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', header = FALSE, float = FALSE)
```

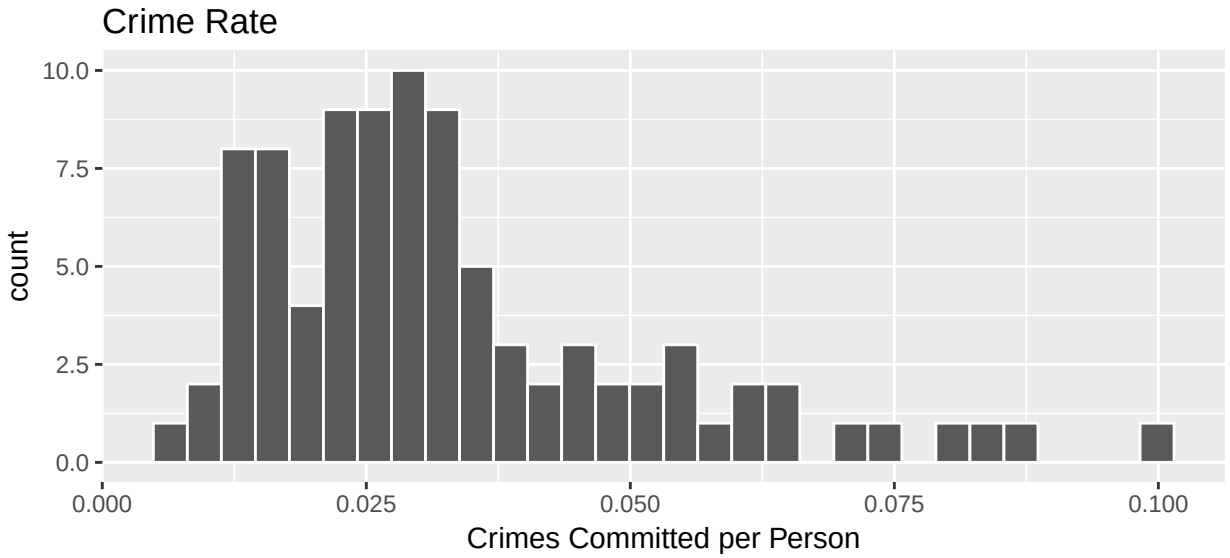
Statistic	N	Mean	St. Dev.	Min	Max
crmrte	91	0.033	0.019	0.006	0.099
prbarr	91	0.295	0.137	0.093	1.091
prbconv	91	0.551	0.352	0.068	2.121
prbpris	91	0.411	0.080	0.150	0.600
avgsen	91	9.647	2.847	5.380	20.700
polpc	91	0.002	0.001	0.001	0.009
density	91	1.429	1.514	0.00002	8.828
taxpc	91	38.055	13.078	25.693	119.761
pctmin80	91	0.255	0.170	0.013	0.643
wcon	91	285.358	47.487	193.643	436.767
wtuc	91	411.668	77.266	187.617	613.226
wtrd	91	211.553	34.216	154.209	354.676
wfir	91	322.098	53.890	170.940	509.466
wser	91	275.564	206.251	133.043	2,177.068
wmfg	91	335.589	87.841	157.410	646.850
wfed	91	442.901	59.678	326.100	597.950
wsta	91	357.522	43.103	258.330	499.590
wloc	91	312.681	28.235	239.170	388.090
mix	91	0.129	0.081	0.020	0.465
pctymle	91	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.

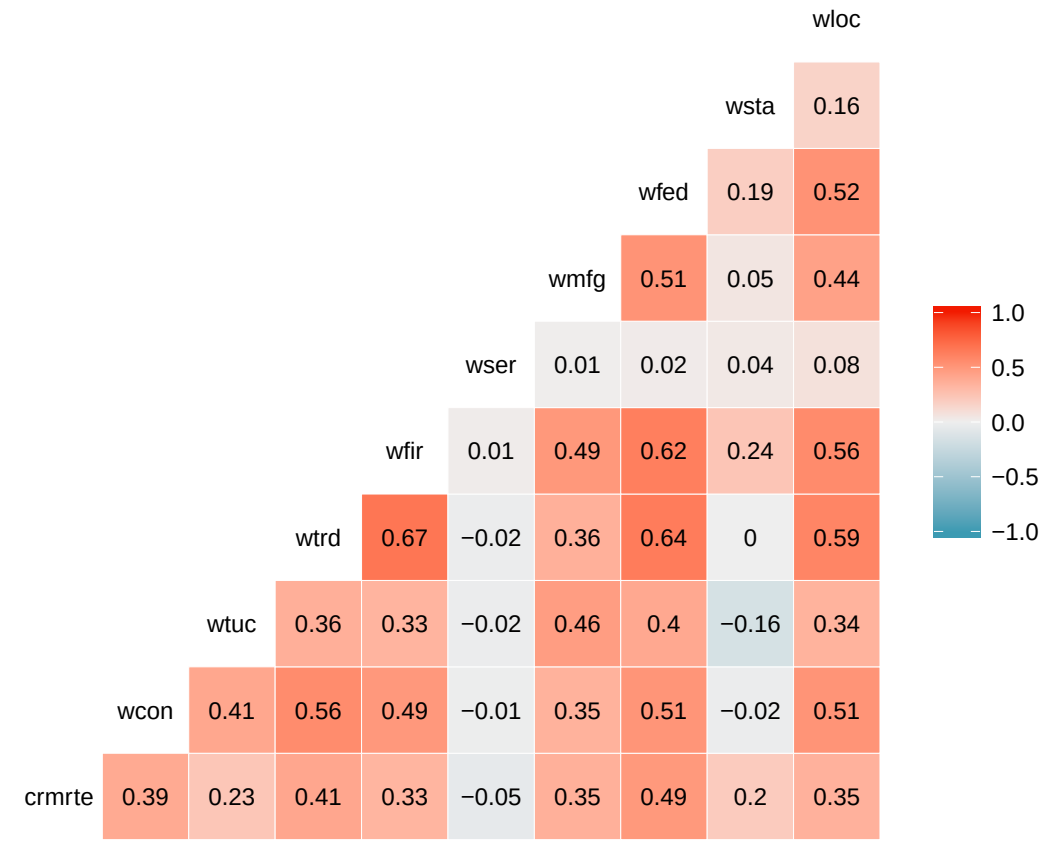
```
qplot(t$crmrte, col = I('white')) +
    labs(title = 'Crime Rate', x = 'Crimes Committed per Person')
```



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')
```



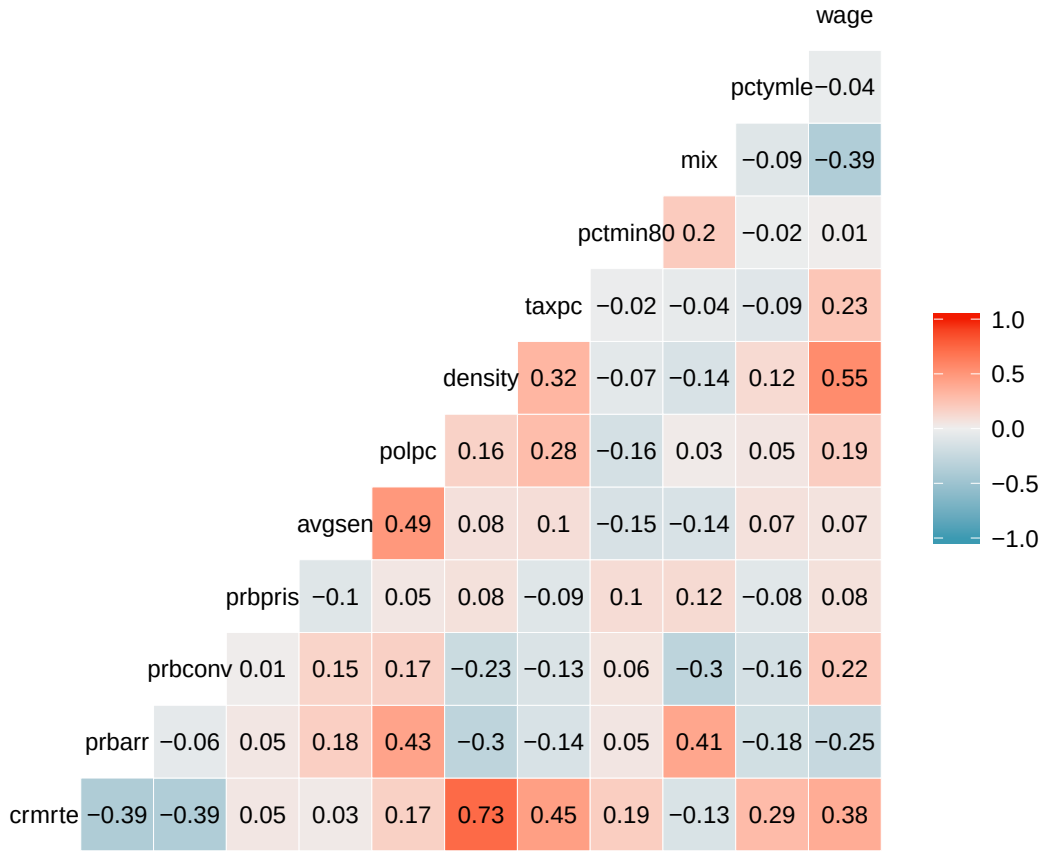
Surprisingly, we find that crime rate is actually positively correlated with all wages except from the service industry. This seems counter to common sentiment that crime is more prevalent in low income areas.

