Lab3 Draft

w203: Statistics for Data Science

Tosin Akinpelu
Stewart Warther
Nishant Hegde

Introduction

How can a political campaign address crime in North Carolina? We believe that crime levels in a North Carolina county are determined by key socio-economic variables.

In this report we examine crime levels in 1987 for 97 North Carolina counties to answer the following questions:
1. Can we understand what the key determinants of crime in North Carolina is?
2. Can we generate suggestions to local government policy based on these determinants?

Data and Processing

Reading in the crime dataset

```
crime = read.csv("crime_v2.csv")
```

We use the variable crmrte (crimes committed per person) to characterize the amount of crime in a county. We first identify variables in the dataset that have a linear relationship with crmrtem and then use Multivariate Ordinary Least Squares (OLS) Regression to identify what predictor variables would likely have a causal effect on the amount of crime rate in a county.

Our report will have 3 models: 1. Model 1 (m1) has only the key explanatory variables of interest. This model does not include covariates, so it might not be the most accurate. 2. Model2 (m2) has the key explanatory variables, as well as only covariates that do not introduce any substantial bias. This model will be the most accurate. 3. Model3 (m3) has all explanatory variables, as well as any covariate whether or not those covariates introduce bias.

The variables identified in these models will be used for policy suggestions that can be applied to local government. The objective of these suggestions will be to reduce crime levels in a county.

Processing

We found that 6 rows do not have any observed data. We remove these rows and create an "_clean" dataset with non-missing data only.

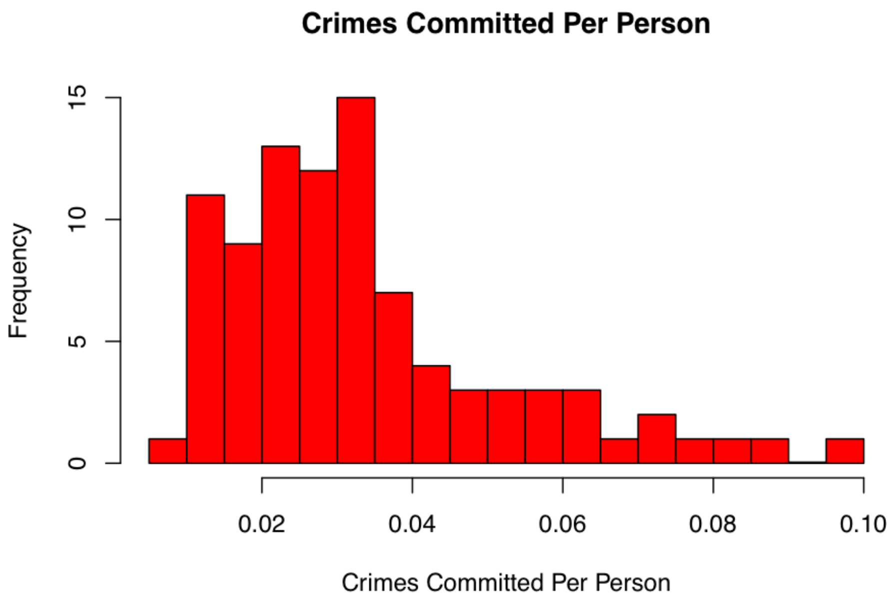
```
crime_clean <- crime[complete.cases(crime), ]
```

We summarize and plot a distribution of the dependent variable. There does not seem to be any nonsensical values in the range of cmrte, and the distribution shows a positive skew. Even though the distribution of the dependent variable is not normal, we decide not perform a log transformation since the skew is not high.

```
summary(crime_clean$crmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020927 0.029986 0.033400 0.039642 0.098966
```

```r
hist(crime_clean$crmrte, breaks = 20
     , col="red", xlab="Crimes Committed Per Person"
     , main="Crimes Committed Per Person")
```



The dataset has prbconv as a factor. We convert this to a numerical vector since we will be using OLS regression to identify key variables.

```r
crime_clean$prbconv_n <- as.numeric(levels(crime_clean$prbconv))[crime_clean$prbconv]
```

```
## Warning: NAs introduced by coercion
```

```r
summary(crime_clean$prbconv_n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

Correlation

To determine which predictors to use, we create a correlation matrix and look at the row for crmrte

```r
sort(cor(crime_clean[ ,c("crmrte", "prbarr", "prbconv_n", "prbpris"
                    ,"avgsen","polpc","density","taxpc", "west","central","urban",
                    "pctmin80","wcon","wtuc","wtrd","wfir"
                    ,"wser","wmfg","wfed","wsta","wloc"
                    ,"mix", "pctymle"
                    )])[1,])
```

```
##      prbarr   prbconv_n        west         mix        wser      avgsen
## -0.39332974 -0.38597236 -0.34938461 -0.13042871 -0.05256884  0.02741132
##     prbpris     central       polpc    pctmin80        wsta        wtuc
##  0.04698428  0.16960244  0.16988485  0.18679652  0.20199129  0.22935756
```

```
##     pctymle        wfir        wloc        wmfg        wcon        wtrd
## 0.29124849  0.32961199  0.34843532  0.35428801  0.39229444  0.41010559
##      taxpc        wfed       urban     density      crmrte
## 0.45097978  0.48615576  0.61560220  0.72896316  1.00000000
```

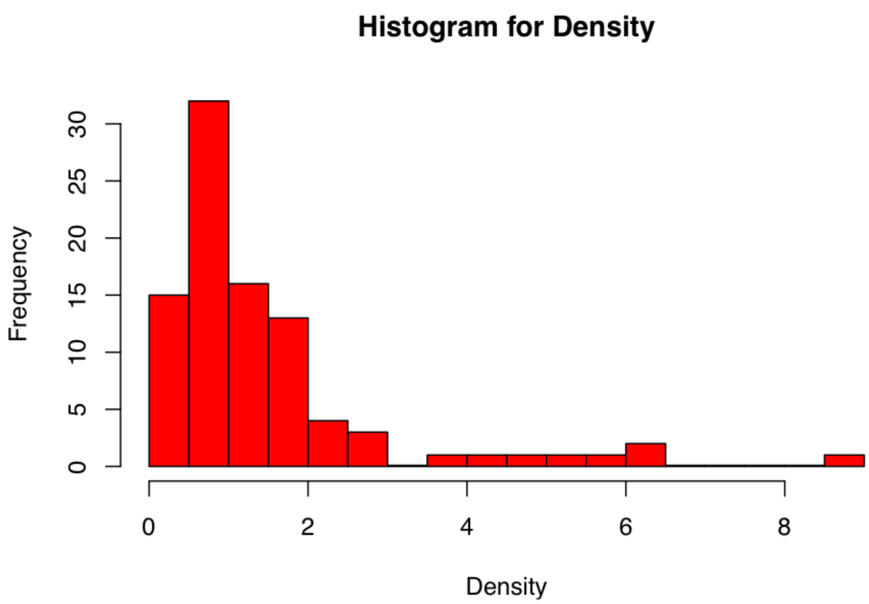
We see that density, urban, wfed, taxpc, wtrd, wcon, prbarr, and prbconv_n are correlated with our dependent variable crmrte. They may be good candidates to include in our model. The high correlation with urban and density makes intuitive sense. Urban areas are denser and will tend to have a high crime rate. Also, the negative correlation with prbarr and prbconv_n is indicative of crime being lower with a higher certainty of punishment.

Univariate Analysis

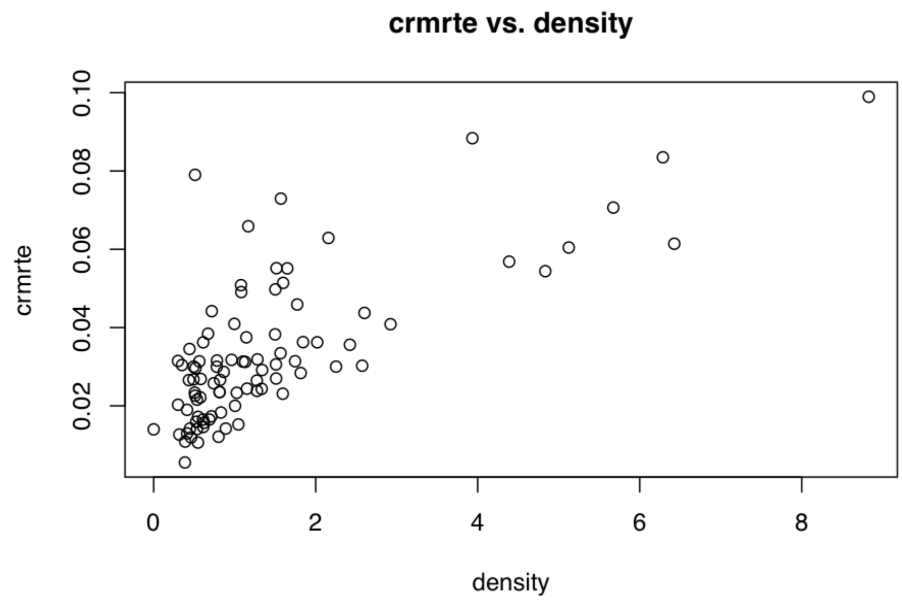
We look at distributions of some of these variables to determine if any transformations need to be made to these variables, so as to better elicit a linear relationship with crmrte

Distribution of density variable

