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 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')
```

Here is a summary of the data.

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

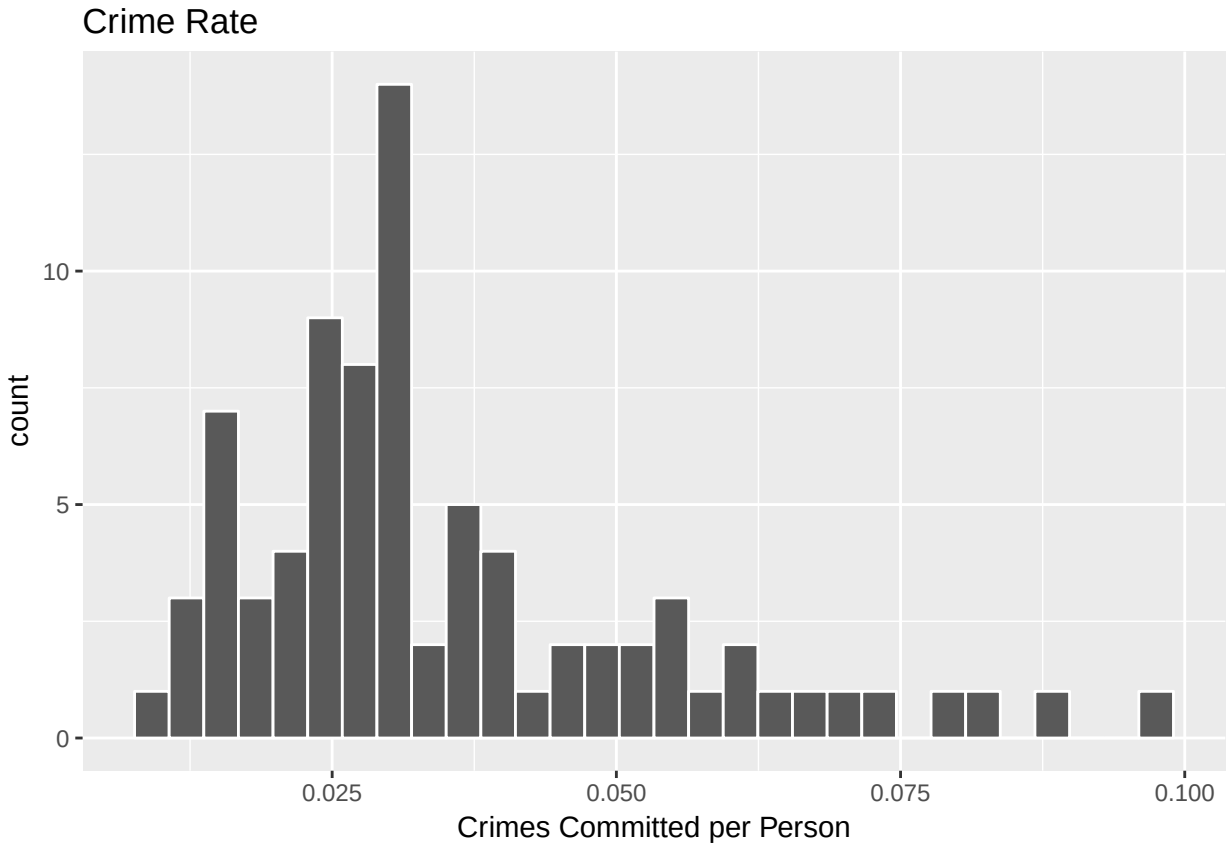
Statistic	Mean	St. Dev.	Min	Max
crmrte	0.035	0.019	0.011	0.099
prbarr	0.297	0.109	0.093	0.689
prbconv	0.448	0.172	0.068	0.973
prbpris	0.412	0.078	0.150	0.600
avgsen	9.362	2.372	5.450	17.410
polpc	0.002	0.001	0.001	0.004
density	1.508	1.580	0.00002	8.828
taxpc	38.042	13.267	25.693	119.761
pctmin80	0.258	0.168	0.015	0.619
wcon	287.879	48.018	193.643	436.767
wtuc	410.875	76.697	187.617	595.372
wtrd	213.146	34.339	154.209	354.676
wfir	322.574	50.684	234.522	509.466
wser	255.201	44.775	133.043	391.308
wmfg	335.661	85.691	157.410	646.850
wfed	445.202	61.039	326.100	597.950
wsta	359.539	42.698	267.780	499.590
wloc	312.081	28.345	239.170	388.090
mix	0.136	0.082	0.051	0.465
pctymle	0.085	0.024	0.064	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')
```