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')
```

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.

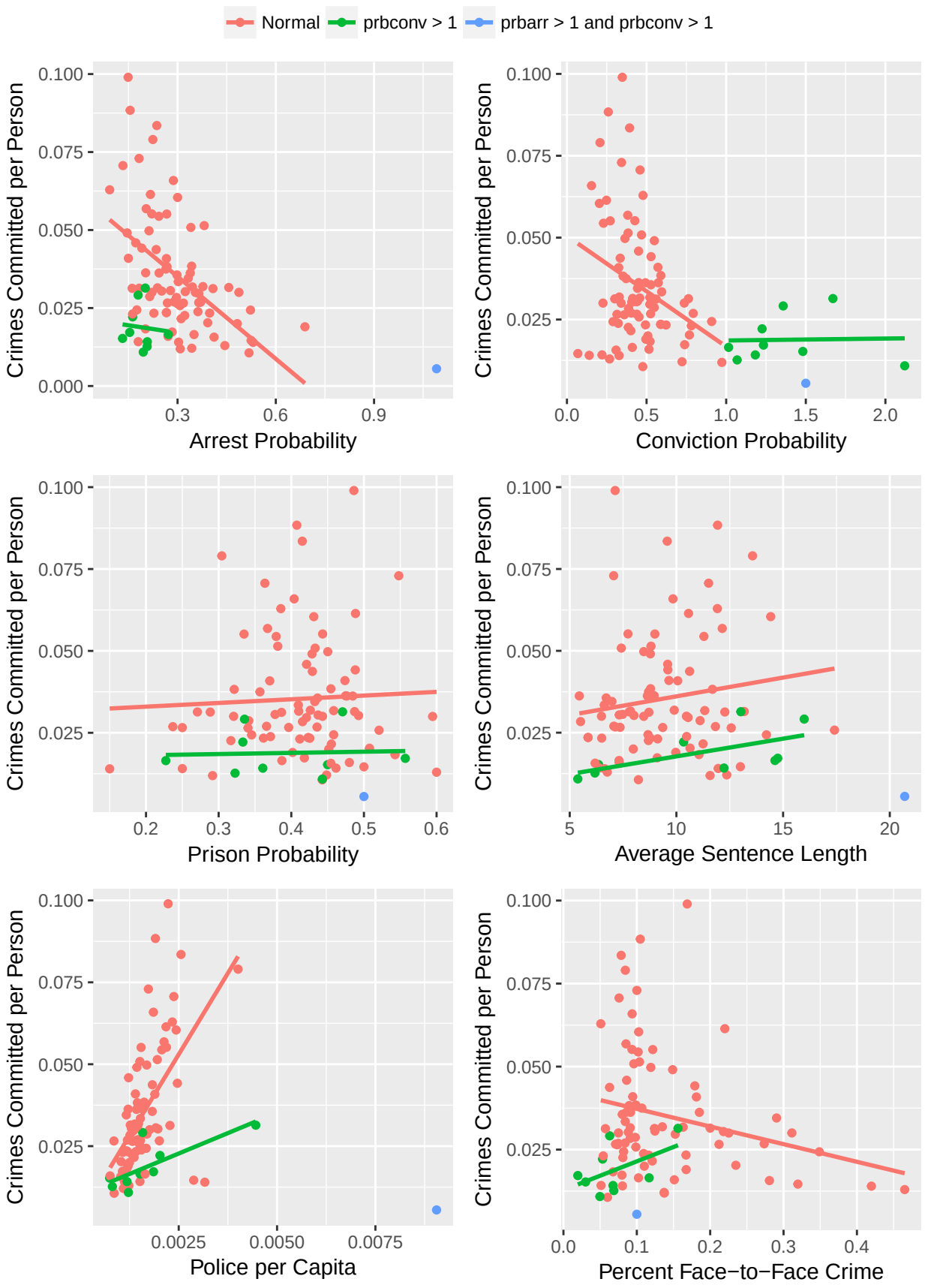
Of the three "certainty of punishment" variables, it looks like arrest probability has a moderate effect (r = -0.39) and conviction probability has a weak effect (r = -0.39), but probability of prison sentence has almost no effect (r = 0.05). It is important to note that these three probabilities seem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

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)
```