```r
hist(crime_clean$density, breaks = 20 , col="red", xlab="Density" , main="Histogram for Density")
```




```r
plot(crime_clean$density, crime_clean$crmrte, xlab = "density", ylab = "crmrte",
     main = "crmrte vs. density")
```

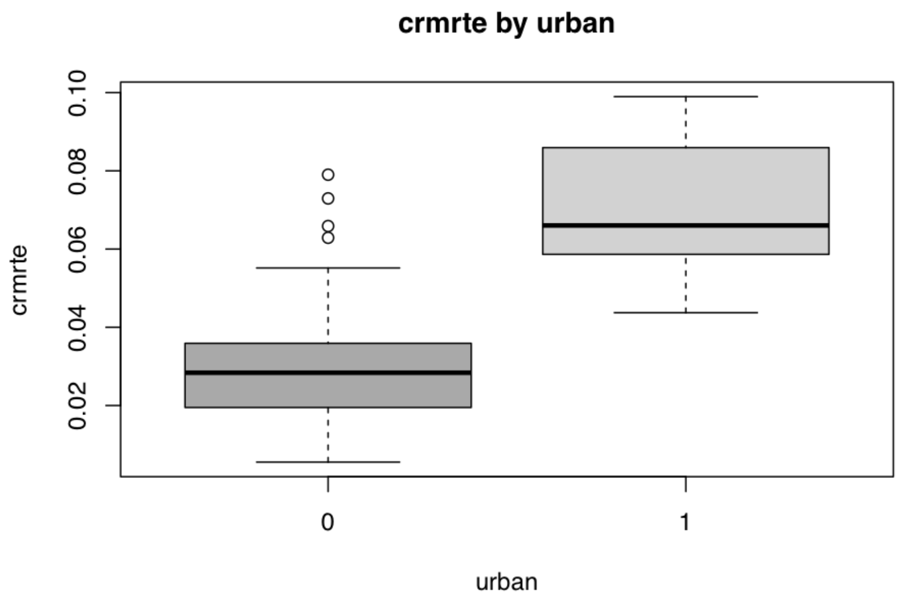



```r
cor(crime_clean$crmrte, crime_clean$density)
```

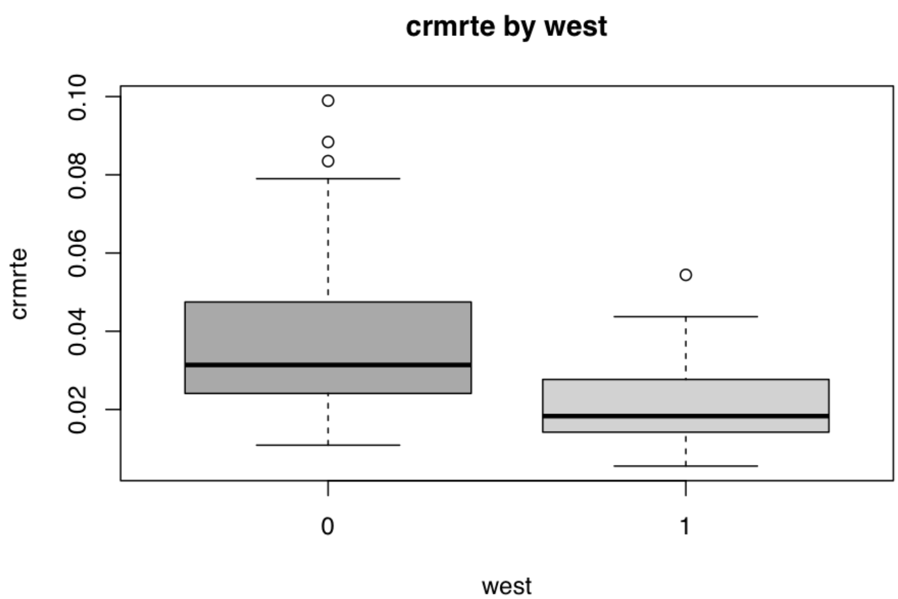
```
## [1] 0.7289632
```

We notice a higher crime rate in counties that are in the SMSA and in counties that are not in western N.C

```r
boxplot(crmrte~urban, data = crime_clean,
        col = c("dark grey", "light grey"),
        xlab = "urban",
        ylab = "crmrte",
        main = "crmrte by urban")
```

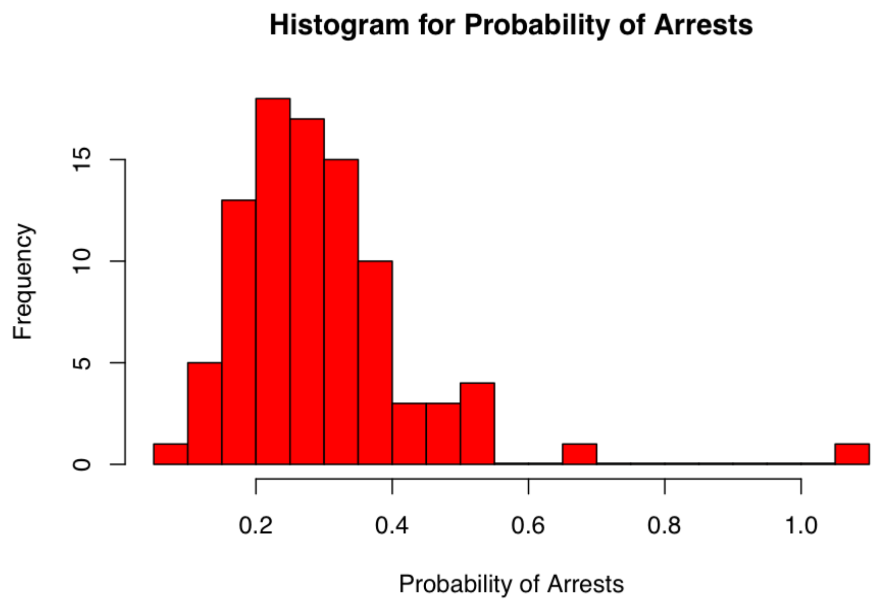