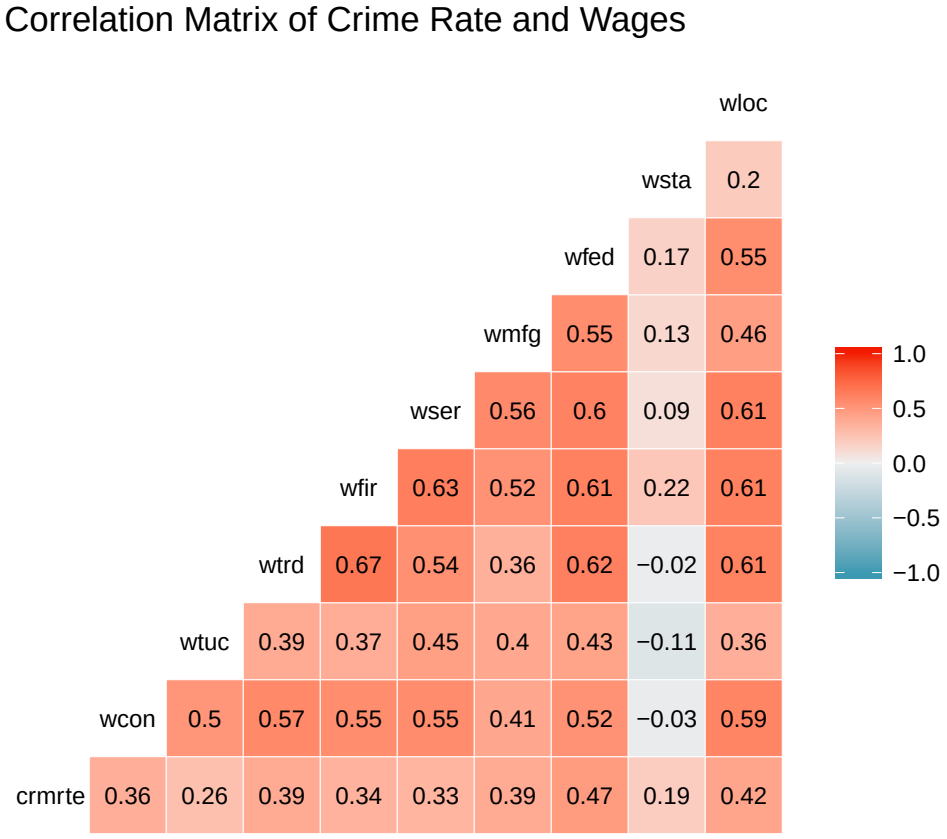



```
summary(t$crmrte)
```

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

We see that 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. First, let us treat the wage variables by themselves.

```
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')
```

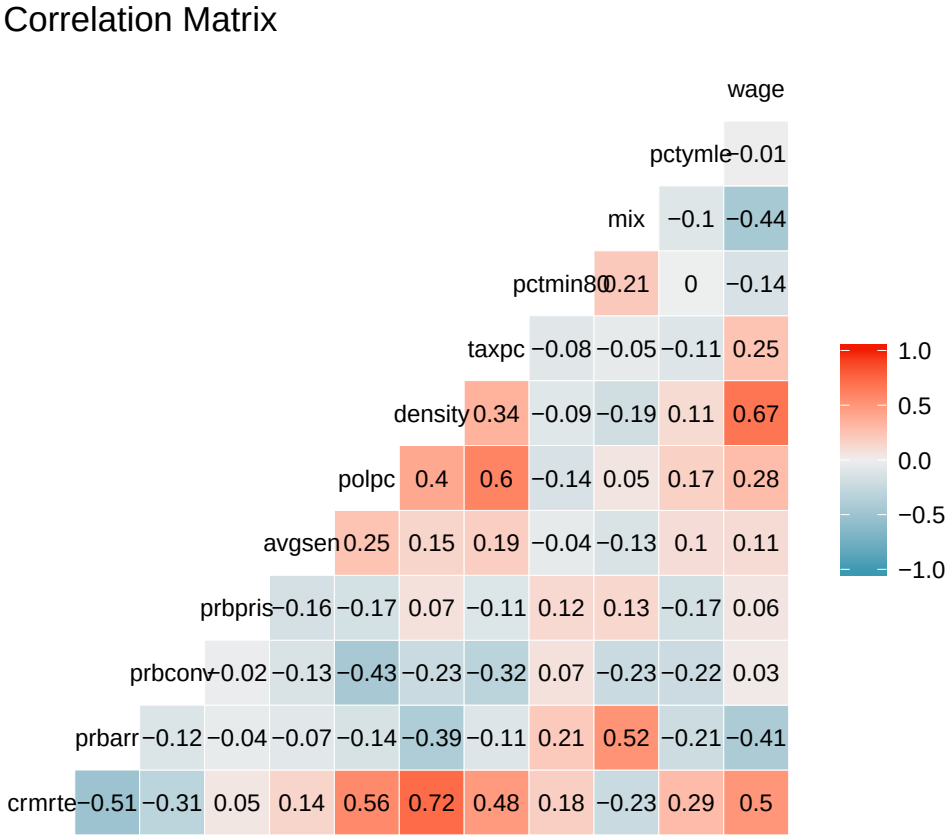


Surprisingly, we find that crime rate is actually positively correlated with all wages. This seems counter to common sentiment that crime is more prevalent in low income areas. Interestingly, we notice federal wages being the most correlated with crime rate (r = 0.47) and state wages being the least correlated (r = 0.19).

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.