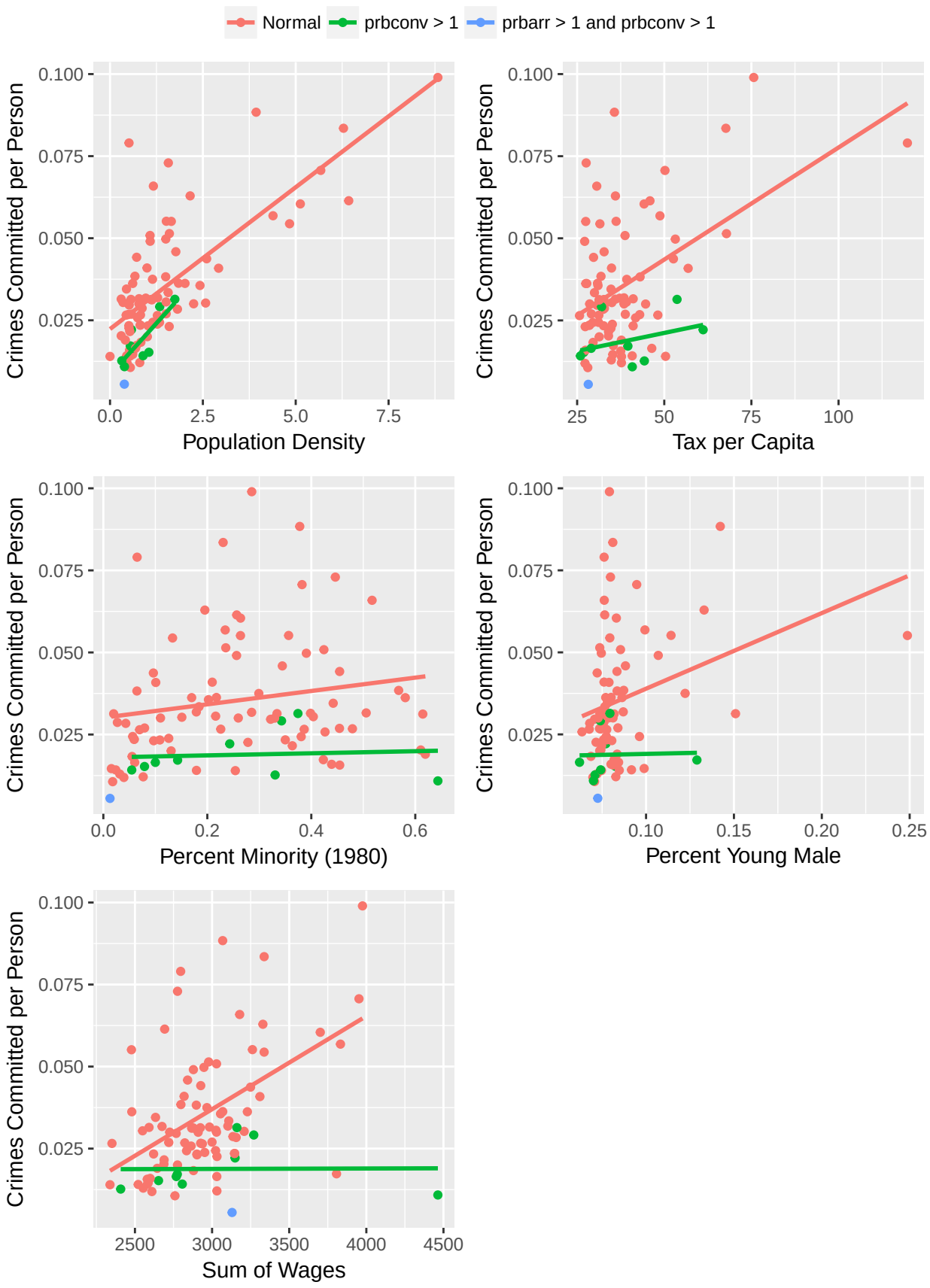

Normal
prbconv > 1
prbarr > 1 and prbconv > 1
Crimes Committed per Person
0.100
0.075
0.050
0.025
0.000
0.3
0.6
0.9
Arrest Probability
0.0
0.5
1.0
1.5
2.0
Conviction Probability
0.2
0.3
0.4
0.5
0.6
Prison Probability
5
10
15
20
Average Sentence Length
0.0025
0.0050
0.0075
Police per Capita
0.0
0.1
0.2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

Immediately, we notice that the counties with anomalous conviction probabilities are problematic. Instead of being uniformly distributed among the data, they are all very good in terms of having low crime rate. County 115 (Madison County[1]) is the point with both arrest probability and conviction probability greater than 1. It has the lowest crime rate among all counties. Madison County also has the longest average prison sentence length and most police per capita by far. In essence, Madison County seems like one in which law enforcement is extremely strict and has very low crime rate as a result.

Socioeconomic/Demographic Variables

```
p7 = qplot(t$density, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Population Density', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p8 = qplot(t$taxpc, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Tax per Capita', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p9 = qplot(t$pctmin80, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Percent Minority (1980)', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p10 = qplot(t$pctymle, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Percent Young Male', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
p11 = qplot(t$wage, t$crmrte, col = t$badprb) +
    labs(title = NULL, col = NULL, x = 'Sum of Wages', y = 'Crimes Committed per Person') +
    geom_smooth(method = 'lm', se = FALSE)
ggarrange(p7, p8, p9, p10, p11, nrow = 3, ncol = 2, common.legend = T)
```

[1]We assume the county numbers in the dataset are Federal Processing Standard Publication (FIPS) numbers. The North Carolina FIPS codes were found at https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina.

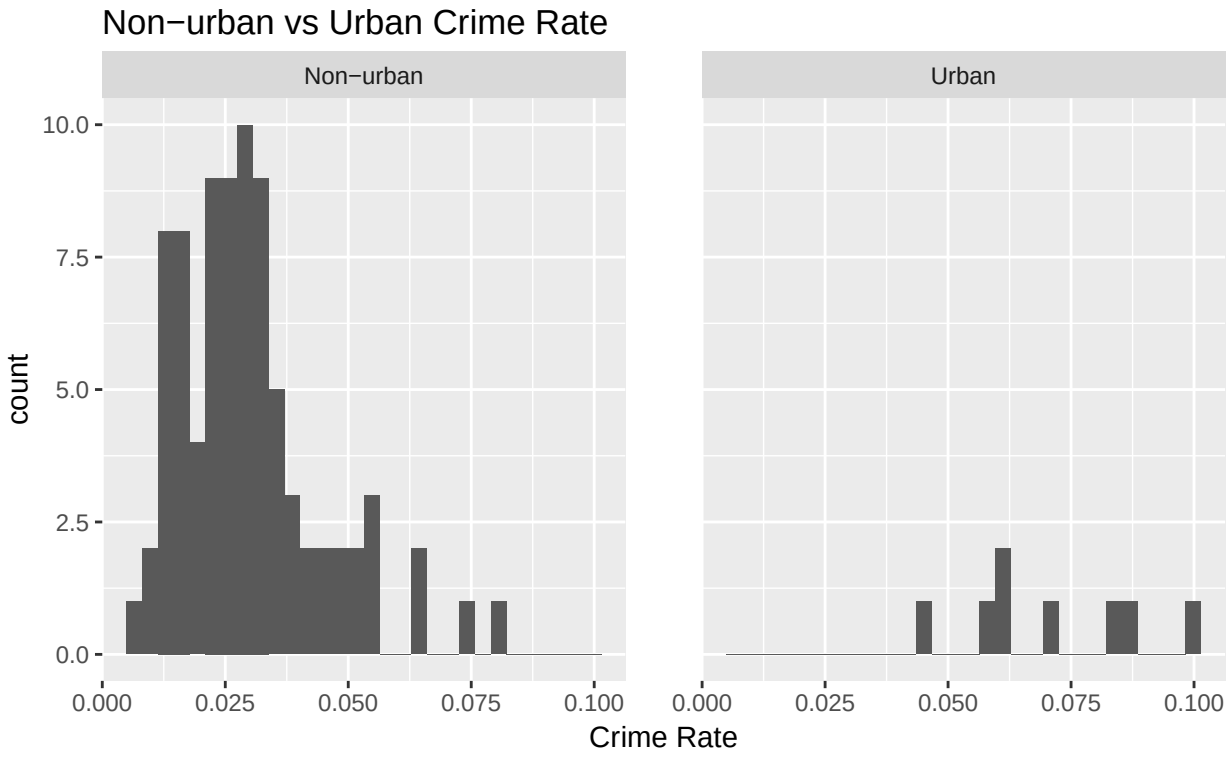

Normal
prbconv > 1
prbarr > 1 and prbconv > 1
Crimes Committed per Person
0.100
0.075
0.050
0.025
0.0
2.5
5.0
7.5
Population Density
25
50
75
100
Tax per Capita
0.2
0.4
0.6
Percent Minority (1980)
0.10
0.15
0.20
0.25
Percent Young Male
2500
3000
3500
4000
4500
Sum of Wages

Again, we notice the problematic high punishment probability counties. This time, we find Madison County is very low in population density, tax per capita, minority population, and percent young male. Indeed, the county is sparsely populated and located in the Appalachian Mountains, within heavy forests[2].

