# Lab3 Crime Statistics

*Kenneth Chen, Peter Trenkwalder, Danielle Salah*

*7/15/2018*

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

## Introduction

We received a crime dataset on North Carolina and would like to explore crime statistics. We like to investigate crime statistics at hands to develop several viable approaches in order to reduce crime. The dataset has **97** observations and **25** variables. Our first approach is to investigate each of the variables and how they relate to the occurrence of crimes in North Carolina in 1987.

## Exploratory Data Analysis

We listed all variables and their descriptions here.

| variable | label |
|---|---|
| 1 county | county identifier |
| 2 year | 1987 |
| 3 crmrte | crimes committed per person |
| 4 prbarr | 'probability' of arrest |
| 5 prbconv | 'probability' of conviction |
| 6 prbpris | 'probability' of prison sentence |
| 7 avgsen | avg. sentence, days |
| 8 polpc | police per capita |
| 9 density | people per sq. mile |
| 10 taxpc | tax revenue per capita |
| 11 west | =1 if in western N.C. |
| 12 central | =1 if in central N.C. |
| 13 urban | =1 if in SMSA |
| 14 pctmin80 | perc. minority, 1980 |
| 15 wcon | weekly wage, construction |
| 16 wtuc | weekly wage, trns, util, commun |
| 17 wtrd | weekly wage, whlelse, retail trade |
| 18 wfir | weekly wage, fin, ins, real est |
| 19 wser | weekly wage, service industry |
| 20 wmfg | weekly wage, manufacturing |
| 21 wfed | weekly wage, fed employees |
| 22 wsta | weekly wage, state employees |
| 23 wloc | weekly wage, local gov emps |
| 24 mix | offense mix: face-to-face/other |
| 25 pctymle | percent young male |

Out of 25 variables, we set our dependent variable to be **`crime rates, crmrte`** because we believe this reflects the frequency of crimes in North Carolina. To create our prediction model precisely and present

clearly, we developed several objectives in our approach and lay our foundational work here.

# Approach

## Sanity check and data cleaning

```
apply(!is.na(crime[1:25]), MARGIN = 2, mean)
```

```
##    county      year    crmrte    prbarr   prbconv   prbpris    avgsen
## 0.9381443 0.9381443 0.9381443 0.9381443 1.0000000 0.9381443 0.9381443
##     polpc   density     taxpc      west   central     urban  pctmin80
## 0.9381443 0.9381443 0.9381443 0.9381443 0.9381443 0.9381443 0.9381443
##      wcon      wtuc      wtrd      wfir      wser      wmfg      wfed
## 0.9381443 0.9381443 0.9381443 0.9381443 0.9381443 0.9381443 0.9381443
##      wsta      wloc       mix   pctymle
## 0.9381443 0.9381443 0.9381443 0.9381443
```

There are 97 observations and 25 variables in our dataset. We checked if there's any empty values in each variables by applying the `!is.na` function. Interestingly, only one variable `prpconv (probability of conviction)` has full observations, i.e., 97. The rest of the variables have 91 observations out of original 97, which give us `91/97 = 0.9381`.

