Lab 4

Shan He, Joanna Huang, Tiffany Jaya

18 December 2017

Introduction

The purpose of this report is to generate policy suggestions based on our understanding of the determinants of crime in North Carolina in 1987. We will list out the limitations of our analysis, including any estimates that suffer from endogeneity bias.

Exploratory Data Analysis

```r
# load the data
data <- read.csv("crime.csv")
# verify that it only contains data from 1987
unique(data$year)
```

```
## [1] 87
```

```r
# list number of counties
length(unique(data$county))
```

```
## [1] 90
```

```r
# list number of western, central, and urban counties
c(sum(data$west == 1), sum(data$central == 1), sum(data$urban == 1))
```

```
## [1] 21 34  8
```

```r
# list number of western & urban counties and central & urban counties
c(sum(data$west == 1 & data$urban == 1), sum(data$central == 1 & data$urban == 1))
```

```
## [1] 1 5
```

```r
# verify number of missing values
colSums(sapply(data, is.na))
```

```
##        X   county     year   crmrte   prbarr  prbconv  prbpris   avgsen
##        0        0        0        0        0        0        0        0
##    polpc  density    taxpc     west  central    urban pctmin80     wcon
##        0        0        0        0        0        0        0        0
##     wtuc     wtrd     wfir     wser     wmfg     wfed     wsta     wloc
##        0        0        0        0        0        0        0        0
##      mix  pctymle
##        0        0
```

The dataset contains 90 counties from North Carolina, all of which is collected in 1987. Out of the 90 counties, 21 are from western NC (out of which 1 is also urban), 34 are from central NC (out of which 5 is also urban), and 8 are considered urban counties. There are no missing values which will make our analysis easier.

For now, we will not take into consideration probabilities that are greater than 1 or less than 0 as well as percentages that are greater than 1 or less than 0. The assumption is that probabilities are in the range [0, 1] and percentages are in the range [0, 100]. Until we know the reason why the values are outside their range, we will not employ datapoints that do not conform to this assumption.

```
# list number of probabilities (prbarr, prbconv, prbpris, mix) that are not in range [0, 1]
c(sum(data$prbarr < 0 | 1 < data$prbarr), sum(data$prbconv < 0 | 1 < data$prbconv),
  sum(data$prbpris < 0 | 1 < data$prbpris), sum(data$mix < 0 | 1 < data$mix))
```

```
## [1]  1 10  0  0
```

```
# list number of percentages (pctymle, pctmin80) that are not in range [0, 100]
c(sum(data$pctymle < 0 | 100 < data$pctymle), sum(data$pctmin80 < 0 | 100 < data$pctmin80))
```

```
## [1] 0 0
```

prbarr and *prbconv* contain 1 and 10 datapoints respectively that do not conform to the probability assumption.