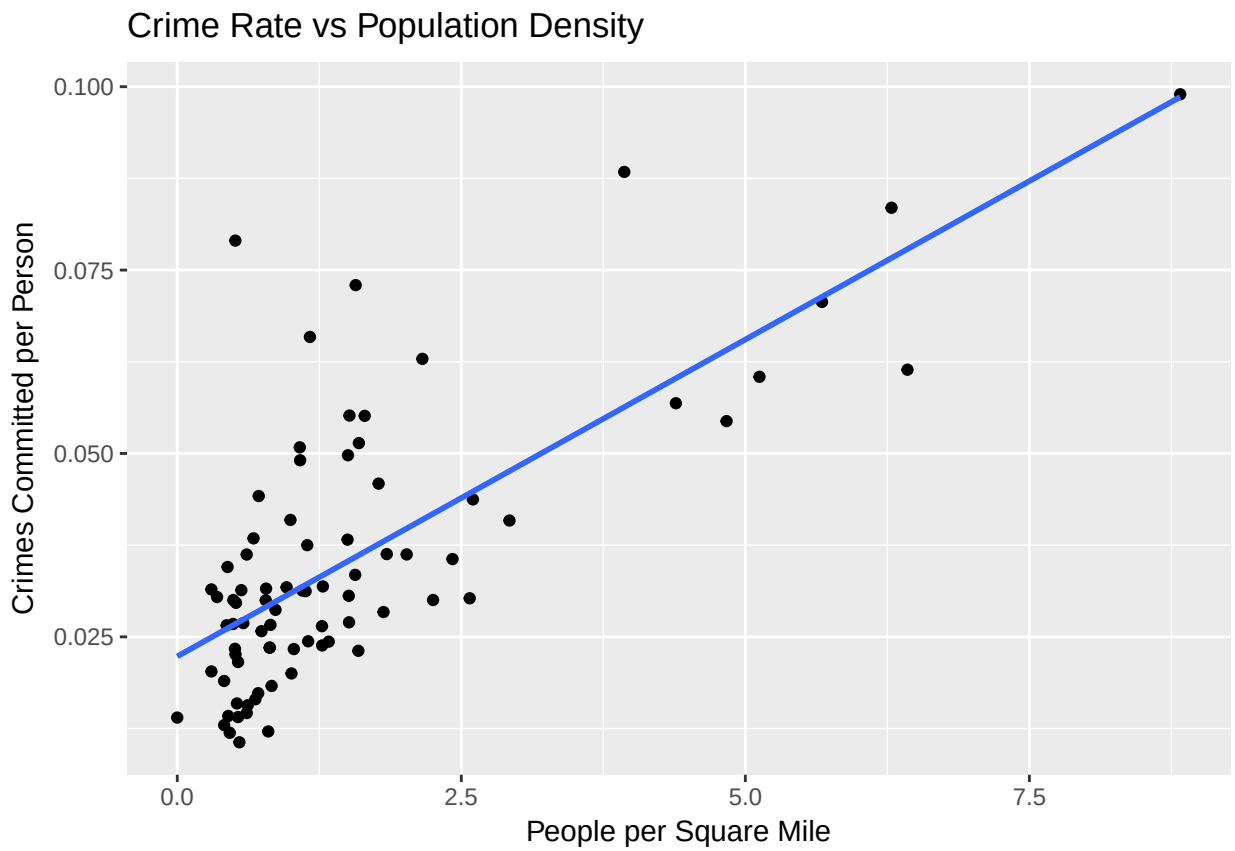
```
t = t %>% mutate(wage = wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc)

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')
```



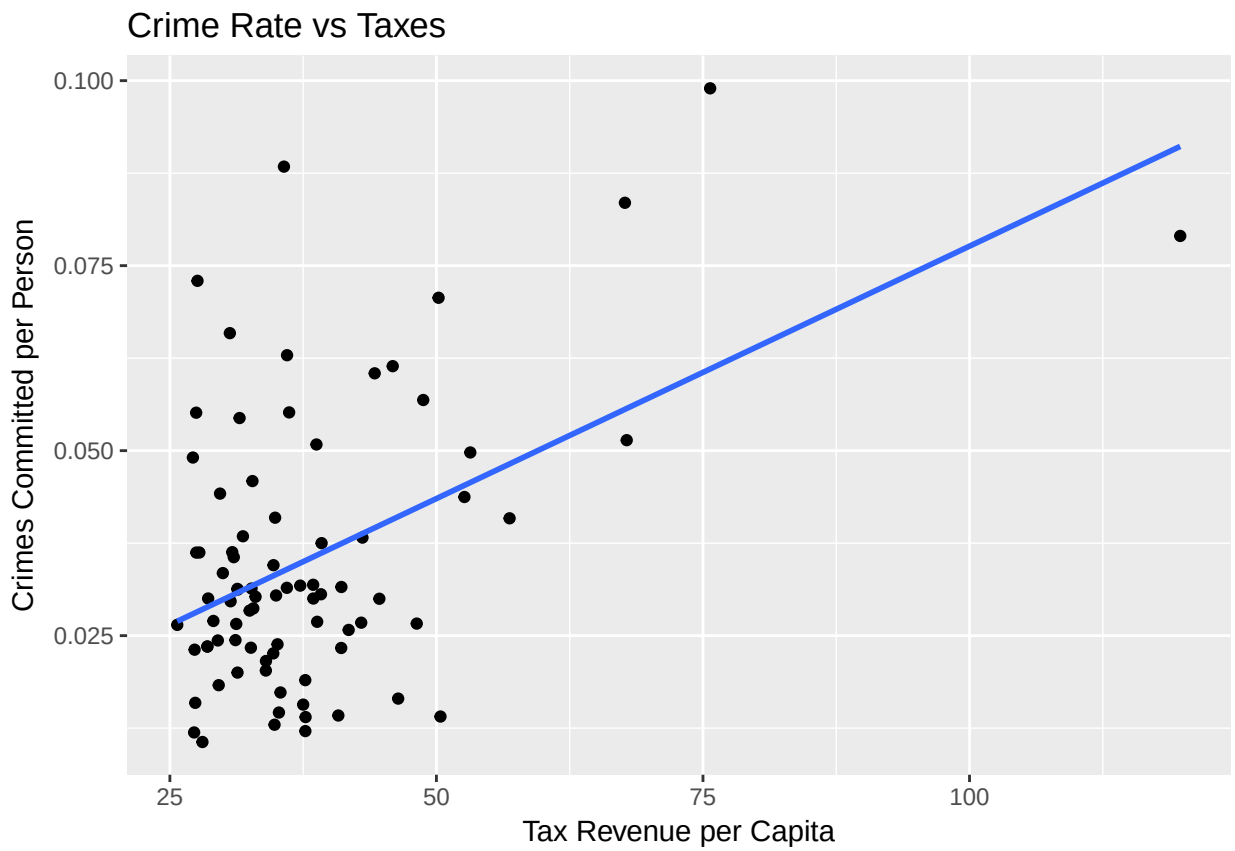
From the correlation matrix, we see that population density stands out as being highly correlated with crime rate (r = 0.72). This variable looks like a good candidate as a predictor for crime rate. One explanation could be that as more people move into an area, the increased number of interactions give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

```
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 total 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 are rising along with crime rate because they are rising along with density.

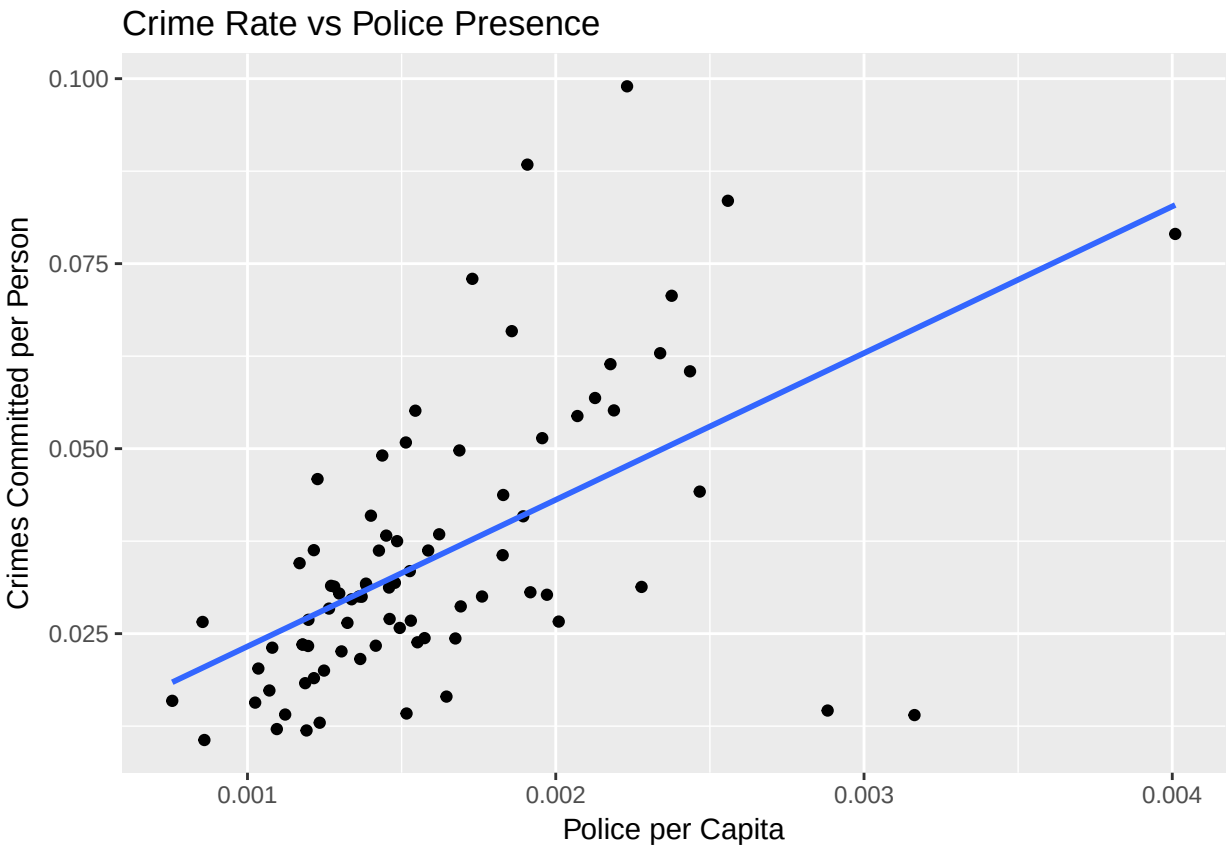
```
qplot(t$taxpc, t$crmrte) +
    labs(title = 'Crime Rate vs Taxes', x = 'Tax Revenue per Capita', y = 'Crimes Committed per Person')
    geom_smooth(method = 'lm', se = FALSE)
```