```
boxplot(crmrte~west, data = crime_clean,
        col = c("dark grey", "light grey"),
        xlab = "west",
        ylab = "crmrte",
        main = "crmrte by west")
```

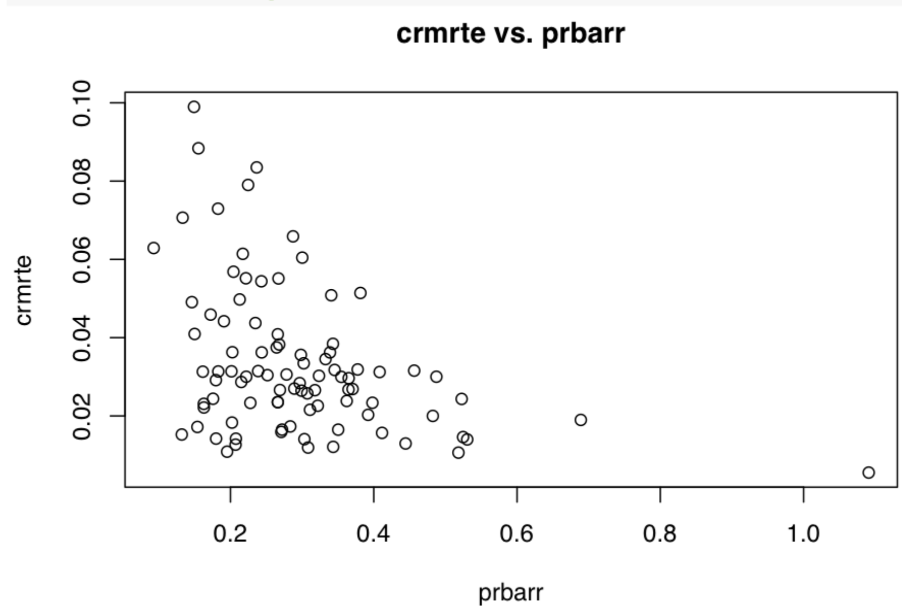


Distribution of prbarr variable

```
hist(crime_clean$prbarr, breaks = 20 , col="red", xlab="Probability of Arrests" , main="Histogram for P:
```




```
plot(crime_clean$prbarr, crime_clean$crmrte, xlab = "prbarr", ylab = "crmrte",
     main = "crmrte vs. prbarr")
```

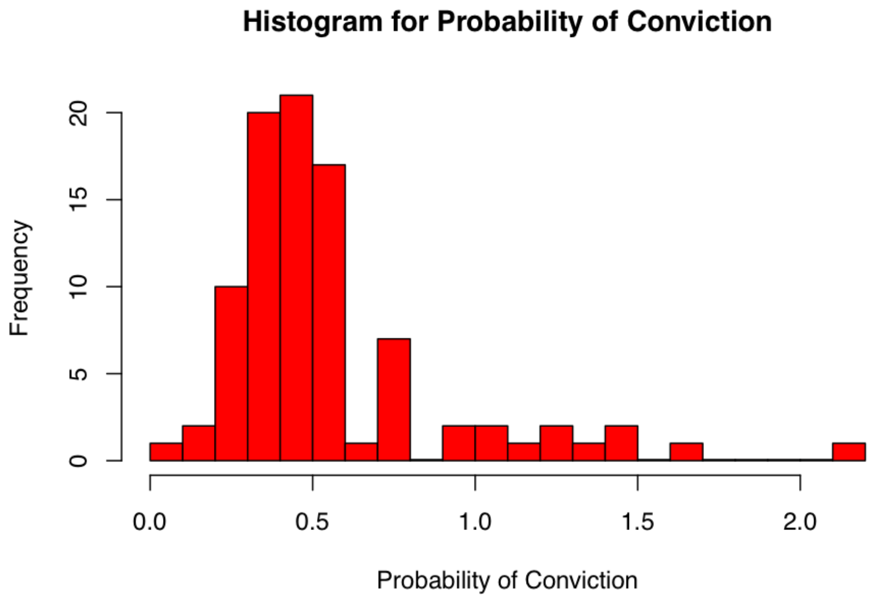


```
cor(crime_clean$crmrte, crime_clean$prbarr)
```

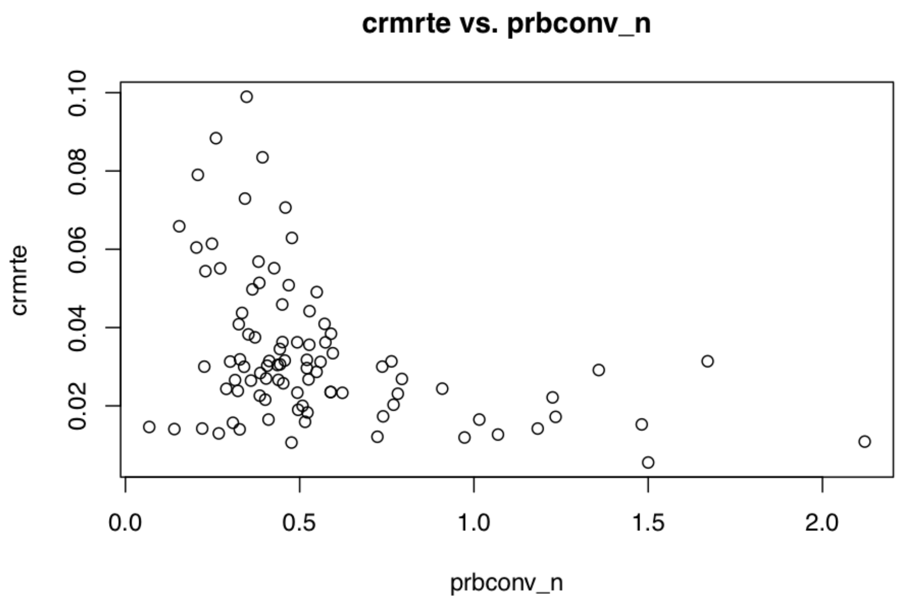
```
## [1] -0.3933297
```

Distribution of prbconv_n variable

```
hist(crime_clean$prbconv_n, breaks = 20 , col="red", xlab="Probability of Conviction" , main="Histogram
```




