Lab 4

Shan He, Joanna Huang, Tiffany Jaya

17 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

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

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

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

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

```
# 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
```

```
# 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
```

```
# 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.

Perusing the variables, we note that there are For now, we take note on variables with probabilities and percentages that fall outside the range. Our assumption is that probabilities are in the range [0, 1] and percentages are in the range [0, 100].

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
```

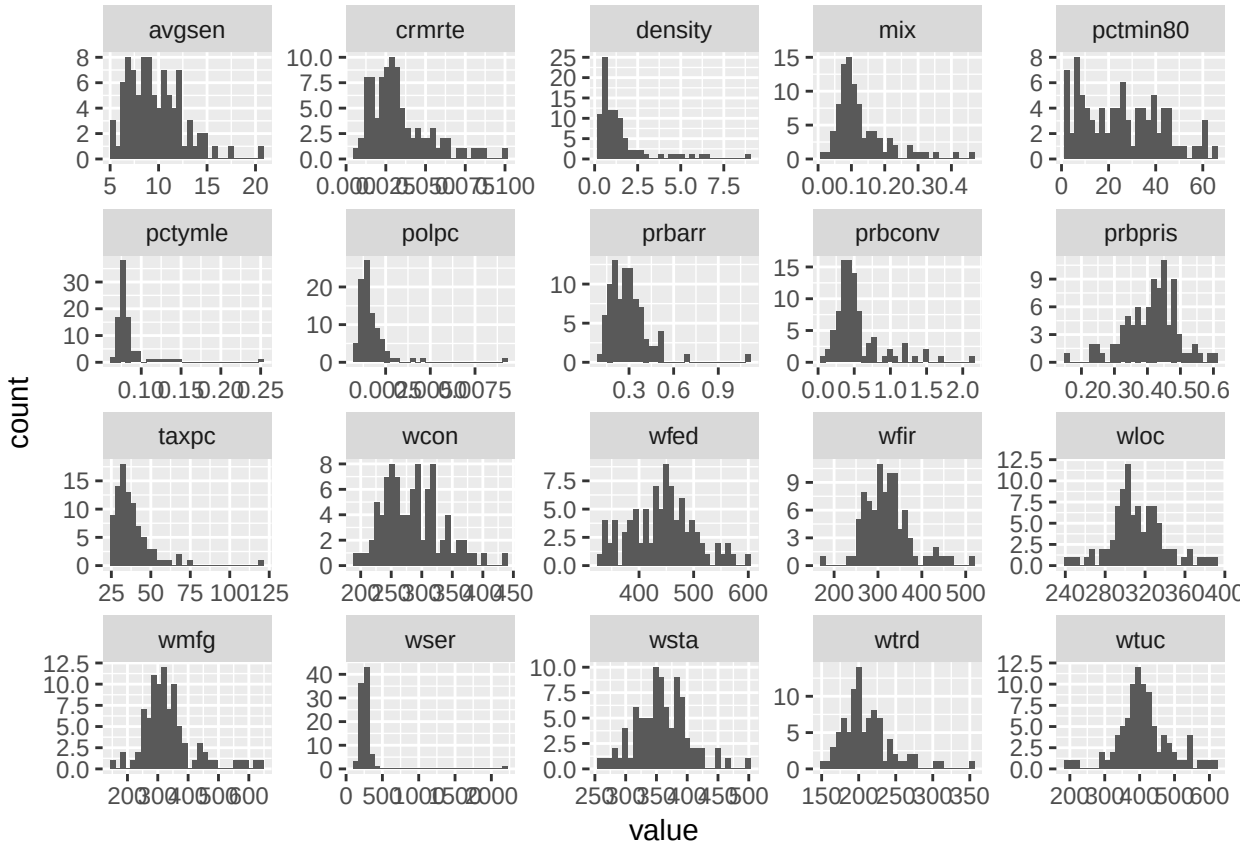
```
# list number of percentages (pctymle, pctmin80) that are not in range [0, 100]
c(sum(data$pctymle < 0 | 100 < data$pctymle), sum(data$pctmin80 < 0 | 100 < data$pctmin80))
```

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

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

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

```
# plot every variable except X, county, year, west, central, urban
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()
```




```
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
##  1.91657046  4.56069073
```

Most of the sample distributions appear to be positively skewed. We will take into consideration of logarithmic transformations, depending whether the interpretations make sense, 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 likely have outlier in another variable. For this reason, we have listed counties which have repeated outliers when we iterate through the entire numeric variables.

```
# 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   5   1   3   2   1   2
##  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  10   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   4   2   1   1   2   1
```

```
# list the most extreme outlier
outlier(num.data)
```

```
##        crmrte        prbarr       prbconv       prbpris        avgsen
##    0.09896590    1.09090996    2.12121010    0.15000001   20.70000076
##         polpc       density         taxpc      pctmin80          wcon
##    0.00905433    8.82765198  119.76145172   64.34819794  436.76663208
##          wtuc          wtrd          wfir          wser          wmfg
##  187.61726379  354.67611694  509.46551514 2177.06811523  646.84997559
##          wfed          wsta          wloc           mix       pctymle
##  597.95001221  499.58999634  388.08999634    0.46511629    0.24871162
```

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.