```
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)
```

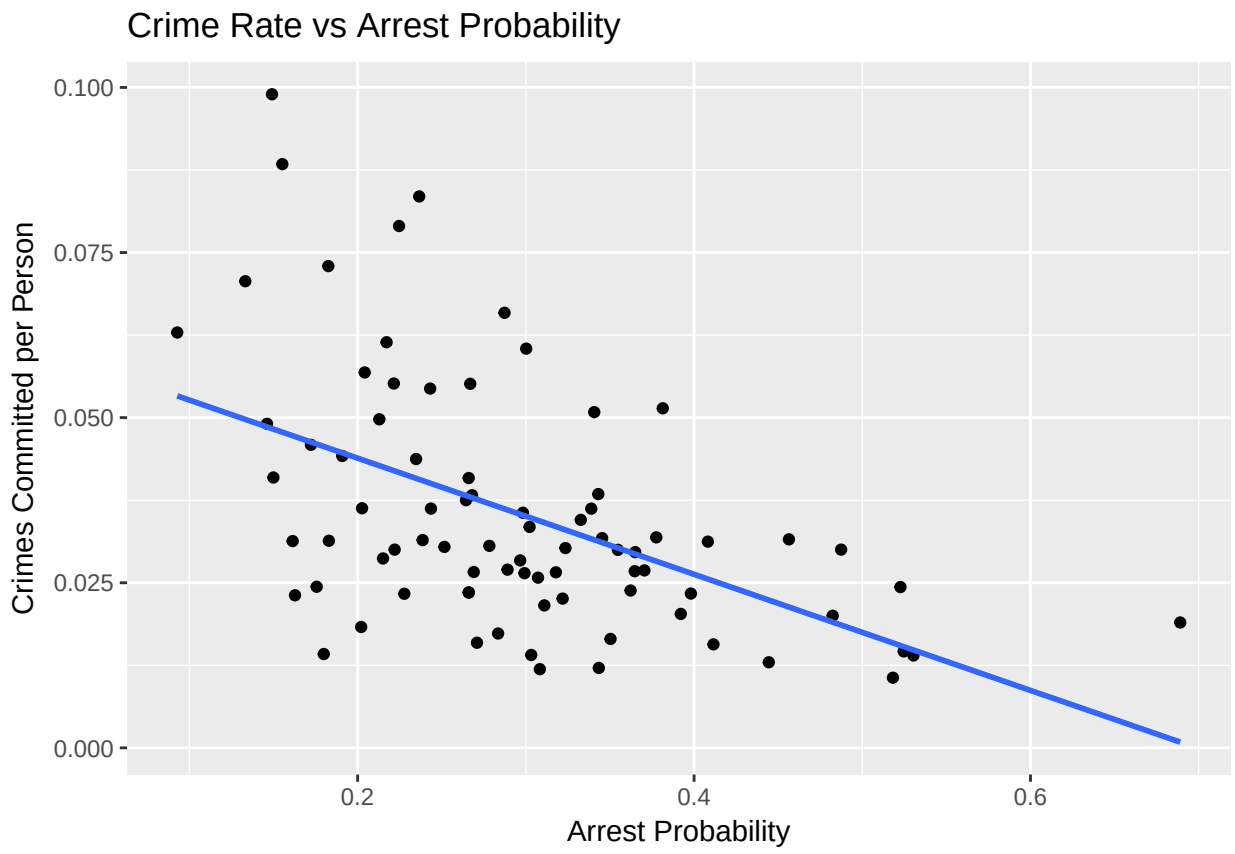
An important finding is that the relationship between police per capita and crime rate is positive and moderately large (r = 0.56). This means that either increasing police presence makes crime rate worse or that crime is causing an increase in police presence rather than vice versa. The latter explanation seems much more logical. Thus, we will not regress crime rate on police per capita, as the direction of causality is questionable.