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')
```

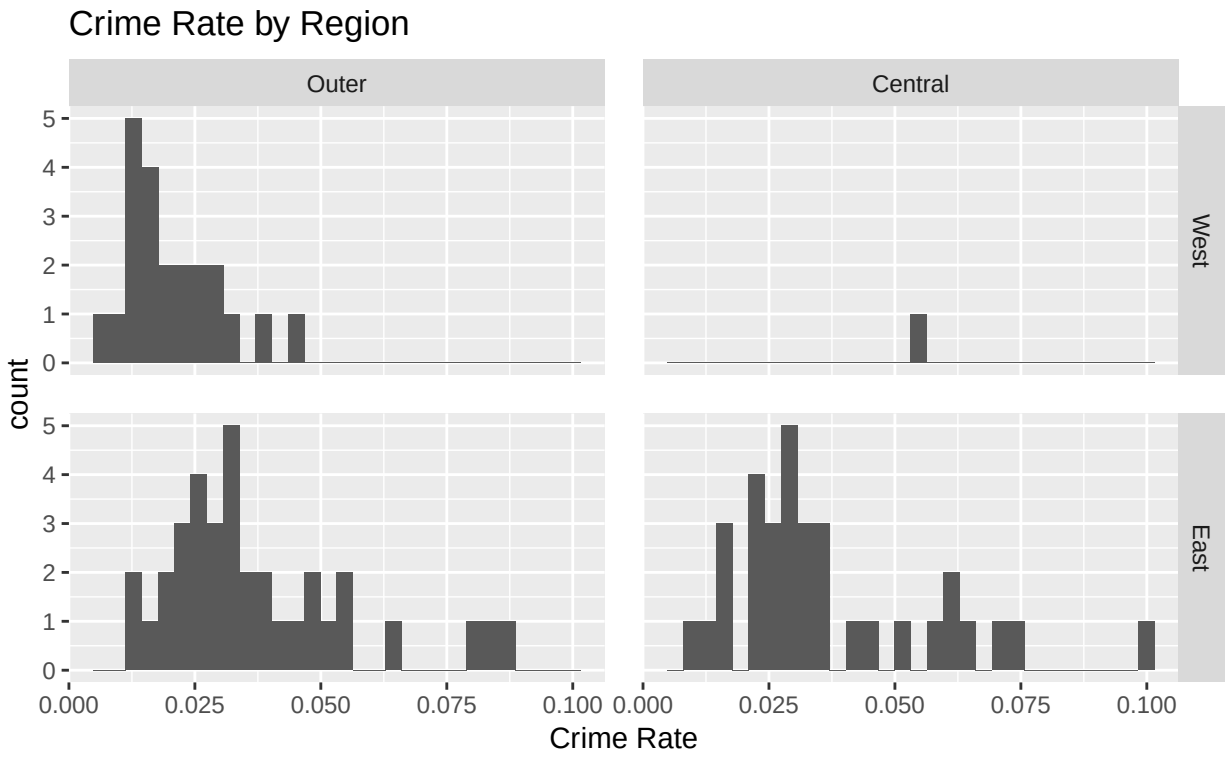


We see that there are only 8 counties coded as urban, which is probably too few to make any sweeping inferences. We will only mention in passing that the crime rate in urban counties does look higher than that in non-urban counties.

Next we examine the differences in geographic region. Here, we assume the 'west' and 'central' variables are non-exclusive. That is, we assume a county can be both a western county and a central county.

```
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')
```

[2]This information from https://en.wikipedia.org/wiki/Madison_County,_North_Carolina.



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.

EDA Summary and Research Questions for Model Building

From our EDA, we think that the best predictor for crime rate is population density. We see that arrest probability is also a decent predictor. These two constitute one law-enforcement variable and one demographic variable. Therefore, we will start by building models with these two predictors and then augment with other variables.

There does not seem to be strong enough evidence to investigate the geographical effects (west vs east, central vs outer, etc.).

We also found in our EDA that there are several very problematic with counties with conviction probabilities greater than 1. From this point forward, we will proceed by removing them from our data analysis (i.e. using t2 for our regressions). The rationale behind the omission are as follows:

1. These counties are obviously special in some way. From the bivariate scatterplots, it is apparent that they tend to cluster low on the y-axis.
2. These counties make up a small number of the data set.
3. These counties have low population density (with Madison County being the most sparsely populated county in the whole data set). With the fewer number of people, it is possible that data collection was more prone to error.
4. Fundamentally, probabilities greater than 1 simply make no sense.

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

Before investigating the main regression models, we first 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(t2$crmrte ~ t2$wfed)
m2_wage = lm(t2$crmrte ~ t2$wcon + t2$wtuc + t2$wtrd + t2$wfir + t2$wser + t2$wmfg +
                t2$wfed + t2$wsta + t2$wloc)
m3_wage = lm(t2$crmrte ~ t2$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 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.

With respect to practical significance, none of the wage models may be that influential. The coefficient on total sum of wages is 2.8×10^{-5}. A rise of $1,000 in weekly wages would only cause 0.028 extra crimes per person.

Main Models

Now we will proceed to build models with all the other variables. We start with our base model, involving only population density and arrest probability.

```
m1 = lm(t2$crmrte ~ t2$density + t2$prbarr)
coeftest(m1, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.03777214  0.00562968  6.7095 2.798e-09 ***
## t2$density   0.00741231  0.00086989  8.5210 9.184e-13 ***
## t2$prbarr   -0.04579393  0.01366522 -3.3511  0.001243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find high statistical significance for the arrest probability coefficient; very high significance for the population density coefficient. Each unit increase in population density seems to bring 0.007 crimes with him or her. Each percent increase in arrest probability takes away 0.0005 crimes. We note that these numbers are a little difficult to comprehend practically. It is nonintuitive how big an effect of 0.0005 crimes. Thus, we propose a second model, where we instead regress the natural logarithm of crime rate ($\ln(crmrte)$). The interpretation of the new output variable is the proportional change in crime rate with respect to differential changes in the input variables. This interpretation makes more intuitive sense (e.g. a 10% increase in crime rate is easier to grasp than 5/1000th of a crime).

```
m2 = lm(log(t2$crmrte) ~ t2$density + t2$prbarr)
coeftest(m2, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.269072   0.161070 -20.2959 < 2.2e-16 ***
## t2$density   0.165723   0.024801   6.6822  3.15e-09 ***
## t2$prbarr   -1.516894   0.420973  -3.6033 0.0005511 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this new model, we now have very high statistical significance in all of our coefficients. It is also at this point that we notice a unit increase in density causes a 16.5% decrease in crime rate. This alarmingly high percentage prompts us to check on the units of population density (something glossed over in our initial EDA). In the dataset, county 119 (Mecklenburg County) has the maximum density at 8.83. In 2018, we see that Mecklenburg has population 944,373 in 546 square miles, or about 1730 people per square mile[3]. Since the data is older, we will assume that density is measured in people X100. Now we see that a single person causes a 0.165% increase in crime rate (more reasonable).

We also see that a percent increase in arrest probability generally causes a 1.52% decrease in crime. Obviously arrest probability can't increase ad infinitum, so the effect cannot continue forever (yet another reason to eliminate the questionable probabilities from earlier). We can consider applying a logit transformation to arrest probability to confine the variable within 0 to 1, but we lose a lot of the interprebality of the arrest probability variable–it is easier to talk about an arrest probability of 50% than to talk about a log-odds of 0. Still, we will very briefly investigate to see if a logit transformation at least improves our fit.

[3]Population data from https://en.wikipedia.org/wiki/List_of_counties_in_North_Carolina.

```
logit = function(x) log(x/(1-x))
summary(m2)$r.square
```

```
## [1] 0.50607
```

```
summary(lm(log(t2$crmrte) ~ t2$density + logit(t2$prbarr)))$r.square
```

```
## [1] 0.506989
```

We find virtually no difference with the logit transformed probability and will forego it in favor of ease of interprebability. Now we return to the diagnostic plots of model 2.

Since $\log(crmrte)$ is a bit easier to interpret than straight number of crimes per person, we will continue building our models with it.

In our third model, we will add in several covariates with density and arrest probability but will not include ones highly correlated with density, arrest probability, or themselves. This is to minimize the absorption of any causal effect. The variables we add are

- Conviction probability
- Prison probability
- Average sentence probability
- Percent minority
- Percent young male

We have added all the law-enforcement variables with the exception of police per capita. As explained above, we think police per capita is responding to crime rate and not the other way around. We also added the two demographic variables, but did not add the two economic ones. This latter omission is due to both taxes and wages being correlated with population density. Our rationale is explained further in the EDA section. We also do not include the mix variable due its high correlation with arrest probability.

```
m3 = lm(log(t2$crmrte) ~ t2$density + t2$prbarr + t2$prbconv + t2$prbpris +
           t2$avgsen + t2$pctmin80 + t2$pctymle)
coeftest(m3, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -3.08845012  0.50784992 -6.0814 4.983e-08 ***
## t2$density   0.14286564  0.03229809  4.4233 3.331e-05 ***
## t2$prbarr   -2.06900334  0.41103896 -5.0336 3.350e-06 ***
## t2$prbconv  -0.66509280  0.32494121 -2.0468   0.04428 *
## t2$prbpris  -0.31896202  0.53564966 -0.5955   0.55337
## t2$avgsen   -0.00081337  0.01633267 -0.0498   0.96042
## t2$pctmin80  1.15926626  0.18826379  6.1577 3.633e-08 ***
## t2$pctymle   1.84588508  1.22223406  1.5103   0.13530
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that both of our original variables retain their high degree of statistical signifance. The two newly added variables that are statistically significant are conviction probability and minority percentage. The former results in a 0.66% decrease in crime rate for each percent increase; the latter results in a 1.16% increase in crime rate for each percent increase.