```
plot(crime_clean$prbconv_n, crime_clean$crmrte, xlab = "prbconv_n", ylab = "crmrte",
     main = "crmrte vs. prbconv_n")
```

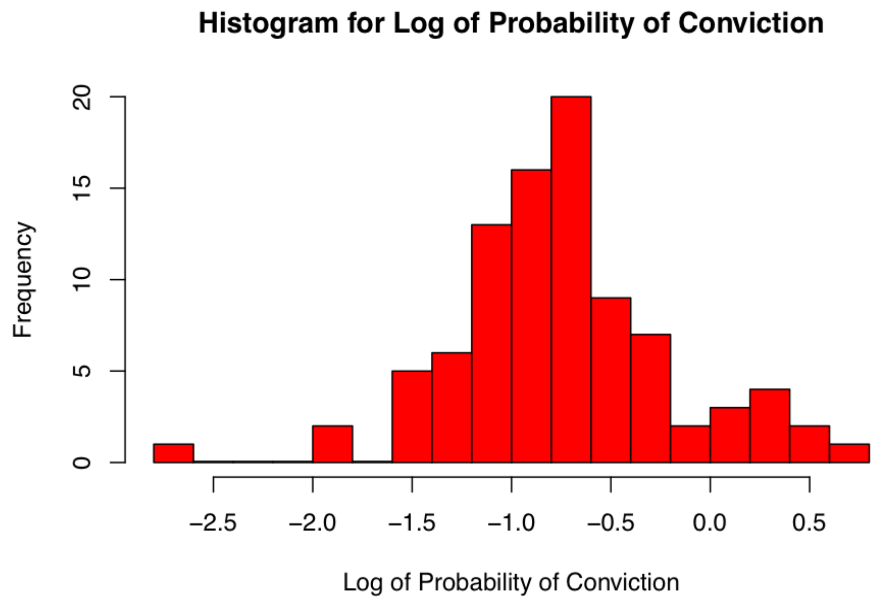



```
cor(crime_clean$crmrte, crime_clean$prbconv_n)
```

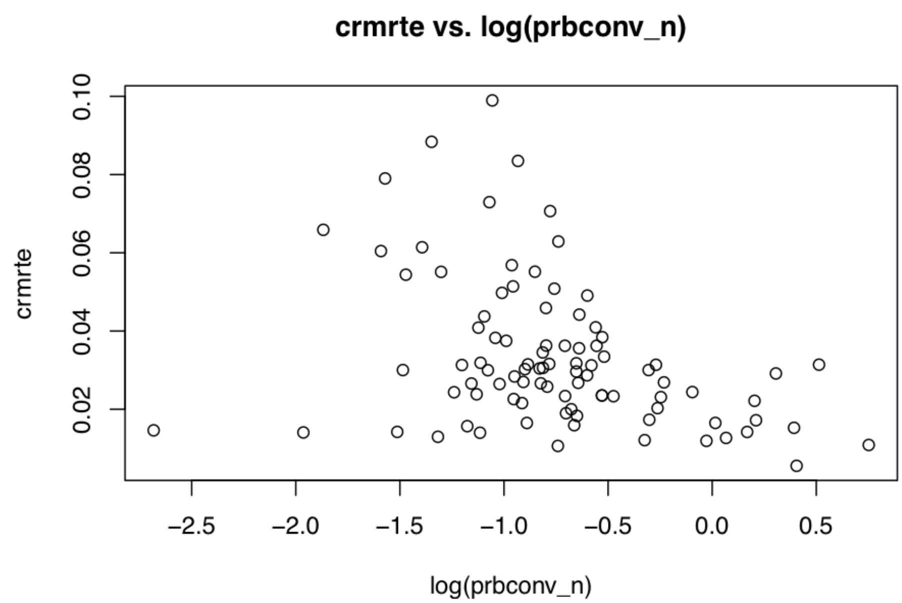
```
## [1] -0.3859724
```

We apply a log transformation to prbconv_n to get a better linear relationship with crmrte

```
hist(log(crime_clean$prbconv_n), breaks = 20 , col="red", xlab="Log of Probability of Conviction" , mai
```




```
plot(log(crime_clean$prbconv_n), crime_clean$crmrte, xlab = "log(prbconv_n)", ylab = "crmrte",
     main = "crmrte vs. log(prbconv_n)")
```

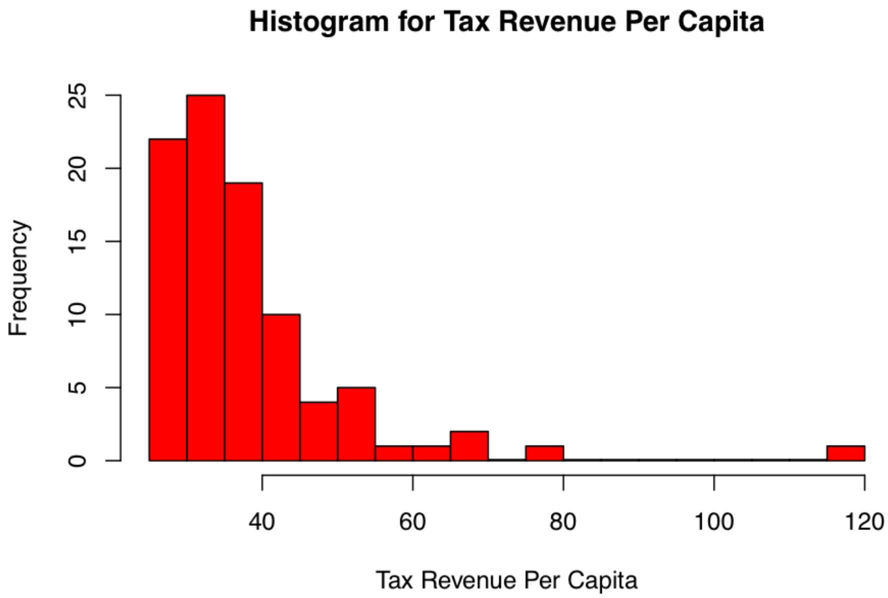


```
cor(crime_clean$crmrte, log(crime_clean$prbconv_n))
```

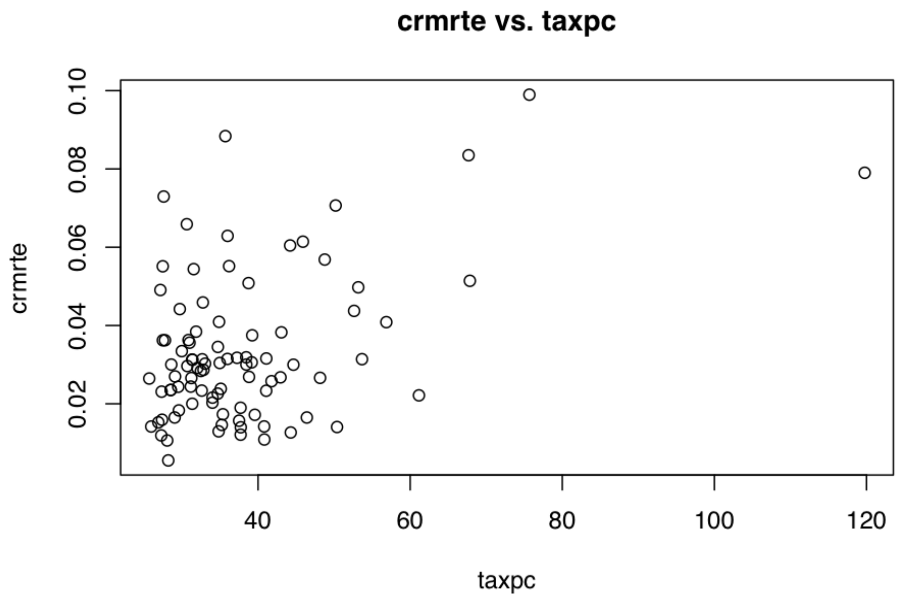
```
## [1] -0.364753
```

Distribution of taxpc variable