```
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)
```

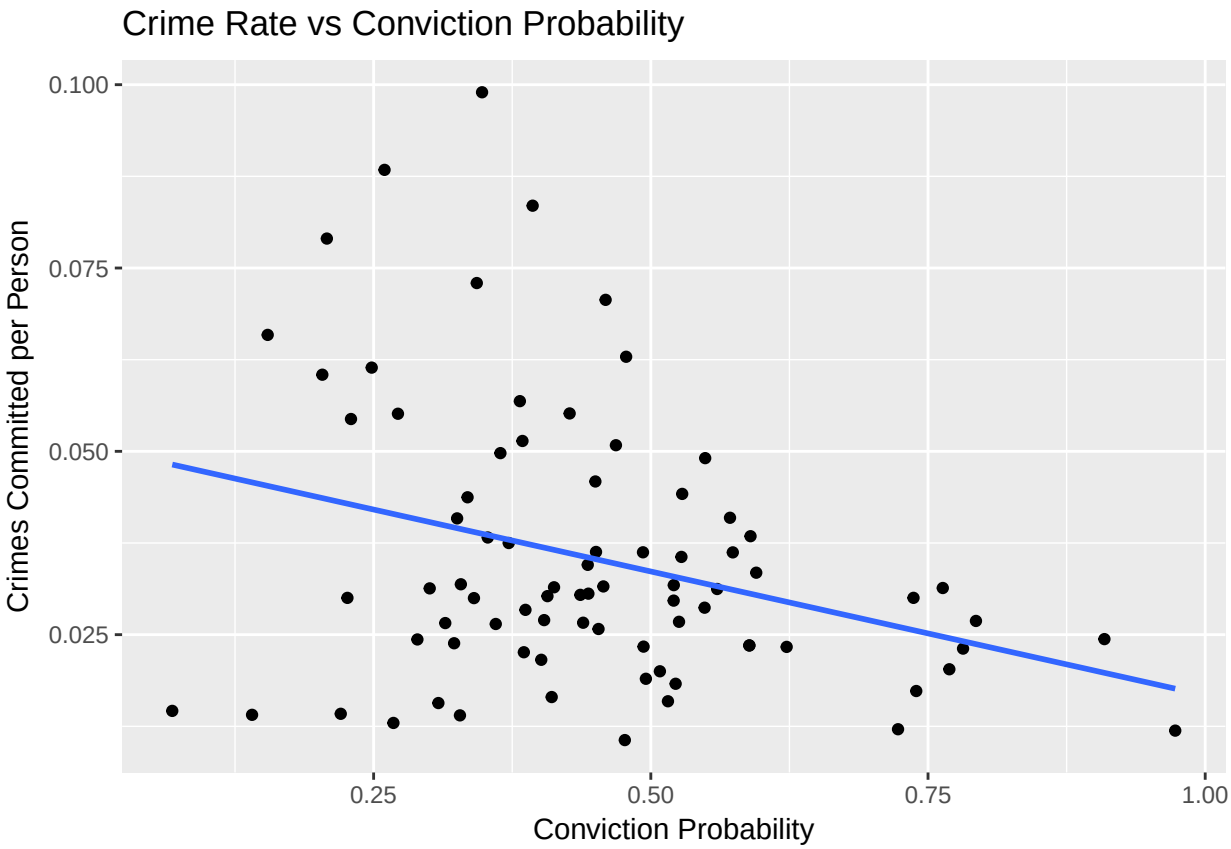


Of the three "certainty of punishment" variables, it looks like arrest probability has a moderate effect (r = -0.51) and conviction probability has a weak effect (r = -0.31), but probability of prison sentence has almost no effect (r = 0.05). It is important to note that these three probabilities seem uncorrelated with one another, so we 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 = 0.14).