We further checked if all 97 observations in `prpconv` is a real value or any of the special characters. As a control, we also check other variables as well.

```
sum(sapply(crime$prbconv, function(x) any(x %in% c("", "`", "?",
    "!", "@", "#", "$", "%", "^", "&", "*", "(", ")"))))
```

```
## [1] 6
```

```
# Checking every column of the dataset if they have any
# strange special characters in their value
for (i in 1:25) {
    special_chars = sum(sapply(crime[i], function(x) any(x %in%
        c("`", "", "?", "!", "@", "#", "$", "%", "^", "&", "*",
            "(", ")"))))
    print(paste("crime variable", i, "has", special_chars, " special characters."))
}
```

```
## [1] "crime variable 1 has 0  special characters."
## [1] "crime variable 2 has 0  special characters."
## [1] "crime variable 3 has 0  special characters."
## [1] "crime variable 4 has 0  special characters."
## [1] "crime variable 5 has 1  special characters."
## [1] "crime variable 6 has 0  special characters."
## [1] "crime variable 7 has 0  special characters."
## [1] "crime variable 8 has 0  special characters."
## [1] "crime variable 9 has 0  special characters."
## [1] "crime variable 10 has 0  special characters."
## [1] "crime variable 11 has 0  special characters."
## [1] "crime variable 12 has 0  special characters."
## [1] "crime variable 13 has 0  special characters."
## [1] "crime variable 14 has 0  special characters."
## [1] "crime variable 15 has 0  special characters."
## [1] "crime variable 16 has 0  special characters."
```

```
## [1] "crime variable 17 has 0  special characters."
## [1] "crime variable 18 has 0  special characters."
## [1] "crime variable 19 has 0  special characters."
## [1] "crime variable 20 has 0  special characters."
## [1] "crime variable 21 has 0  special characters."
## [1] "crime variable 22 has 0  special characters."
## [1] "crime variable 23 has 0  special characters."
## [1] "crime variable 24 has 0  special characters."
## [1] "crime variable 25 has 0  special characters."
```

We found that there are **6** special characters in **prpconv** variable, which left us 91 observations from 97. The **crmrte, crime rate** variable does not containe any of the special characters.

## Selection of Key variables

Out of 25 variables, we set

**Dependent variable** = **crmrte** crime rate

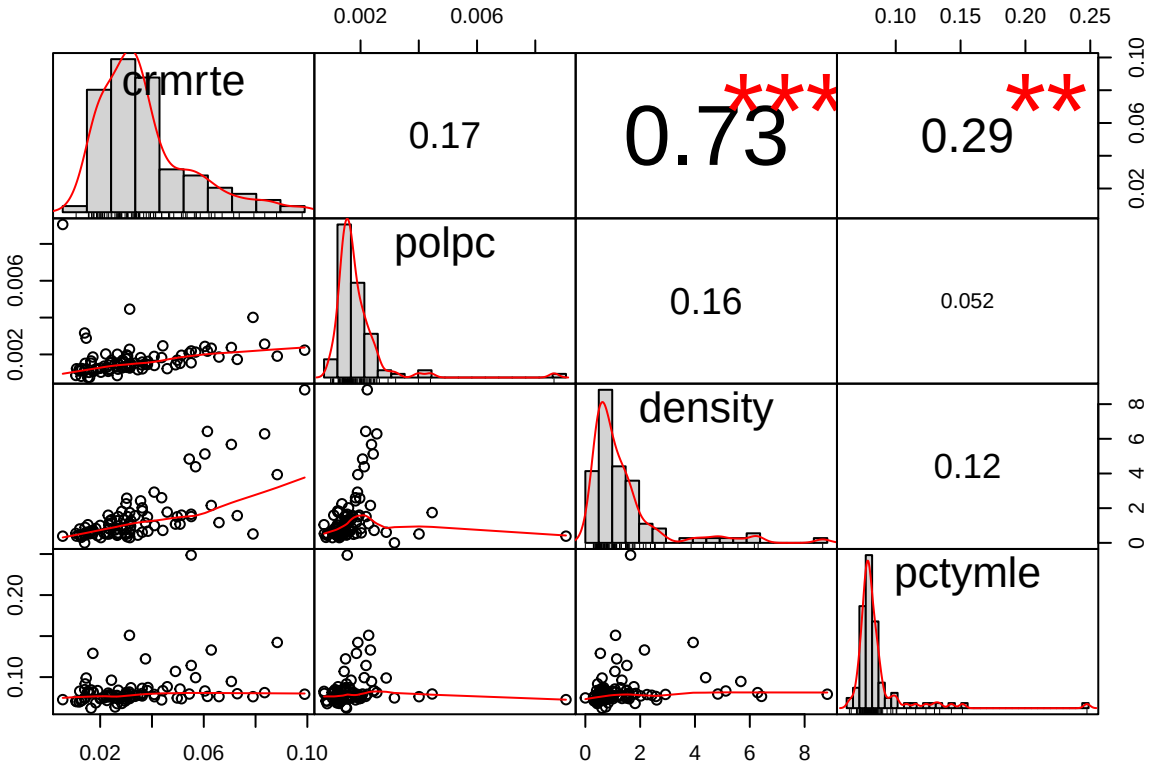
**Key independent variables**

1. **polpc** police per capita
2. **density** people per sq.mile
3. **pctymle** percent young male

We first checked key variable correlation matrix.