```
hist(crime_clean$taxpc, breaks = 20 , col="red", xlab="Tax Revenue Per Capita" , main="Histogram for Ta
```




```
plot(crime_clean$taxpc, crime_clean$crmrte, xlab = "taxpc", ylab = "crmrte",
     main = "crmrte vs. taxpc")
```

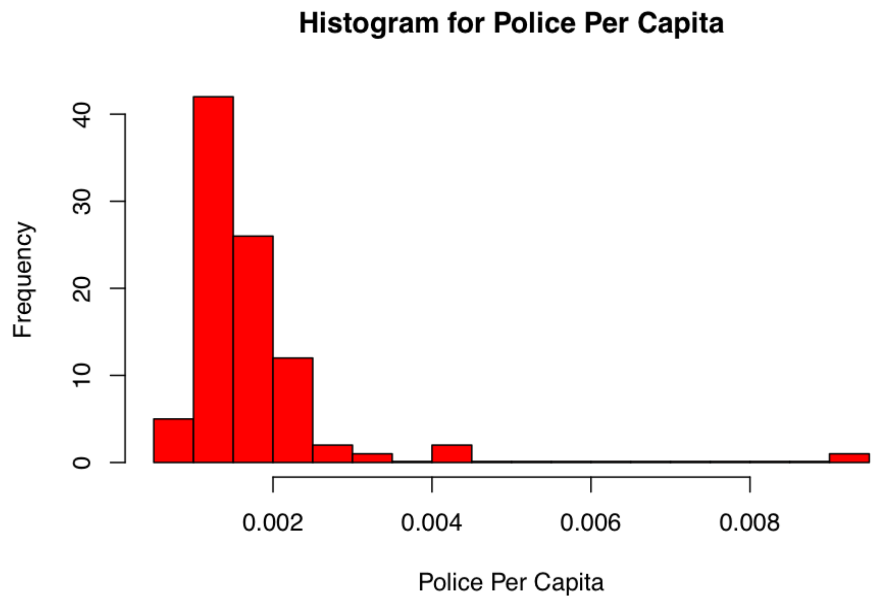



```
cor(crime_clean$crmrte, crime_clean$taxpc)
```

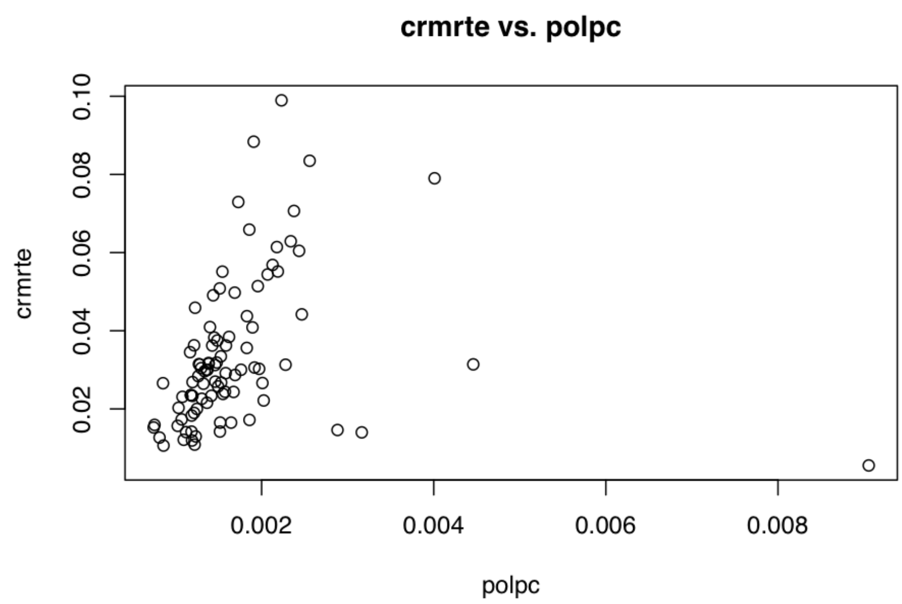
```
## [1] 0.4509798
```

Distribution of polpc variable

```
hist(crime_clean$polpc, breaks = 20 , col="red", xlab="Police Per Capita" , main="Histogram for Police I
```




```
plot(crime_clean$polpc, crime_clean$crmrte, xlab = "polpc", ylab = "crmrte",
     main = "crmrte vs. polpc")
```

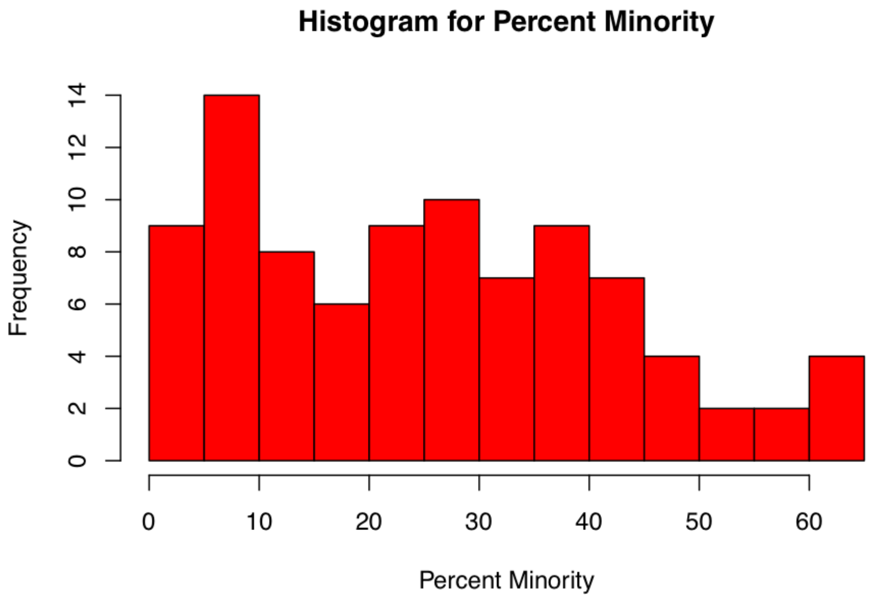


```
cor(crime_clean$crmrte, crime_clean$polpc)
```

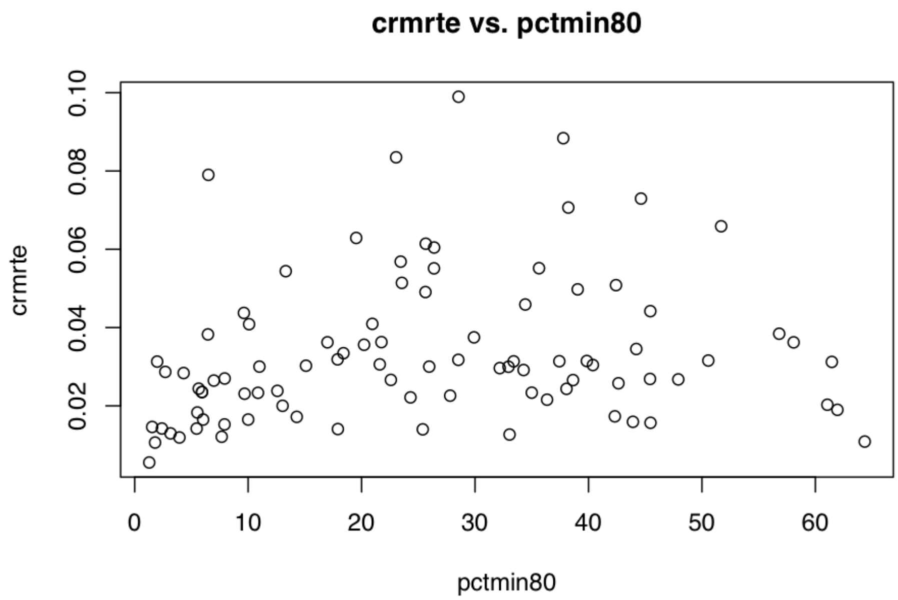
```
## [1] 0.1698849
```

Distribution of pctmin80 variable