```
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)
```

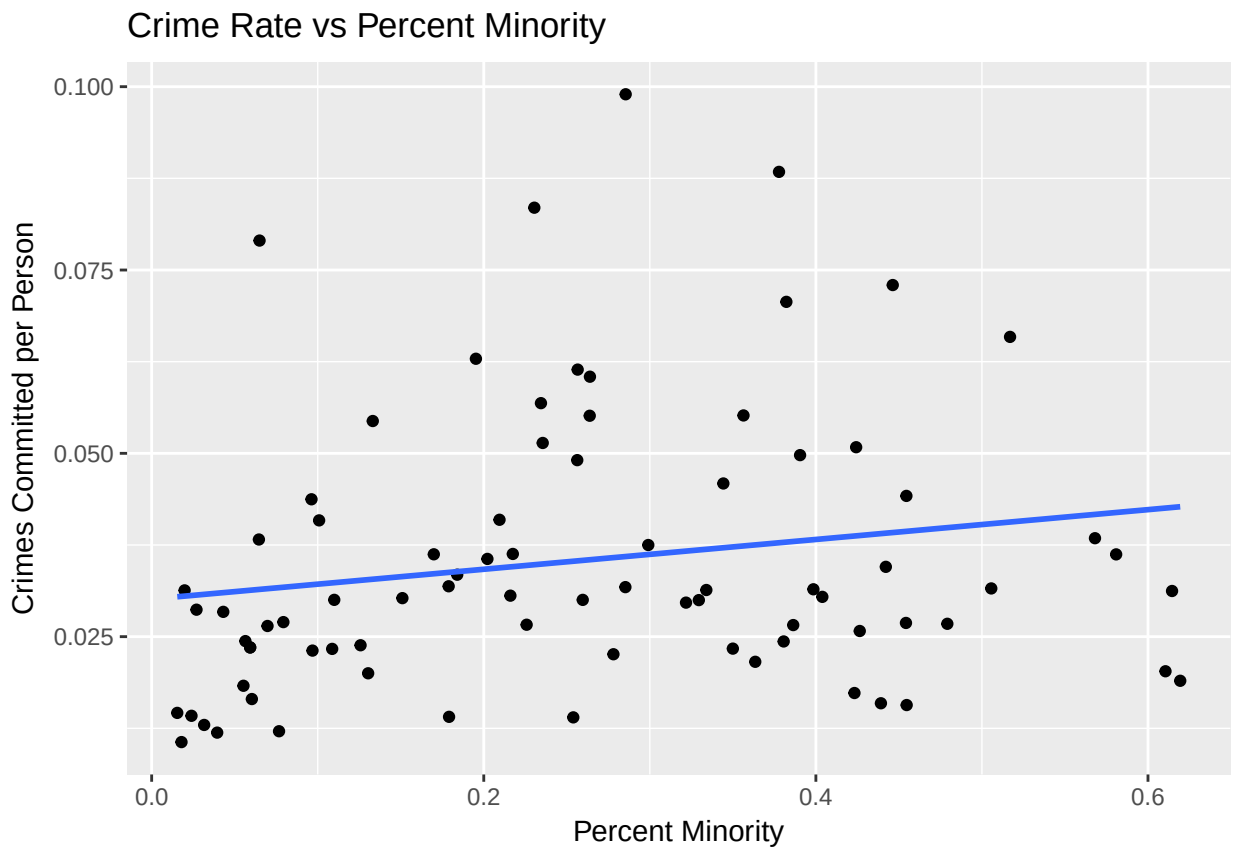



```
qplot(t$prbconv, t$crmrte) +
    labs(title = 'Crime Rate vs Conviction Probability', x = 'Conviction Probability', y = 'Crimes Comm
    geom_smooth(method = 'lm', se = FALSE)
```

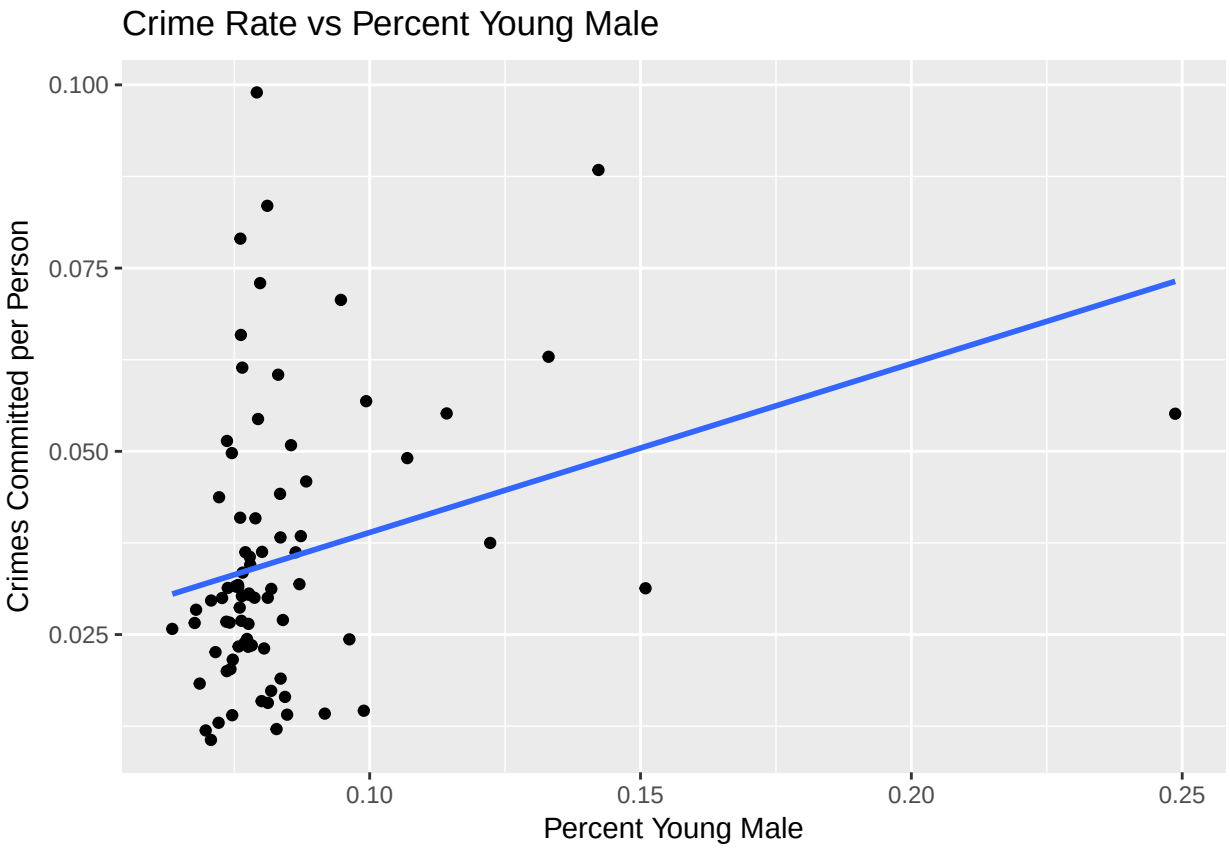


Finally, the two demographic variables seem to have relatively weak correlations with crime rate. However, their directions are at least in line with historic sentiment (young male minorities are commonly associated with crime).