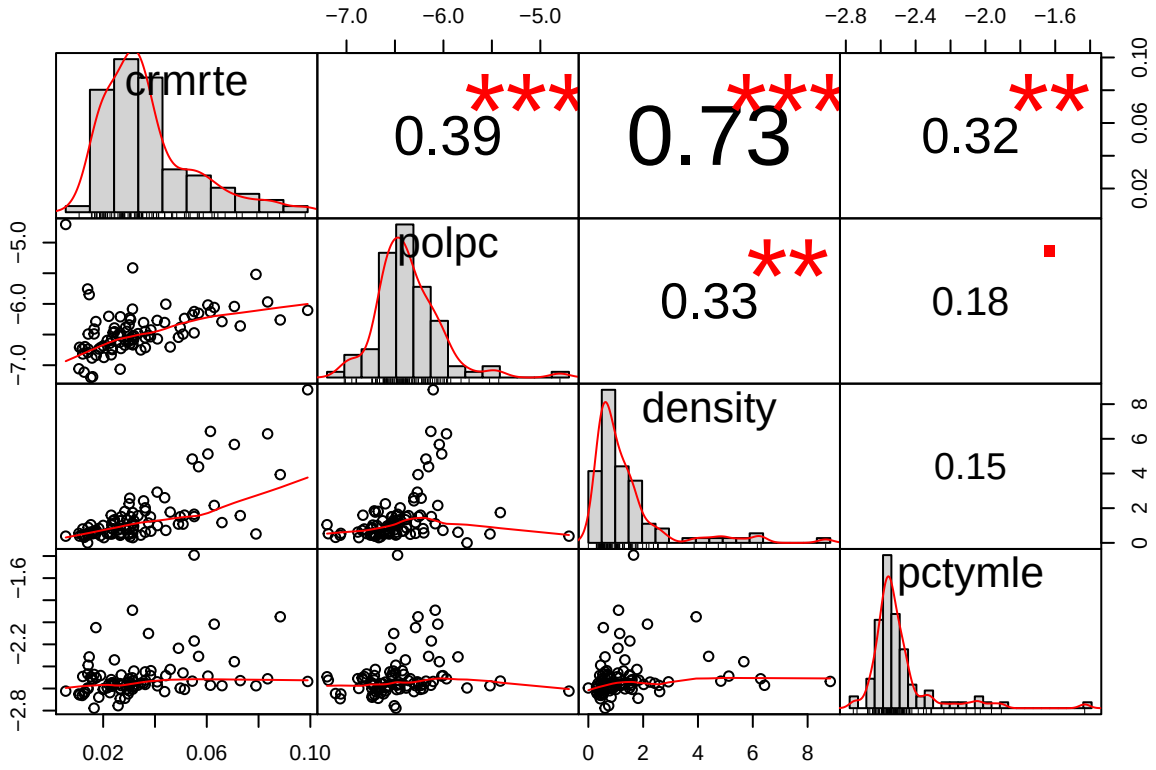
```
# transform the entire 'crime' table to numeric
crime_num = as.data.frame(lapply(crime, as.numeric))

table1 = cbind(crime_num[3], crime_num[8:9], crime_num[25])
chart.Correlation(table1, histogram = TRUE, pch = 19)
```