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

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

Research Question

James Q. Wilson and George Kelling's "broken windows theory" in 1982 led to a nation-wide movement for stricter crime-fighting policies between the 1980s and 1990s. The theory states:

> *if the first broken window in a building is not repaired, then people who like breaking windows will assume that no one cares about the building and more windows will be broken. Soon the building will have no windows. . . .*

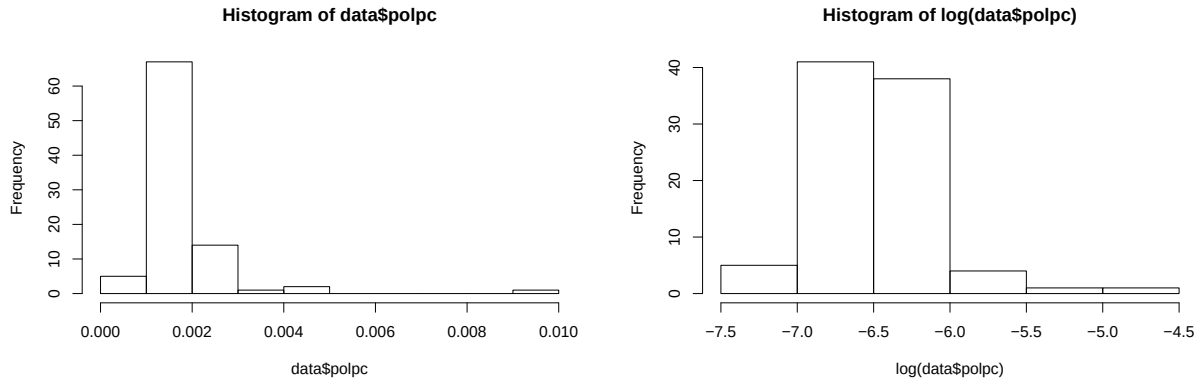
The belief was that by adopting a zero tolerance approach that enforced even the lowest level offenses, crime rates would subsequently go down. While New York City notably enforced this more stringent approach, San Francisco went the opposite direction of less strident law enforcement policies that reduced arrests, prosecutions and incarceration rates. Both sides experienced considerable declines in crime rates. Thus we hope to test the "broken windows theory" for the counties of South Carolina in 1987 and answer the question: Does the conservative approach of deterrence through arrests, incapacitation through imprisonment, harsh sentencing and higher police per capita lead to lower crime rates?

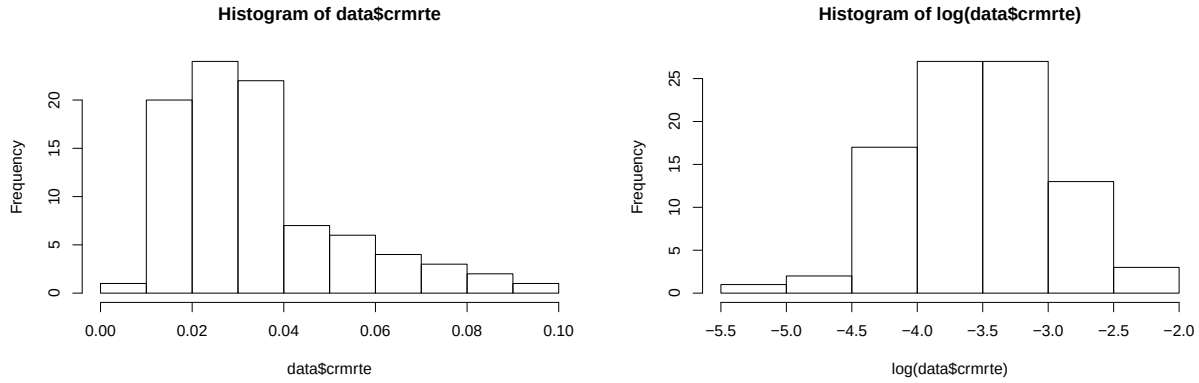
Model 1: only the explanatory variables of key interest

Based on the research question, our initial proposed model will include all variables related to stricter law enforcement policies: *prbarr*, *prbconv*, *prbpris*, *avgsen*, and *polpc*. Assuming the "broken windows theory" is valid, we expect generally negative coefficients for all variables.

Given that the histogram of *polpc* has a significant positive skew, we noted that it would do well to have a log transformation applied to *polpc* since its values are non-zero and positive. The same can be said about the dependent variable *crmrte* where its histogram is positively skewed and its values are non-zero and positive.