```
qplot(t$pctmin80, t$crmrte) +
    labs(title = 'Crime Rate vs Percent Minority', x = 'Percent Minority', y = 'Crimes Committed per Pe
    geom_smooth(method = 'lm', se = FALSE)
```



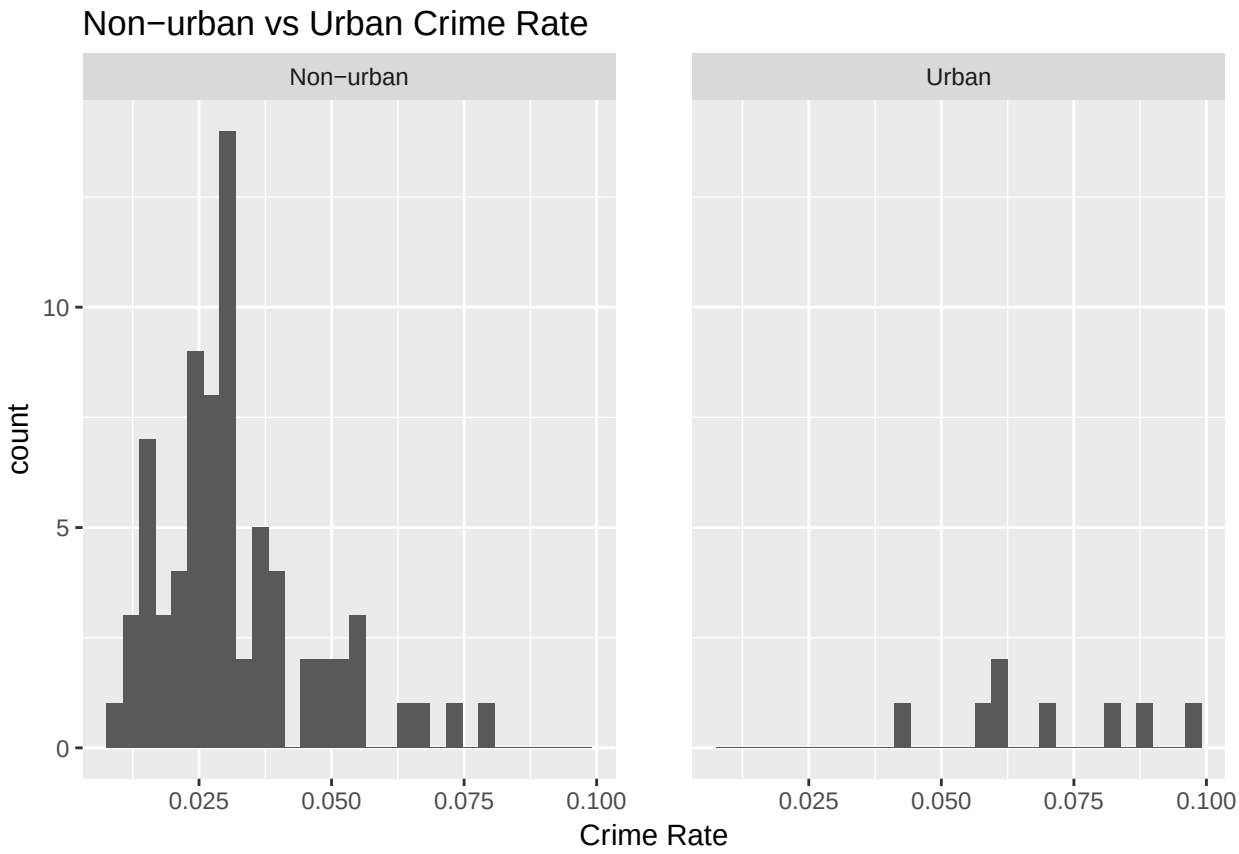