In practical perspective, arrest probability, percent young male, and percent minority have relatively higher practical significances among the variables in changing crime rates. As probability arrest increases by 1%, crime rate falls by 2.1%, and as percentage of young male increases by 1%, crime rate increases by 1.9%. As percentage of minority increases by 1%, crime rate goes up by 1.16%.

For our last model, we include the remaining covariates that were excluded due to correlation with existing covariates.

```
m4 = lm(log(t2$crmrte) ~ t2$density + t2$prbarr + t2$prbconv + t2$prbpris + t2$avgsen
        + t2$pctmin80 + t2$pctymle + t2$taxpc + t2$mix + t2$wage)
coeftest(m4, vcov. = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -3.91455252  1.02275799 -3.8274 0.0002787 ***
## t2$density   0.10888008  0.04874831  2.2335 0.0287128 *
## t2$prbarr   -1.75016501  0.34033055 -5.1425 2.358e-06 ***
## t2$prbconv  -0.56508986  0.30510021 -1.8521 0.0682218 .
## t2$prbpris  -0.03460577  0.43077906 -0.0803 0.9362016
## t2$avgsen   -0.00718878  0.01566980 -0.4588 0.6478231
## t2$pctmin80  1.19822214  0.20382753  5.8786 1.281e-07 ***
## t2$pctymle   3.05723085  1.07232182  2.8510 0.0057219 **
## t2$taxpc     0.00760464  0.00827755  0.9187 0.3614036
## t2$mix      -0.52554468  0.55302938 -0.9503 0.3452287
## t2$wage      0.00011898  0.00025091  0.4742 0.6368417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients of density, arrest probability, percent minority are still statistically significant, despite adding in multiple correlated variables. We did gain additional high significance predictors, such as percent young male, but conviction probability lost statistical significance.

We notice that percent young male has gained more practical significance, a percent increase resulting in 3.05% increase in crime rates. We also see that other coefficients lost practical significance, including percent young male. Arrest probability's practical significance has been reduced as well, resulting in 1.75% decrease in crime for a percent increase.

Below is a summary of the models

```
stargazer(m1, m2, m3, m4, 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', 'log(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	log(Crime Rate)		
	(1)	(2)	(3)	(4)
Population Density	0.007***	0.166***	0.143***	0.109***
	(0.001)	(0.025)	(0.032)	(0.030)
Arrest Probability	−0.046***	−1.517***	−2.069***	−1.750***
	(0.014)	(0.421)	(0.411)	(0.374)
Conviction Probability			−0.665*	−0.565*
			(0.325)	(0.219)
Prison Probability			−0.319	−0.035
			(0.536)	(0.428)
Average Prison Sentence			−0.001	−0.007
			(0.016)	(0.014)
Tax per Capita			1.159***	1.198***
			(0.188)	(0.194)
Percent Minority			1.846	3.057*
			(1.222)	(1.467)
Offense Mix				0.008**
				(0.003)
Percent Young Male				−0.526
				(0.495)
Sum of Wages				0.0001
				(0.0001)
Constant	0.038***	−3.269***	−3.088***	−3.915***
	(0.006)	(0.161)	(0.508)	(0.560)
R^2	0.584	0.506	0.695	0.737
Adjusted R^2	0.573	0.493	0.666	0.699
Residual Std. Error	0.012 (df = 78)	0.360 (df = 78)	0.292 (df = 73)	0.277 (df = 70)
Note:				*p<0.05; **p<0.01; ***p<0.001

We see that the level-level model had a higher R^2 than the log-level model. However, we maintain that the ease of interprebability makes model 2 superior. In any case, adding in the non-correlated covariates allowed us to explain over 2/3 of the variation in crime rate. Obviously, adding in all the additonal covariates will increase R^2 even further (though not by much), but we do not think model 4 is the best model due to the extra complexity.

Classical Linear Model Assumptions

In this section we will investigate whether any of the 6 CLM assumptions were violated. Mostly, we will be analyzing model 1, model 2 and model 3.

Model 1: $$crmrte = \beta_0 + \beta_1\ density + \beta_2\ prbconv + u$$

Model 2: $$\ln(crmrte) = \beta_0 + \beta_1\ density + \beta_2\ prbconv + u$$

Model 3: $$\ln(crmrte) = \beta_0 + \beta_1\ density + \beta_2\ prbconv + \beta_3\ prbarr + \beta_4\ prbpris + \beta_5\ avgsen + \beta_6\ pctmin80 + \beta_7\ pctyml$$

Linearity

All models are specified as an output variable in relation to a linear combination of input variables.

Random Sampling

While we've seen several problems with data integrity within this data set, we have no reason to believe that the counties were not sampled randomly from the same population (North Carolina).

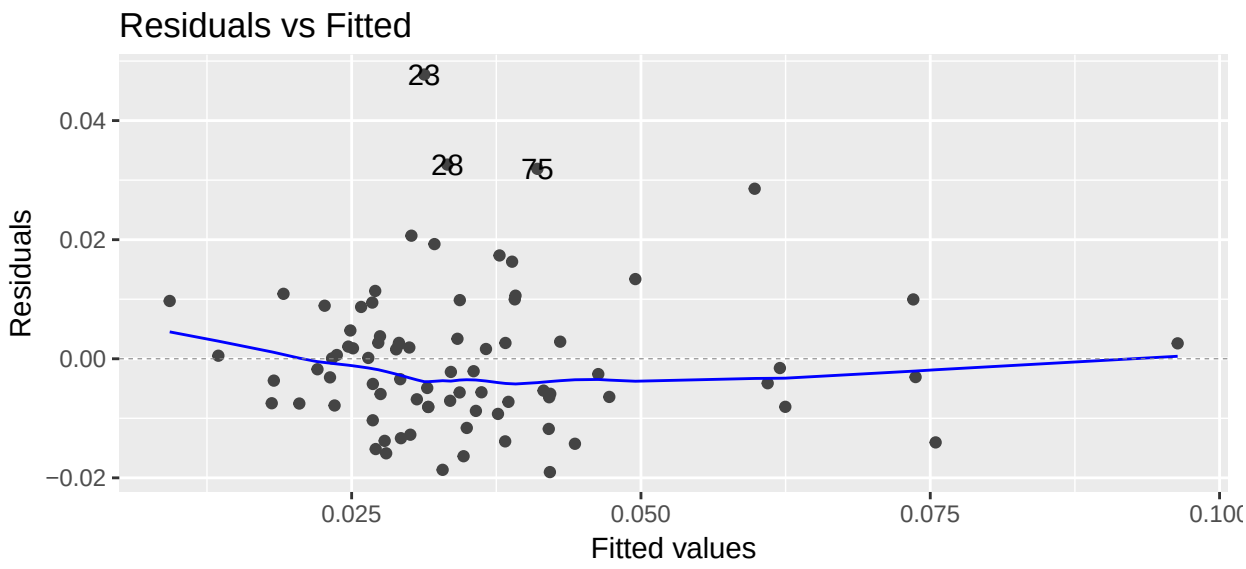
Perfect Multicolinearity

Density and and conviction probability are obviously not perfectly colinear. In fact, none of the variables in the data set with the exception of the manufactured wage-sum variable has perfect colinearity with each other.

