient but conviction probablity does not. We have already explained over 50% of the variation in crime rate with these two variables alone (probably mostly from density).

For model 2, we added in variables that increase the accuracy of our result without introducing substantial bias. For model 3, we added the remaining variables.

Omitted Variables

We identified seven omitted variables that may introduce bias to the crime rate outcome. The seven variables are a person's morals (Morals), a healthy diet (Diet), a person's mental health (MH), a person's happiness (Happiness), a person's family stability (FS), the amount of drugs in the area (Drugs), and the probability a person will report a crime (prbrc).

The table below shows omitted variables' effect on both the measure variables and the outcome (crime rate). A value or (1) represents that the omitted variable has a positive correlation with the measured or outcome variable, a (-1) represents that the omitted variable has a negative correlation with the measured or outcome variable, and a (0) represents the omitted variable has no impact on the measured or outcome variable.

Omitted Variable	Morals	Diet	MH	Happiness	FS	Drugs	prbrc
crmrate (B1)	-1	0	-1	-1	-1	1	1
prbarr	-1	0	-1	-1	-1	1	1
prbconv	-1	0	-1	-1	-1	1	1
density	0	-1	0	0	0	1	0
taxpc	0	1	1	1	0	0	0
pctmin80	0	0	0	0	0	0	0
pctymle	0	-1	0	0	0	1	0

The equation for model_2 is:

$$crmrate = \beta_0 + \beta_1 \cdot prbarr + \beta_2 \cdot prconv + \beta_3 \cdot density + \beta_4 \cdot taxpc + \beta_5 \cdot pctymle + \beta_6 \cdot pctymle + error$$

$$Omitted\ variables = Morals, Diet, MH, Happiness, FS, Drugs, and\ prbrc$$

As shown in the table above, the first row displays the impact the omitted variables have on the outcome variable crmrate.

###Morals omitted

$$B_1 = (-)$$

$$Morals = \alpha_0 + \alpha_1 \cdot prbarr + \alpha_2 \cdot prbcov$$

$$\alpha_1 \cdot prbarr > 0\ and\ \alpha_2 \cdot prbcov > 0$$

The OLS coeffecient will be less negative, therefore losing statistical significance.

###Mental Health (MH) Omitted

$$B_1 = (-)$$

$$Mental\ Health = \alpha_0 + \alpha_1 \cdot prbarr + \alpha_2 \cdot prbcov + \alpha_3 \cdot taxpc$$

$$\alpha_1 \cdot prbarr > 0\ and\ \alpha_2 \cdot prbcov > 0\ and\ \alpha_3 \cdot taxpc > 0$$

The OLS coeffecient will be less negative, therefore losing statistical significance.

###Happiness Omitted

$$B_1 = (-)$$

$$Happiness = \alpha_0 + \alpha_1 \cdot prbarr + \alpha_2 \cdot prbcov + \alpha_3 \cdot taxpc$$

$$\alpha_1 \cdot prbarr > 0\ and\ \alpha_2 \cdot prbcov > 0\ and\ \alpha_3 \cdot taxpc > 0$$

The OLS coeffecient will be less negative, therefore losing statistical significance.

###Family Stability (FS) Omitted

$$B_1 = (-)$$
$$Family\ Stability = \alpha_0 + \alpha_1 \cdot prbarr + \alpha_2 \cdot prbcov$$
$$\alpha_1 \cdot prbarr > 0\ and\ \alpha_2 \cdot prbcov > 0$$

The OLS coeffecient will be less negative, therefore losing statistical significance.

Drugs in area (Drugs) Omitted

$$B_1 = (+)$$

$$Drugs = \alpha_0 + \alpha_1 \cdot prbarr + \alpha_2 \cdot prbcov + \alpha_3 \cdot density + \alpha_4 \cdot pctymle$$
$$\alpha_1 \cdot prbarr < 0\ and\ \alpha_2 \cdot prbcov < 0\ and\ \alpha_3 \cdot density > 0\ and\ \alpha_4 \cdot pctymle > 0$$

The OLS coeffecient will be more positive, therefore gaining statistical significance.

Probability of Reported Crimes (prbrc) Omitted

$$B_1 = (+)$$

$$prbrc = \alpha_0 + \alpha_1 \cdot prbarr + \alpha_2 \cdot prbcov$$
$$\alpha_1 \cdot prbarr < 0\ and\ \alpha_2 \cdot prbcov < 0$$

The OLS coeffecient will be less positive, therefore losing statistical significance.

Conclusion

We found that the population density is the best predictor we have available for crime rate. However, it is unlikely that any political platform could make a direct effect to the how closely people live together. The other variable more tractable. Of the three probabilities of punishment, arrest has the highest impact on crime rate. Simply increasing police per capita does not seem to be having an effect of decreasing crime rate significantly. It may be more prudent to dedicate more resources for the existing police force so that more arrests can be made without necessarily increasing police presence.