```
qplot(t$pctymle, t$crmrte) +
    labs(title = 'Crime Rate vs Percent Young Male', x = 'Percent Young Male', y = 'Crimes Committed pe
    geom_smooth(method = 'lm', se = FALSE)
```



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.

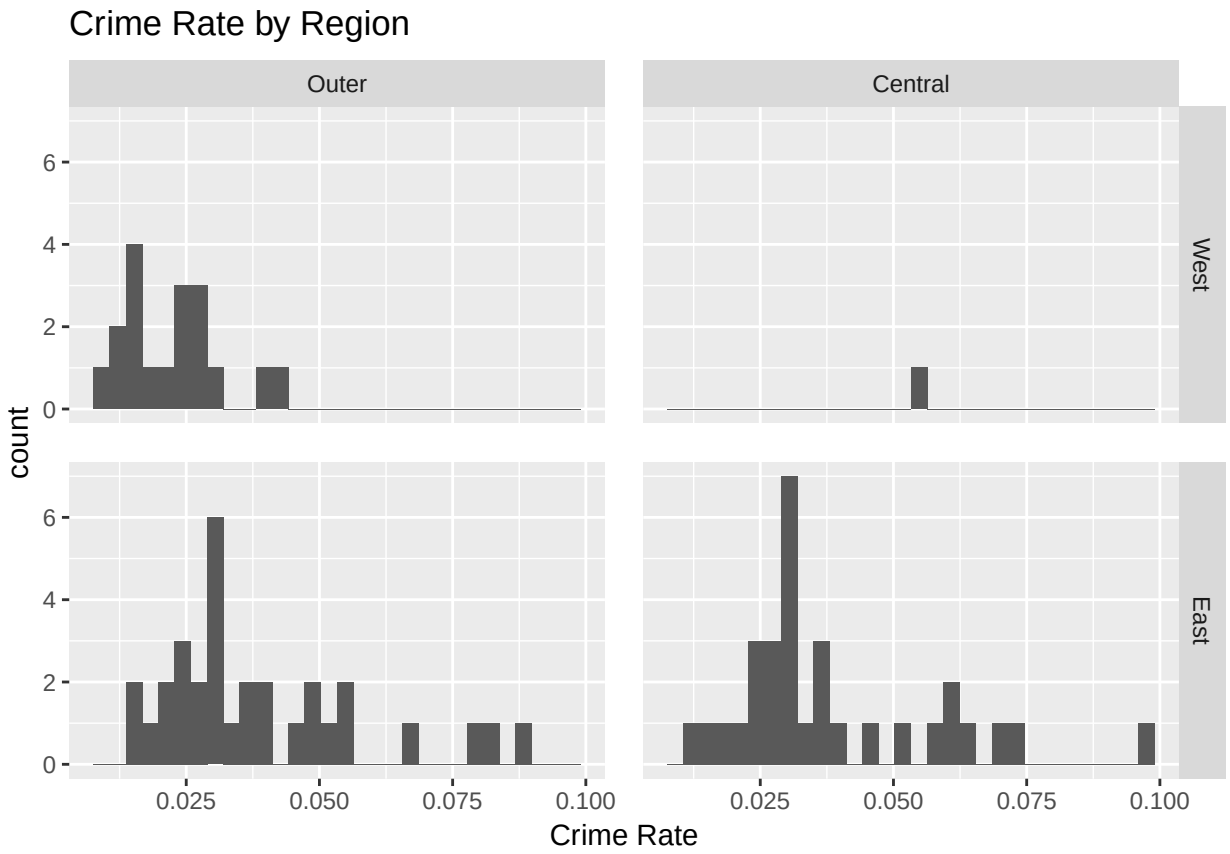
```
ggplot(t, aes(crmrte)) +
    geom_histogram() +
    facet_grid(. ~ urban) +
    theme(panel.spacing = unit(2, "lines")) +
    labs(title = 'Non-urban vs Urban Crime Rate', x = 'Crime Rate')
```



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 19 western counties, with a mere single county in the wester 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 = -1.406, df = 63.25, p-value = 0.1646
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01469592  0.00255640
## sample estimates:
##  mean of x  mean of y
## 0.03296647 0.03903623
```

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.

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		0.00004	
		(0.0001)	
Trans, Util, Commun		0.00001	
		(0.00003)	
Whlesle, Retail, Trade		0.0001	
		(0.0001)	
Fin, Ins, Real Est		−0.0001	
		(0.0001)	
Service		−0.00004	
		(0.0001)	
Manuacturing		0.00004	
		(0.00004)	
Federal	0.0001***	0.0001	
	(0.00004)	(0.0001)	
State		0.0001	
		(0.00004)	
Local		0.0001	
		(0.0001)	
Total Sum			0.00003***
			(0.00001)
Constant	−0.030	−0.071*	−0.048*
	(0.019)	(0.028)	(0.022)
R^2	0.224	0.305	0.249
Adjusted R^2	0.215	0.217	0.239
Residual Std. Error	0.017 (df = 79)	0.017 (df = 71)	0.016 (df = 79)
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***	0.006***	0.005**
	(0.001)	(0.001)	(0.002)
Arrest Probability		−0.059***	−0.054***
		(0.013)	(0.012)
Conviction Probability	−0.016	−0.015	−0.017
	(0.010)	(0.009)	(0.011)
Prison Probability			0.005
			(0.014)
Average Prison Sentence			−0.0002
			(0.0005)
Tax per Capita		0.0004	0.0004
		(0.0003)	(0.0003)
Percent Minority		0.036***	0.037***
		(0.007)	(0.008)
Offense Mix			−0.012
			(0.017)
Percent Young Male		0.135*	0.140*
		(0.056)	(0.061)
Sum of Wages			0.00000
			(0.00001)
Constant	0.030***	0.015	0.011
	(0.006)	(0.015)	(0.029)
R^2	0.545	0.801	0.805
Adjusted R^2	0.534	0.785	0.777
Residual Std. Error	0.013 (df = 78)	0.009 (df = 74)	0.009 (df = 70)
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 the bias that each omitted variable has on both the measured variables, and the output (crmrate).

Omitted Variable	Morals	Diet	MH	Happiness	FS	Drugs	prbrc
crmrate	-1	0	-1	-1	-1	1	1
prbarr	-1	0	-1	-1	-1	1	1
prbconv	-1	0	-1	-1	-1	1	1
prbpris	-1	0	-1	-1	-1	1	1
avgsen	-1	0	-1	-1	-1	1	1
polpc	0	0	0	0	0	1	0
density	0	-1	0	0	0	1	0
taxpc	0	1	1	1	0	0	0
pctmin80	0	0	0	0	0	0	0
wage	0	1	1	1	1	-1	1
mix	0	0	-1	-1	-1	1	0
pctymle	0	-1	0	0	0	1	0

Conclusion