Zero Conditional Mean

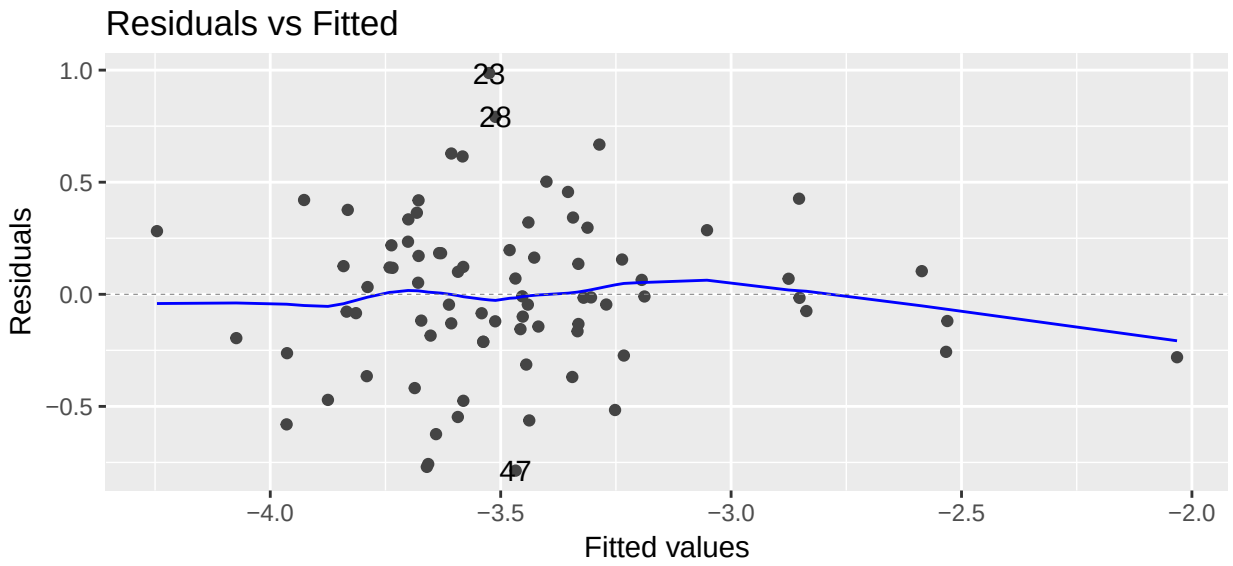
At this point, let us examine the diagnostic plots for our models.

```
autoplot(m1, which = 1)
```



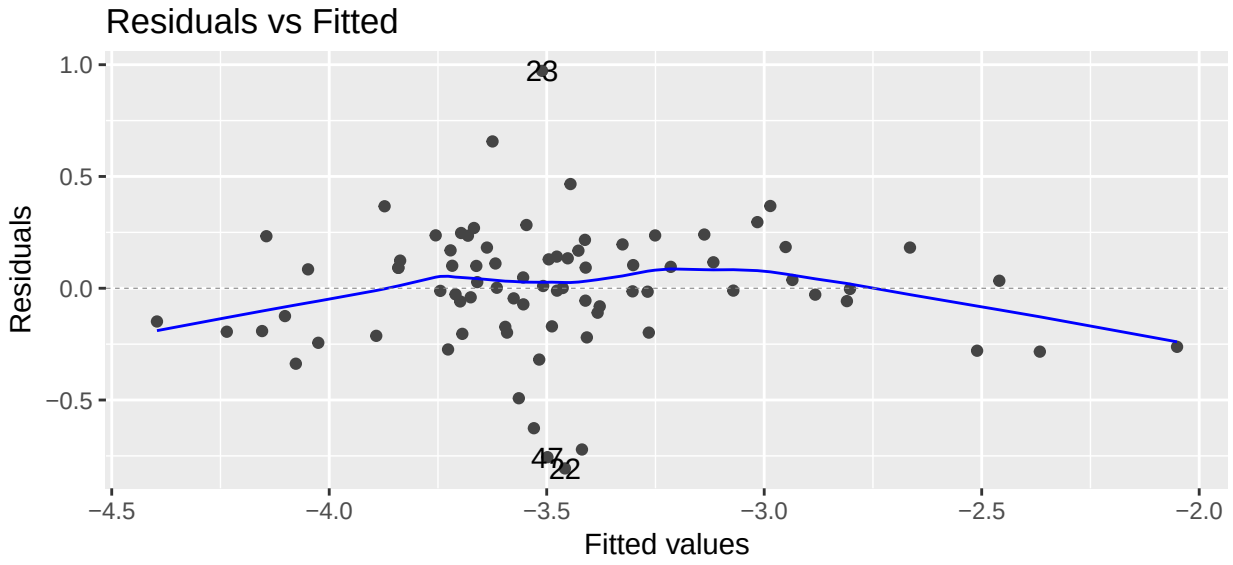
For model 1, we see from the residuals vs fitted values plot that there is some evidence of non-zero conditional error mean. This means our coefficients will be biased. We can try and troubleshoot this with model specification, which is model 2 in this case.

```
autoplot(m2, which = 1)
```



We now see a better residuals vs fitted values plot (much closer to zero conditional mean) for model 2.