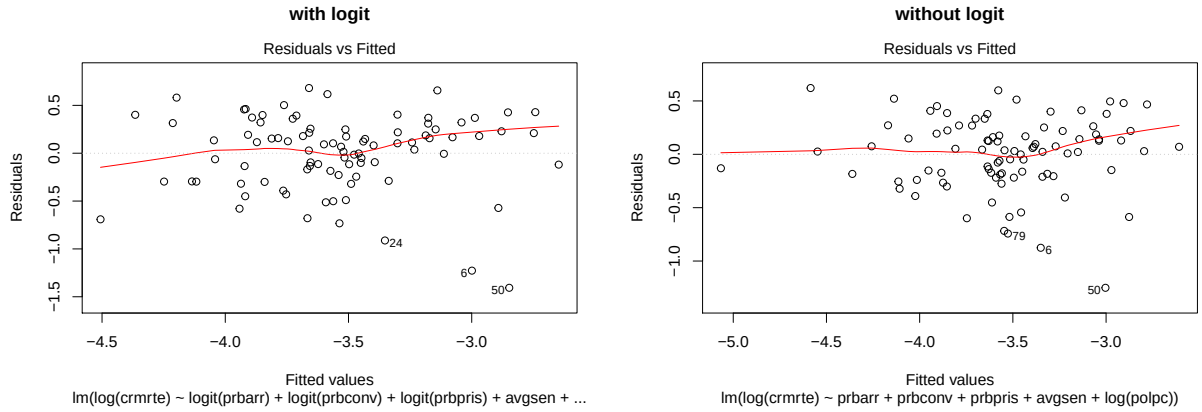
```
# before and after log transformation
hist(data$polpc); hist(log(data$polpc))
hist(data$crmrte); hist(log(data$crmrte))
```





Although the marginal distributions of *prbarr*, *prbconv*, and *prbpris* could benefit from a logit transformation, applying the transformation will make the residuals less normally distributed as seen from the residuals vs fitted values plot. We steered away from taking the log of these variables because it can make the values more extreme. For this reason, we kept *prbarr*, *prbconv*, and *prbpris* as is for the linear regression model.

```
# comparing with and without logit transformation on probability variables
plot(lm(log(crmrte) ~ logit(prbarr) + logit(prbconv) + logit(prbpris) + avgsen + log(polpc), data=data)
     which=1, main="with logit\n")
plot(lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + log(polpc), data=data),
     which=1, main="without logit\n")
```



We also kept *avgsen* as is for easier interpretation. Plus, applying the log on *avgsen* does not seem to affect the normality of the residual distribution.

For this reason, we propose our first model as follows which contains all explanatory variables of key interest:

$$log(crmrte) = \beta_0 + \beta_1 \cdot prbarr + \beta_2 \cdot prbconv + \beta_3 \cdot prbpris + \beta_4 \cdot avgsen + \beta_5 \cdot log(polpc) + u$$

We will now run the model and test the validity of the 6 CLM assumptions:

```
m1 <- lm(log(crmrte) ~ prbarr + prbconv + prbpris + avgsen + log(polpc), data=data)
```

CLM 1 - A linear model

The model is specified such that the dependent variable is a linear function of the explanatory variables. ## CLM 2 - Random Sampling

We do not know how the survey is collected. We assume that the variables are representative of the entire population distribution, but we cannot assess this assumption perfectly. Since there is a possibility that the individuals who collect the survey reach out to only one municipal police department instead of the county police department, the data collected this way are not representative of the county. There is nothing we can do to correct this, so we note this as a potential weakness in the analysis.

CLM 3 - Multicollinearity

As a quick test of the multicollinearity condition, we check the correlation of the explanatory variables and their Variance Inflation Factors (VIF):

```
# correlation matrix of explanatory variables
data$log.polpc <- log(data$polpc)
cor(data.matrix(subset(data, select=c("prbarr", "prbconv", "prbpris", "avgsen", "polpc", "log.polpc"))))
```

```
##                prbarr      prbconv     prbpris      avgsen      polpc
## prbarr     1.00000000 -0.055796206  0.04583324  0.17869425 0.42596481
## prbconv   -0.05579621  1.000000000  0.01102265  0.15585232 0.17186516
## prbpris    0.04583324  0.011022645  1.00000000 -0.09468083 0.04820783
## avgsen     0.17869425  0.155852319 -0.09468083  1.00000000 0.48815230
## polpc      0.42596481  0.171865155  0.04820783  0.48815230 1.00000000
## log.polpc  0.21624362 -0.007574581  0.01041348  0.43729326 0.90577332
##              log.polpc
## prbarr     0.216243622
## prbconv   -0.007574581
## prbpris    0.010413481
## avgsen     0.437293258
## polpc      0.905773320
## log.polpc  1.000000000
```

```
# verify VIFs are less than 10
vif(m1)
```

```
##     prbarr    prbconv    prbpris     avgsen log(polpc)
##   1.068228   1.039388   1.016889   1.310152   1.277425
```

The explanatory variables (logarrperc, logconvperc, logpolpc, logavgsen and prisperc) are not perfectly correlated and the VIFs are low (i.e. less than 10), so there is no perfect multicollinearity of the independent variables.

Is the assumption valid? **Yes Response**: No response required.

References:

"Shattering"Broken Windows": An Analysis of San Francisco's Alternative Crime Policies", CENTER ON JUVENILE AND CRIMINAL JUSTICE, October 1999 http://www.cjcj.org/uploads/cjcj/documents/shattering.pdf