```
hist(crime_clean$pctmin80, breaks = 20 , col="red", xlab="Percent Minority" , main="Histogram for Percer
```




```
plot(crime_clean$pctmin80, crime_clean$crmrte, xlab = "pctmin80", ylab = "crmrte",
     main = "crmrte vs. pctmin80")
```




```r
cor(crime_clean$crmrte, crime_clean$pctmin80)
```

```
## [1] 0.1867965
```

Building our Model

A quasi Forward Stepwise Regression approach was adopted which entails successively adding or removing variables based on the t-statistics of their estimated coefficients, Adjusted R-Squared and p-values suggesting the relative strength of each variable's relationship with crime rate observed from the correlation matrix.

Model 1

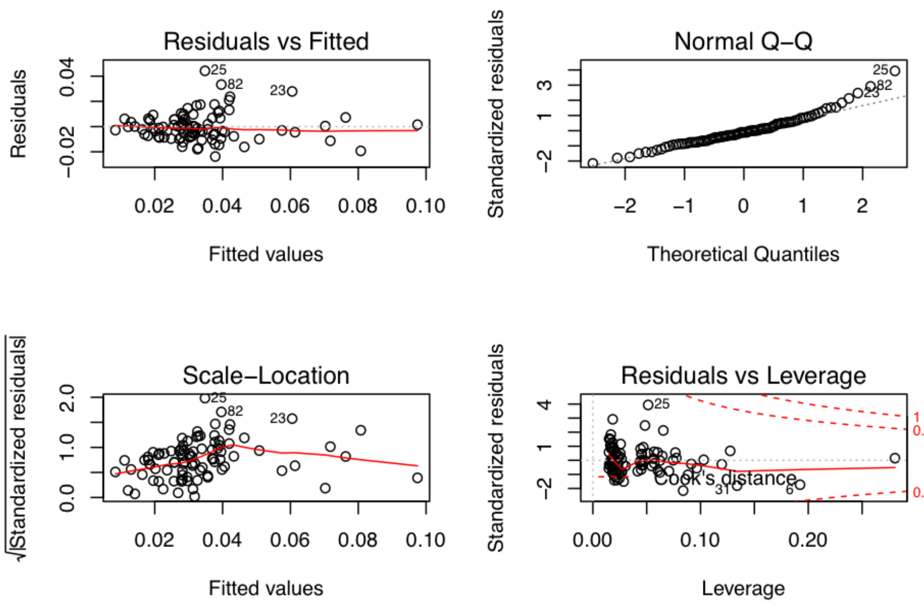
Model 1 explores relationship between crime rate, population density, a geographic dummy variable (west) and log-transformed 'probability' of conviction.

```r
m1 <- lm(crmrte ~ density + west + log(prbconv_n), data = crime_clean)
summary(m1)
```

```
##
## Call:
## lm(formula = crmrte ~ density + west + log(prbconv_n), data = crime_clean)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.023776 -0.007674 -0.001236  0.005655  0.044208
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      0.019342   0.002341   8.264 1.44e-12 ***
## density          0.007983   0.000834   9.572 3.03e-15 ***
## west            -0.011095   0.002811  -3.947  0.00016 ***
## log(prbconv_n) -0.007245   0.002248  -3.223  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01153 on 87 degrees of freedom
## Multiple R-squared:  0.6367, Adjusted R-squared:  0.6242
## F-statistic: 50.83 on 3 and 87 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(m1)
```



Adjusted R-squared of 0.6242 and p-value below 0.01 suggest statistically significant relationship between crime rate and the variables.

Model 2

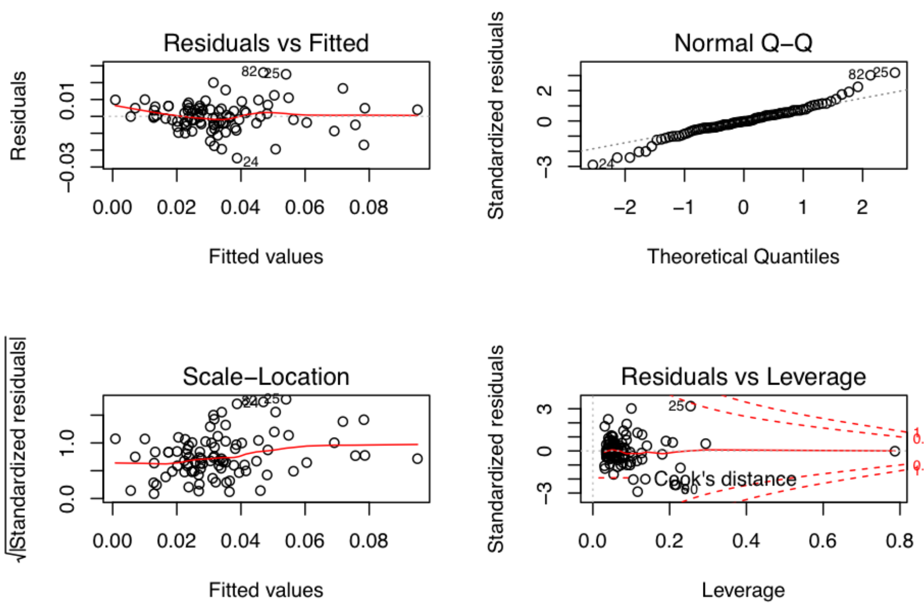
Model 2 adds crime deterrent variables (prbarr + polpc) and demographic variable (pctmin80) to Model 1 and evaluates the variables' relationship with crime rate.