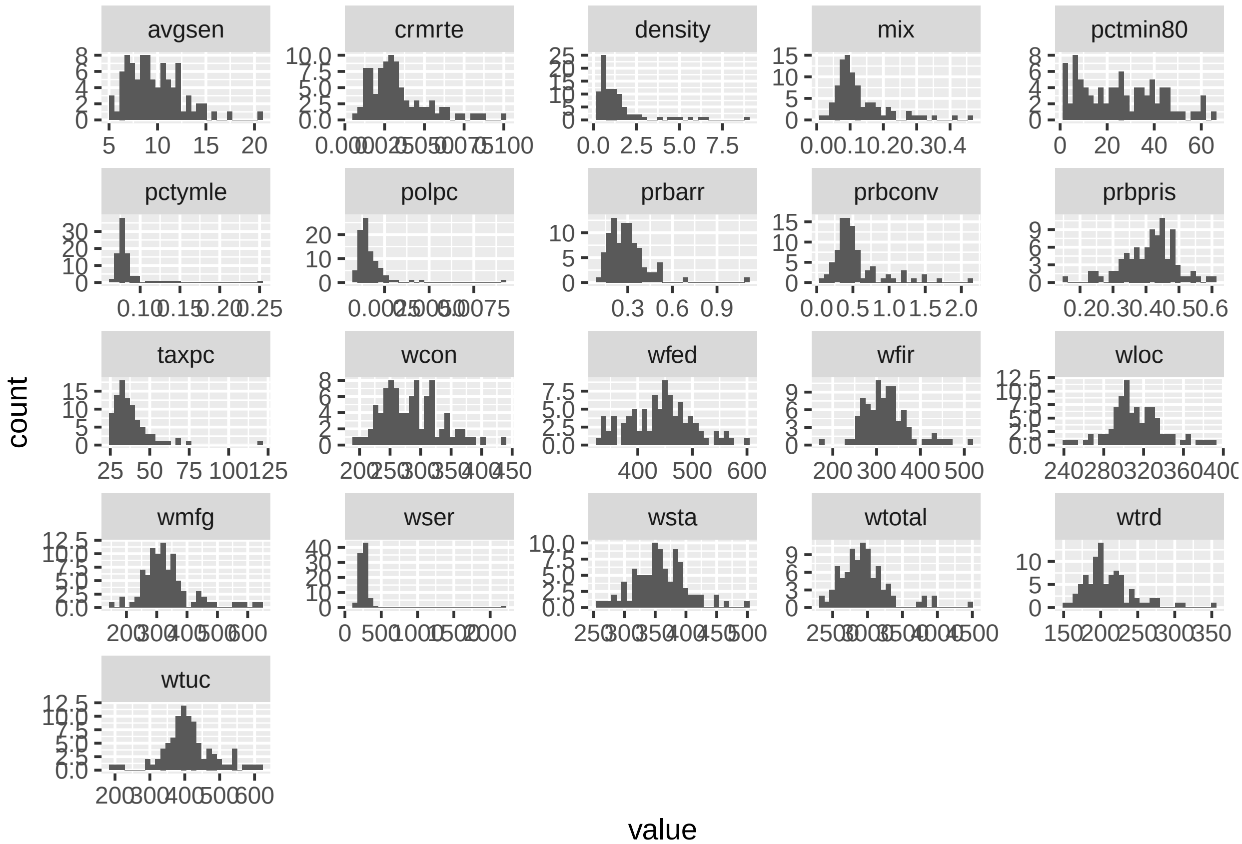
We have also decided to create an additional column that adds up all the wages to see if wages collectively can be considered as a predictor variable for the regression.

```
# create a column that adds up all the wages
data$wtotal <- rowSums(subset(data,
  select=c("wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg", "wfed", "wsta", "wloc")))
```

We then plot each numeric variable in a histogram to see its sample distribution.

```
# plot every variable except X, county, year, west, central, urban
library(ggplot2) # ggplot
library(tidyr) # gather
num.data <- data[!(names(data) %in% c("X", "county", "year", "west", "central", "urban"))]
ggplot(gather(num.data), aes(value)) +
       facet_wrap(~key, scales="free") +
       geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
library(moments) # skewness
skewness(num.data)
```

```
##      crmrte      prbarr     prbconv     prbpris      avgsen       polpc
##  1.28174888  2.52529596  2.03950599 -0.45254022  1.00116340  4.98348795
##     density       taxpc    pctmin80        wcon        wtuc        wtrd
##  2.65301071  3.29057447  0.36566169  0.60680223  0.06819768  1.46120657
##        wfir        wser        wmfg        wfed        wsta        wloc
##  0.82063146  8.69918165  1.42253166  0.13223761  0.36236826  0.29513808
##         mix     pctymle      wtotal
##  1.91657046  4.56069073  1.42770014
```

Most of the sample distributions appear to be positively skewed. We will take into consideration a logarithmic transformation when it is time to include the variables into the regression model.