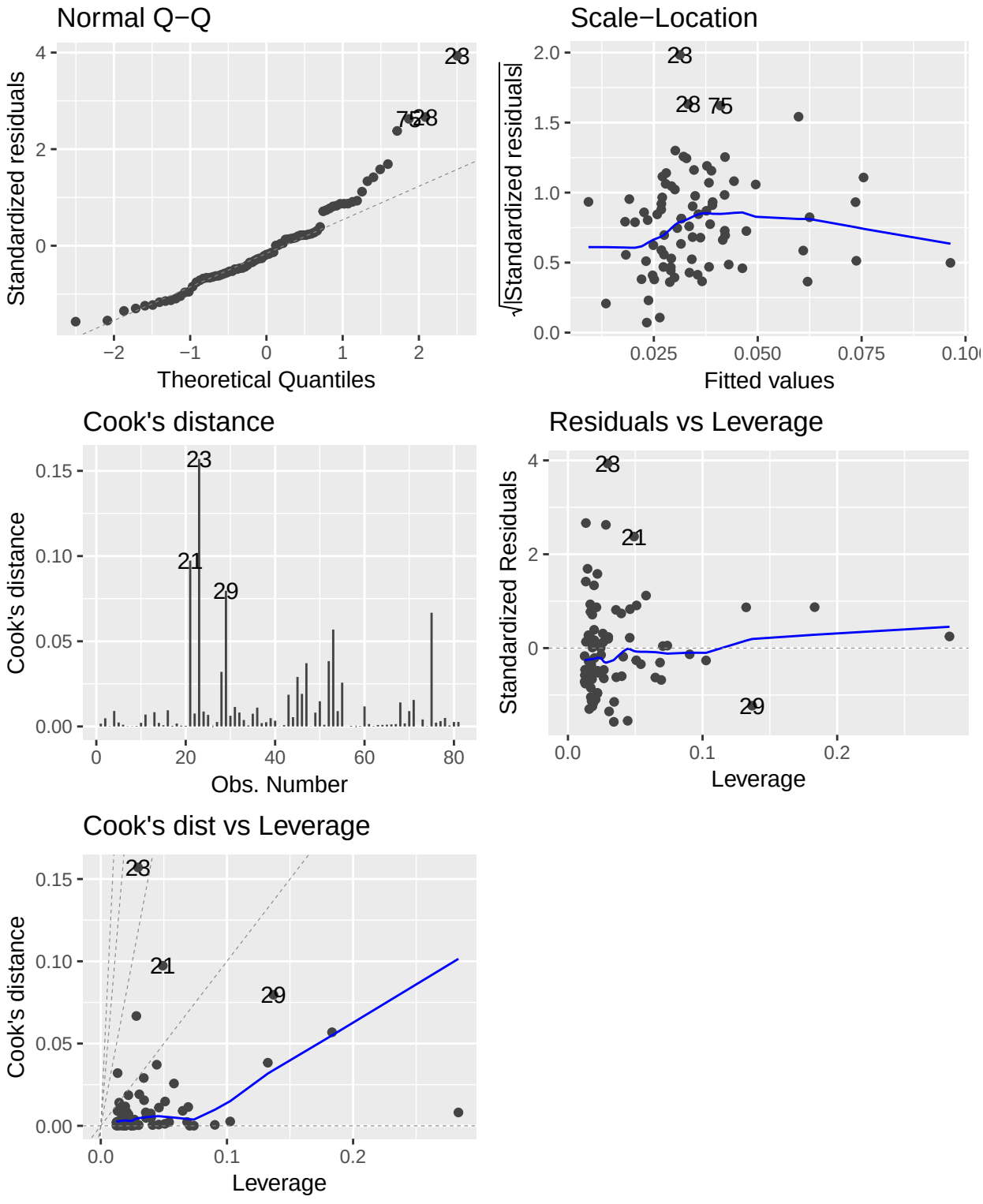
```
autoplot(m3, which = 1)
```



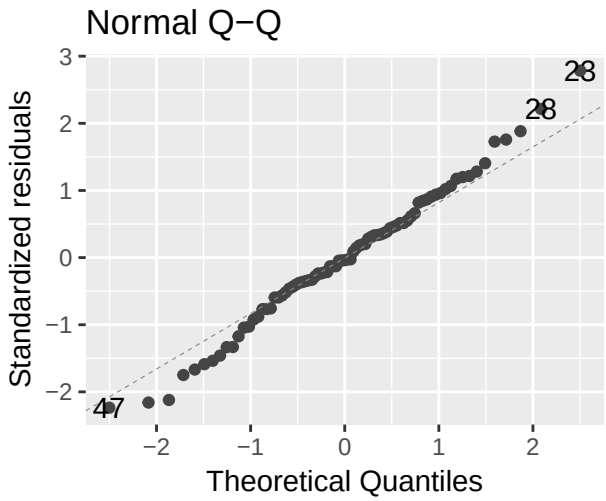
Even with log transformation of crime rates variable, reversed parabolic shape gives a strong evidence that zero-conditional mean assumption is violated in model 3. This can be explained by omitted variables that are hidden in u and correlated with our independent variables, which will be explained in details in the next section.

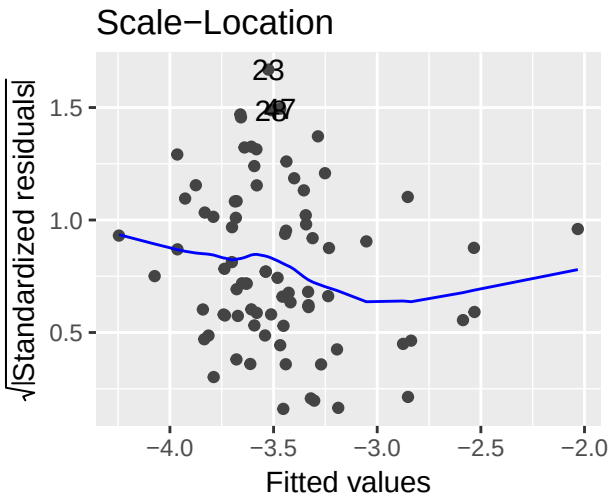
Homoskedasticity

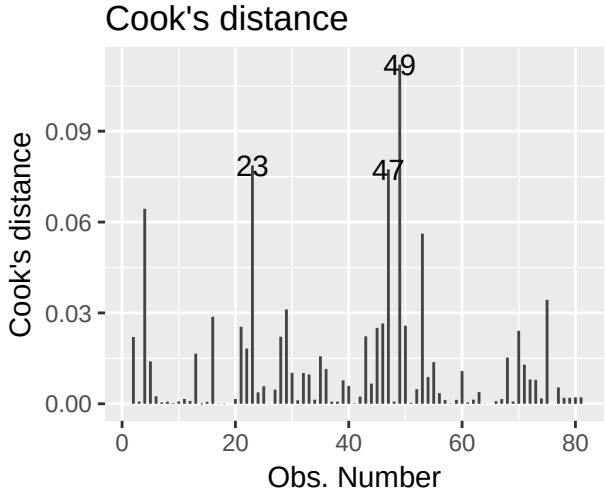
```
autoplot(m1, which = 2:6)
```

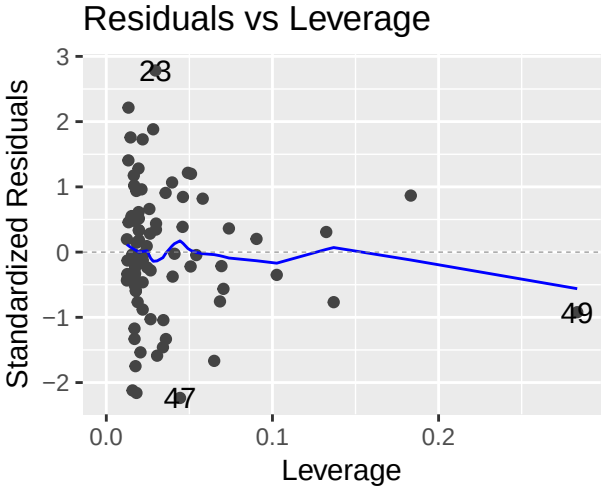


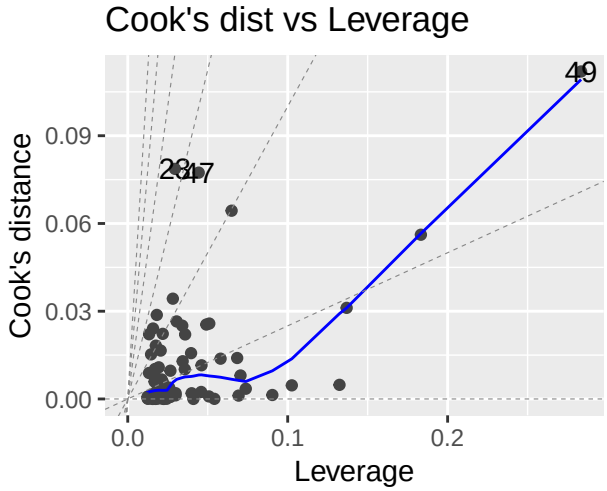
```
autoplot(m2, which = 2:6)
```



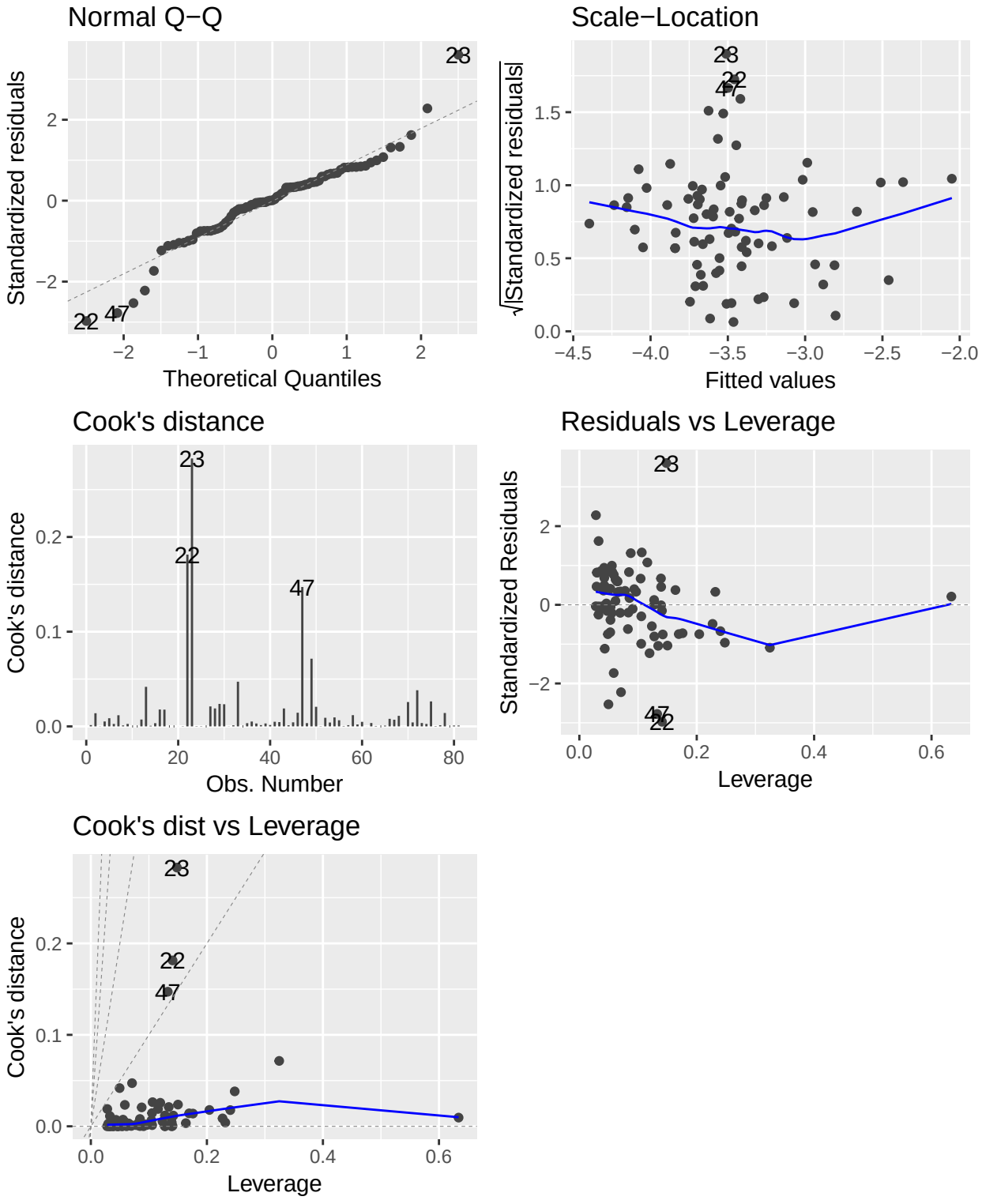









```
autoplot(m3, which = 2:6)
```