```r
m2 <- lm(crmrte ~ density + west + log(prbconv_n) + central + prbarr + polpc + pctmin80, data = crime_c
summary(m2)
```

```
##
## Call:
## lm(formula = crmrte ~ density + west + log(prbconv_n) + central +
##     prbarr + polpc + pctmin80, data = crime_clean)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.024720 -0.004077 -0.000179  0.004672  0.025956
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.051e-02  4.319e-03   4.749 8.45e-06 ***
## density         6.456e-03  7.935e-04   8.136 3.49e-12 ***
## west           -7.617e-03  3.536e-03  -2.154  0.03415 *
## log(prbconv_n) -1.068e-02  1.866e-03  -5.721 1.63e-07 ***
## central        -5.470e-03  2.501e-03  -2.187  0.03153 *
## prbarr         -5.633e-02  9.133e-03  -6.167 2.42e-08 ***
## polpc           6.154e+00  1.200e+00   5.129 1.88e-06 ***
## pctmin80        2.248e-04  8.505e-05   2.643  0.00981 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009062 on 83 degrees of freedom
## Multiple R-squared:  0.786,  Adjusted R-squared:  0.7679
## F-statistic: 43.55 on 7 and 83 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m2)
```



An increase in Adjusted R-squared to 0.7679 and low individual p-values of the variables suggest statistically significant relationship with crime rate.

Model 3

Our third model includes previous covariates, and most, if not all, other covariates. The key purpose of this model is to demonstrate the robustness of our results to model specification. Adding the federal wage covariate increases the adjusted r-squared slightly to 72.8% versus model two although the p-value for this coefficient only meets the 10% threshold, not 5%.

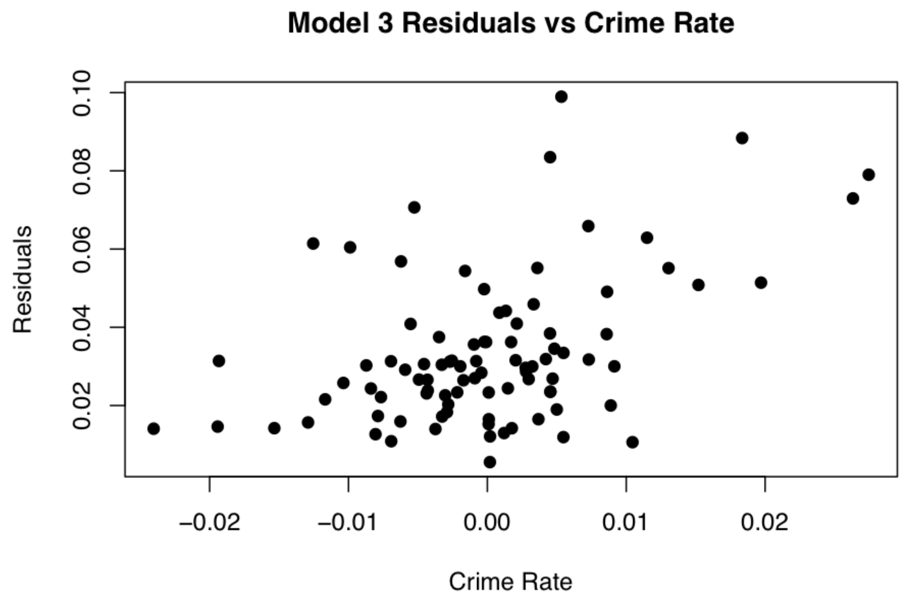
The 3rd model is specified as such: crmrte=density+west+log(prconv_n)+central+prbarr+polpc+pctmin80+wfed+u

Below are the steps to generate the model, inspect the coefficients, determine goodness of fit, and identify omitted variable bias.

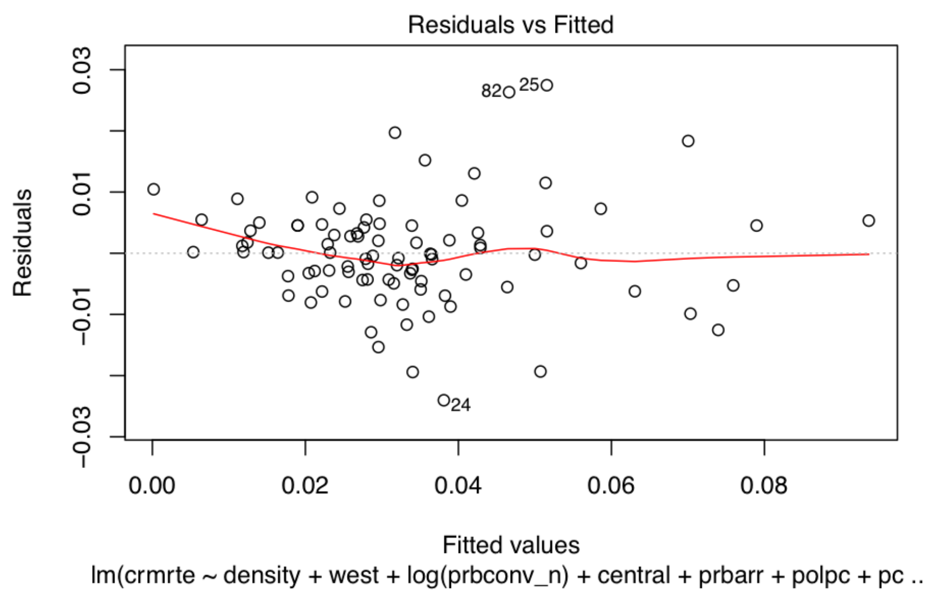
```
m3 <- lm(crmrte ~ density + west + log(prbconv_n) + central + prbarr + polpc + pctmin80 + wfed, data=cr

summary(m3)
```

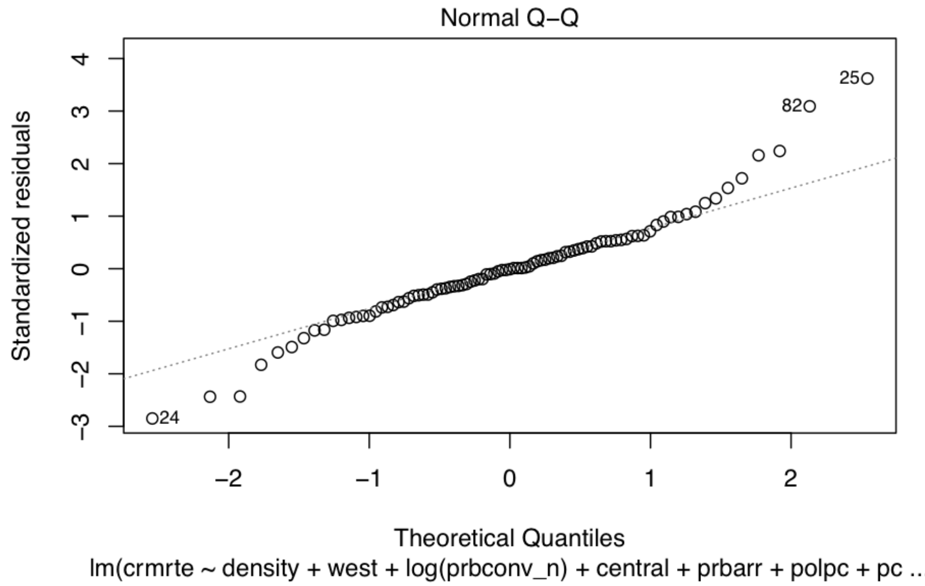
```
##
## Call:
## lm(formula = crmrte ~ density + west + log(prbconv_n) + central +
##     prbarr + polpc + pctmin80 + wfed, data = crime_clean)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0240297 -0.0044671 -0.0000902  0.0045169  0.0274627
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.650e-03  9.242e-03   0.828   0.4102
## density         5.799e-03  8.909e-04   6.509 5.64e-09 ***
## west           -7.787e-03  3.507e-03  -2.220   0.0292 *
## log(prbconv_n) -1.106e-02  1.866e-03  -5.928 6.97e-08 ***
## central        -6.149e-03  2.517e-03  -2.444   0.0167 *
## prbarr         -5.525e-02  9.079e-03  -6.086 3.55e-08 ***
## polpc           5.916e+00  1.199e+00   4.933 4.17e-06 ***
## pctmin80        2.127e-04  8.466e-05   2.513   0.0139 *
## wfed            3.206e-05  2.042e-05   1.570   0.1203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008983 on 82 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.772
## F-statistic: 39.08 on 8 and 82 DF,  p-value: < 2.2e-16
```