From the histograms, we also see several notable outliers. We are under the impression that a county which has outlier in one variable will have outlier in another variable. For this reason, we have listed counties which have repeated outliers when we iterate through the entire numeric variables.

```r
# iterate through each numeric variable and list the outlier counties and their respective frequency
county.ids <- c()
for(var in num.data) {
  var.out <- boxplot.stats(var)$out
  county.ids <- c(county.ids, data[var %in% var.out, ]$county)
}
table(county.ids)
```

```
## county.ids
```

```
##   1   3   5   7  11  19  35  39  49  51  53  55  63  67  69  71  79  81
##   1   1   1   1   2   4   2   2   1   3   1   3   6   1   4   2   1   3
##  85  87  93  99 105 111 113 115 119 123 127 129 131 133 135 137 139 143
##   1   1   1   2   1   1   1   5  11   1   2   3   1   1   2   2   1   2
## 147 149 169 173 175 181 183 185 187 189 195 197
##   1   1   1   4   1   2   5   3   1   1   2   1
```

One outlier that is interesting to note is that the weekly wage in the service industry for county with id 185 is $2177.10, which is approximately eight times higher than the median. We do not know if the value is inputted incorrectly or if the county in general is making a weekly wage of $2177.10 in the service industry.

```r
summary(data$wser)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   133.0   229.3   253.1   275.3   277.6  2177.1
```

Variable Selection

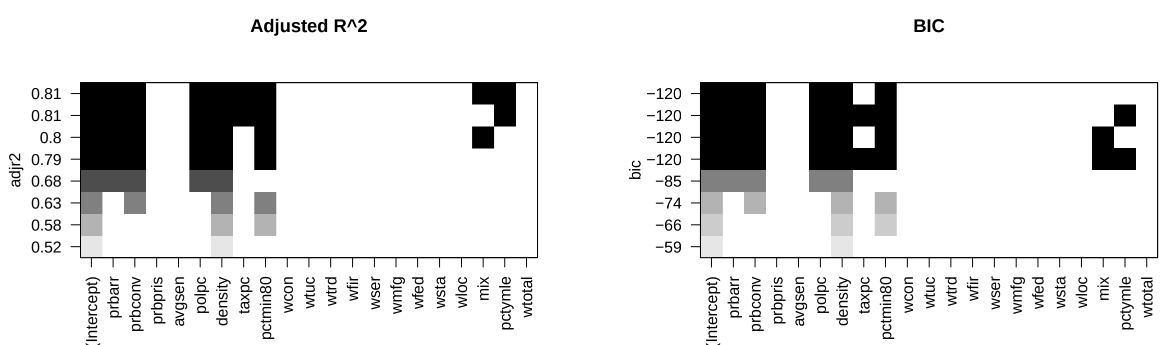
We have determined that our dependent variable will be *crmrte*, the crimes committed per person. We chose this variable because we want to understand what determinants of crime lead to the increase in crime rate.

Selecting the independent variables for a regression model and comparing the different regression models will give us a better understanding which determinants of crime lead to increase crime rate. We decide to perform all subsets regression to pick out our independent variables for the regression model. All subsets regression will consider all possible combinations of potential variables and finds the model that best fits the data. We chose all subsets regression over stepwise regression due to the criticism involved in stepwise regression.

```r
# perform all subsets regression on all numeric variables
library(leaps) # regsubsets
regsubsets.out <- regsubsets(crmrte ~ ., data=num.data)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found
```

```r
plot(regsubsets.out, scale="adjr2", main="Adjusted R^2")
plot(regsubsets.out, scale="bic", main="BIC")
```



Black indicates that a variable is included in the model, and the higher it is on the y axis, the better. Also, since the algorithm returns a best model of each size, it does not make a difference if we used AIC or BIC. For this reason, we will use the second best model shared by both the highest adjusted Rˆ2 and lowest BIC as our first regression model: $crmrte \sim prbarr + prbconv + polpc + density + taxpc + pctmin80 + pctymle$.

Model 1: only the explanatory variables of key interest

Model 2: key explanatory variables and covariates that increase accuracy without introducing bias

Model 3: most, if not all, other covariates

Summary of Models

Discussion of Causality

Conclusion