The scale-location plots of all models actually look fairly decent in terms of residual spread. However, Q-Q plot for model 1 and model 3 show skewness to the right and to the left, respectively. In model 2, we see a clear improvement in Q-Q plot from model 1 to model 2. Even though the scale-location plots for all 3 models look fairly homoskedastic, they are probably not convincing enough for us to use non-robust standard errors.

We see that row 49 has a fairly high leverage for model 2, but its Cook's distance is not high enough for us to

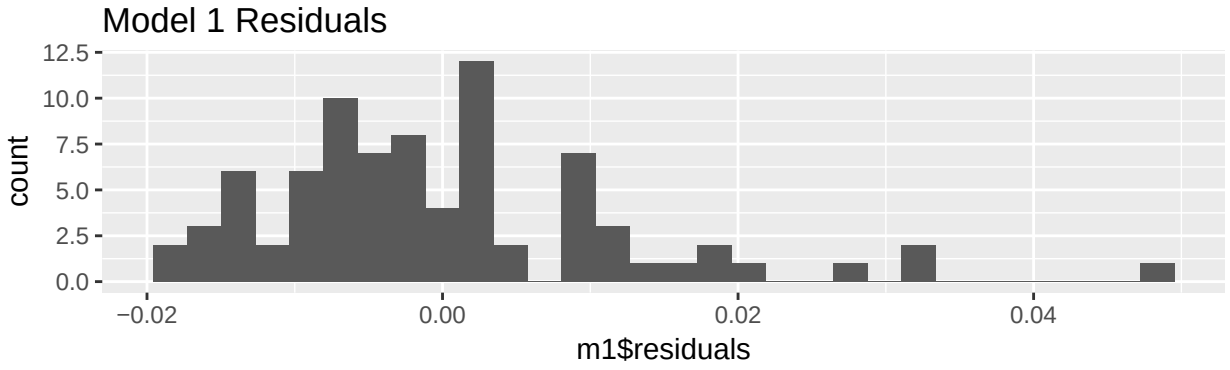
warrant investigating it further.

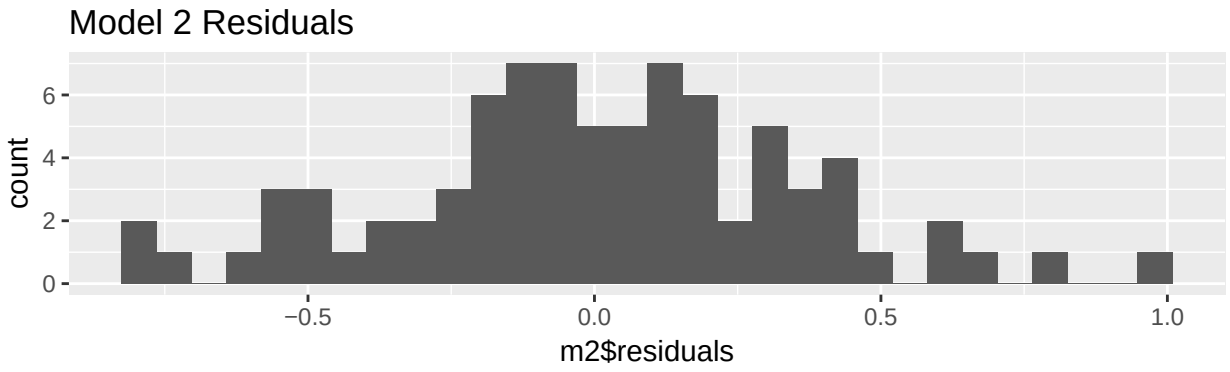
Normality of Errors

Here we see an advantage of taking the logarithm in our second model. Our residuals seem much closer to normality in the second model, as shown in the Q-Q plots. This advantage is not huge, as we could have relied on asymptotics for model 1 as well.

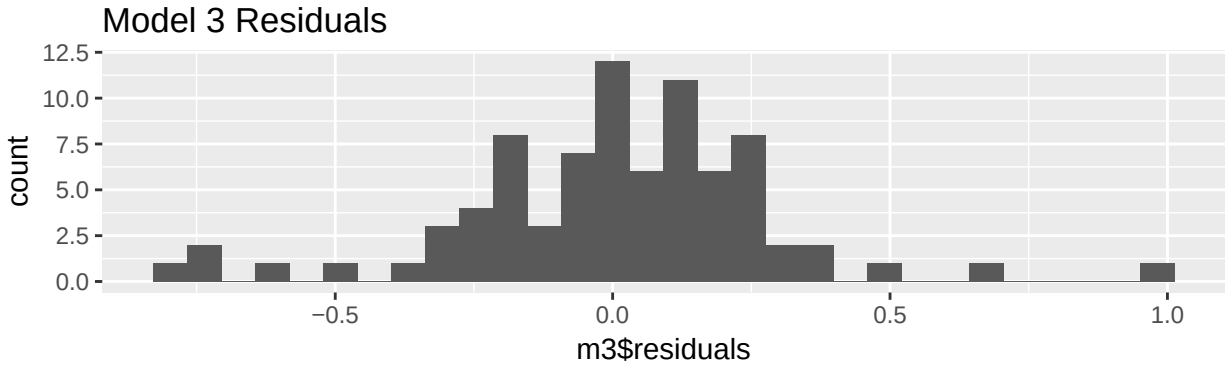
```
ggarrange(
    qplot(m1$residuals, bins = 30) + ggtitle('Model 1 Residuals'),
    qplot(m2$residuals, bins = 30) + ggtitle('Model 2 Residuals'),
    qplot(m3$residuals, bins = 30) + ggtitle('Model 3 Residuals'),
    nrow = 2)
```

```
## $`1`
```






```
##
## $`2`
```



```
##
## attr(,"class")
## [1] "list"      "ggarrange"
```

Model 1 does not show normal distribution of residuals. However, with model transformation, model 2 shows pretty decent normal distribution of residuals. Model 3 shows fairly normal distribution as well.

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. Still, it may behoove the political campaign to visit high density areas and address the crime problem to the citizens. People living in high density areas seem to see the most crime, so it is more likely for it to be an important issue for them.

The law enforcement variables are more tractable. Of the three probabilities for punishment, we found conviction probability to have the highest impact on crime rate. Simply increasing police per capita does not seem to do a good job of reducing crime rate in isolation. This is perhaps why a high number of arrests does not reduce crime as much as a high number of convictions. It may be more prudent to make existing ordinances more harsh towards petty crime. Certainly intuitive, harsher punishments for lesser crimes should work well in reducing the number of offenses.

For further research, it would be a great idea to investigate the counties with anomalous arrest and convicition probabilities. Counties like Madison seem to have very low crime rate, but addtional investigation is required to figure out whether these counties are special in some way. If nothing else, one must elucidate why probabilities greater than 1 were recorded.