```
df <- data.frame(crime_clean$crmrte,m3$residuals)
plot(df$m3.residuals, df$crime_clean.crmrte, main="Model 3 Residuals vs Crime Rate",
    xlab="Crime Rate", ylab="Residuals", pch=19)
```

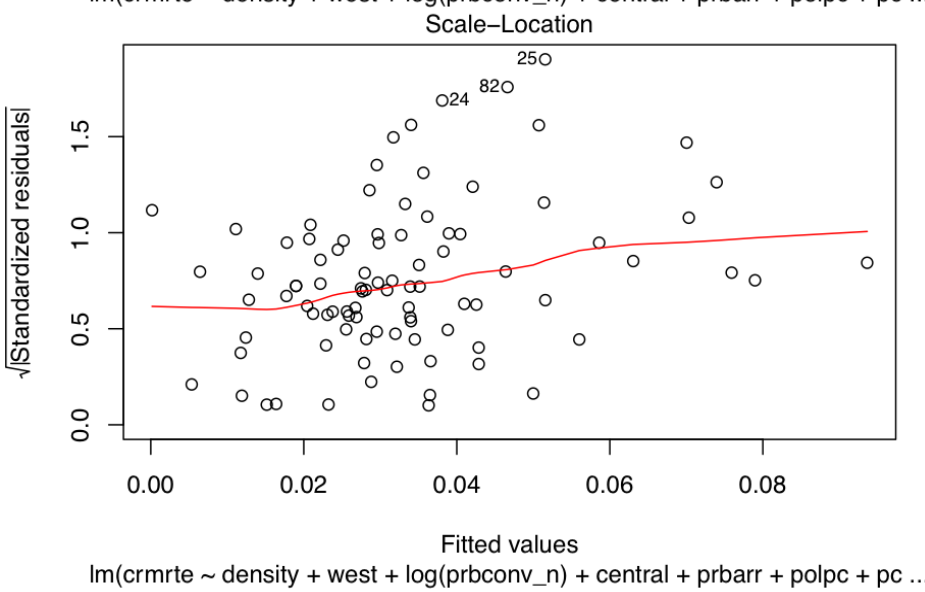

Model 3 Residuals vs Crime Rate
Residuals
0.02
0.04
0.06
0.08
0.10
−0.02
−0.01
0.00
0.01
0.02
Crime Rate

```r
plot(m3)
```

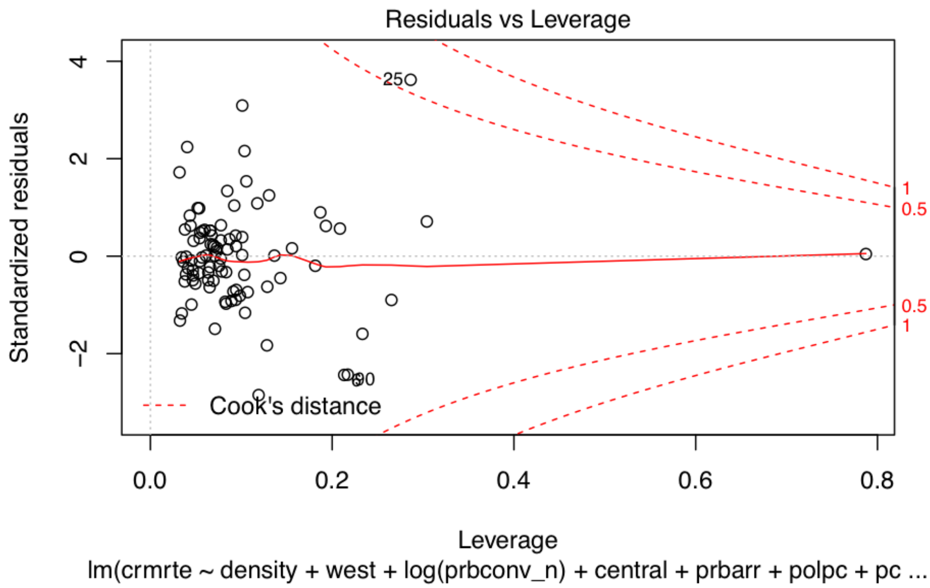

Residuals vs Fitted
Residuals
−0.03
−0.01
0.01
0.03
82
25
24
0.00
0.02
0.04
0.06
0.08
Fitted values
lm(crmrte ~ density + west + log(prbconv_n) + central + prbarr + polpc + pc ...



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(crmrte ~ density + west + log(prbconv_n) + central + prbarr + polpc + pc ...
25
82
24
−3
−2
−1
0
1
2
3
4



Scale−Location
√|Standardized residuals|
Fitted values
lm(crmrte ~ density + west + log(prbconv_n) + central + prbarr + polpc + pc ...
25
82
24
0.0
0.5
1.0
1.5
0.00
0.02
0.04
0.06
0.08



Model 3 shows a loose linear relationship between model residuals and the dependent variable, crime rate. This suggests the presence of omitted variable bias. Crime rates of high crime areas are underestimated by the model. Given that key variables and other covariates are all included in this model, it is likely that omitted variables were not included in the panel data. This may have included factors related to educational attainment. Another explanation may be that crime rates exhbibit non-linear tendancies.

Conclusion

We developed three linear regression models to explain crime rates by county in North Carolina during the year 1987. Using data provided in the panel, we sought to find the best combination of predictors which could help inform policy aimed at reducing crime.

From the data, we found that demographic and social factors were most important as was the perceived probability of conviction. Crimes were more likely to be commited in the western part of the State and were also more likely to be commited in dense urban areas. The perception of higher conviction probabilities was meaningful in reducing crime rates.

We expanded the model to other covariates and found that adding such factors could improve the adjusted R-squared of the model. However, even in the case of model 3, we see that residuals are positively related with the dependent variable, i.e. the model under-estimates high crime rate counties.

This linear relationship between dependent variable and residuals suggests the presence of omitted variable bias. We attempted the use of all covariates in the data provided and thus believe that any omitted variable is outside the scope of this exercise.

As a policy recommendation, we would suggest community outreach, more frequent PSAs, and higher conviction rates in densely populated areas and in the western part of the state. Notably, higher police concentration and higher government employee wages were positively related to crime rates. Thus, it would seem that increasing these may not be meaningful for reducing crime.

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.1. https://CRAN.R-project.org/package=stargazer
```

```r
stargazer(m1, m2, m3, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting Crime Rate in North Carolina Counties",
          keep.stat = c("adj.rsq", "n"),
          omit.table.layout = "n") # Omit more output related to errors
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Apr 02, 2018 - 15:32:25

Table 1: Linear Models Predicting Crime Rate in North Carolina Counties

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
density	0.008	0.006	0.006
west	−0.011	−0.008	−0.008
log(prbconv_n)	−0.007	−0.011	−0.011
central		−0.005	−0.006
prbarr		−0.056	−0.055
polpc		6.154	5.916
pctmin80		0.0002	0.0002
wfed			0.00003
Constant	0.019	0.021	0.008
Observations	91	91	91
Adjusted R^2	0.624	0.768	0.772