There is a high positive correlation between `crmrte` and `density` with `0.73` with high significance. We transformed `polpc` and `pctymle` by taking log. Transformation of these variables give us a high correlation of three variables to our dependent variables `crmrte`.

```
table2 = cbind(crime_num[3], log(crime_num[8]), crime_num[9],
    log(crime_num[25]))
chart.Correlation(table2, histogram = TRUE, pch = 19)
```



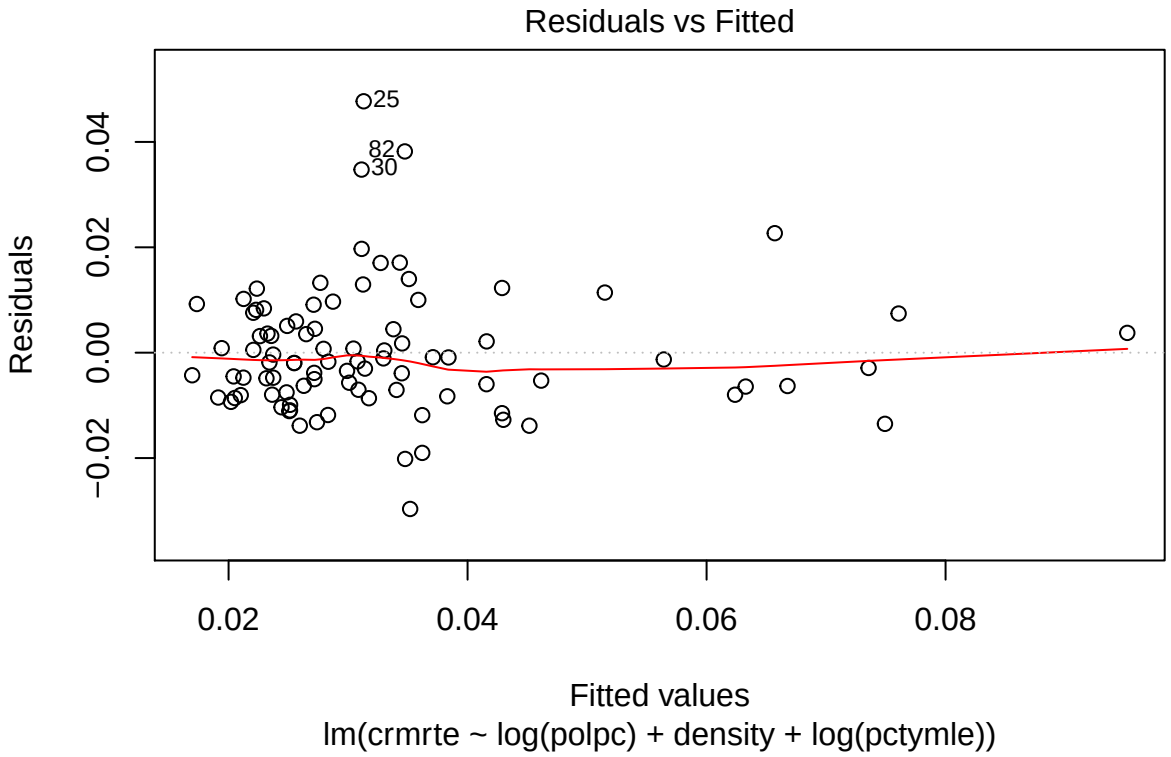

```
regress = lm(crmrte ~ log(polpc) + density + log(pctymle), data = crime)
regress$coefficients
```

```
##   (Intercept)    log(polpc)       density log(pctymle)
##  0.116175032  0.007168444  0.008099331  0.019202422
```

```
summary(regress)$r.squared
```

```
## [1] 0.5965125
```

```
plot(regress, which = 1)
```




### 3. Omitted Variable Bias (OVB)

We are concerned that the key variable we are currently interested, `polpc`, `density` and `pctymle` have other variables that are highly corrrelated to each other such as the location in North Carolia, and if there's a multicollinearity between `density` and `pctymle` which indicates the population distribution. If that's the case, we will need to modify our model to fine tune our key variables.