Lab3 Crime Statistics

Kenneth Chen, Peter Trenkwalder, Danielle Salah

7/15/2018

```r
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 propose a better policy in our political campaign on North Carolina. 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

There are 97 observations and 25 variables in our dataset. We checked if there are any empty values in each variable by applying the `!is.na` function. Interestingly, only one variable `prbconv (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`.

```
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
```

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

```
# Checking special characters such as 'a white space' etc
(apply(crime[1:25], MARGIN = 2, FUN = function(x) sum(x %in%
    c("`", "", "?", "!", "@", "#", "$", "%", "^", "&", "*", "(",
        ")"))))
```

```
##   county     year   crmrte   prbarr  prbconv  prbpris   avgsen    polpc
##        0        0        0        0        6        0        0        0
##  density    taxpc     west  central    urban pctmin80     wcon     wtuc
##        0        0        0        0        0        0        0        0
##     wtrd     wfir     wser     wmfg     wfed     wsta     wloc      mix
##        0        0        0        0        0        0        0        0
##  pctymle
##        0
```

We found that there are `6` special characters in `prbconv` variable, which left us 91 observations from 97. The rest of the variables do not contain special characters. Further check upon `prbconv` shows that the variable contains `5` white space and a special character `backtick`, ‘.

Before we continue our analysis, we removed all empty rows and changed the variable type into `numeric` for developing our model.

```
# So 97 observations end up at 91 observations.
crime_full = crime[complete.cases(crime), ]

# Changing the data type into 'numeric' for our data analysis
crime_num = as.data.frame(lapply(crime_full, as.numeric))
```

Selection of Key variables

Out of 25 variables, in order to understand the `key determinants` of the crime, we set our

1. **Dependent variable** = `crmrte` crime rate
2. **Key independent variables**

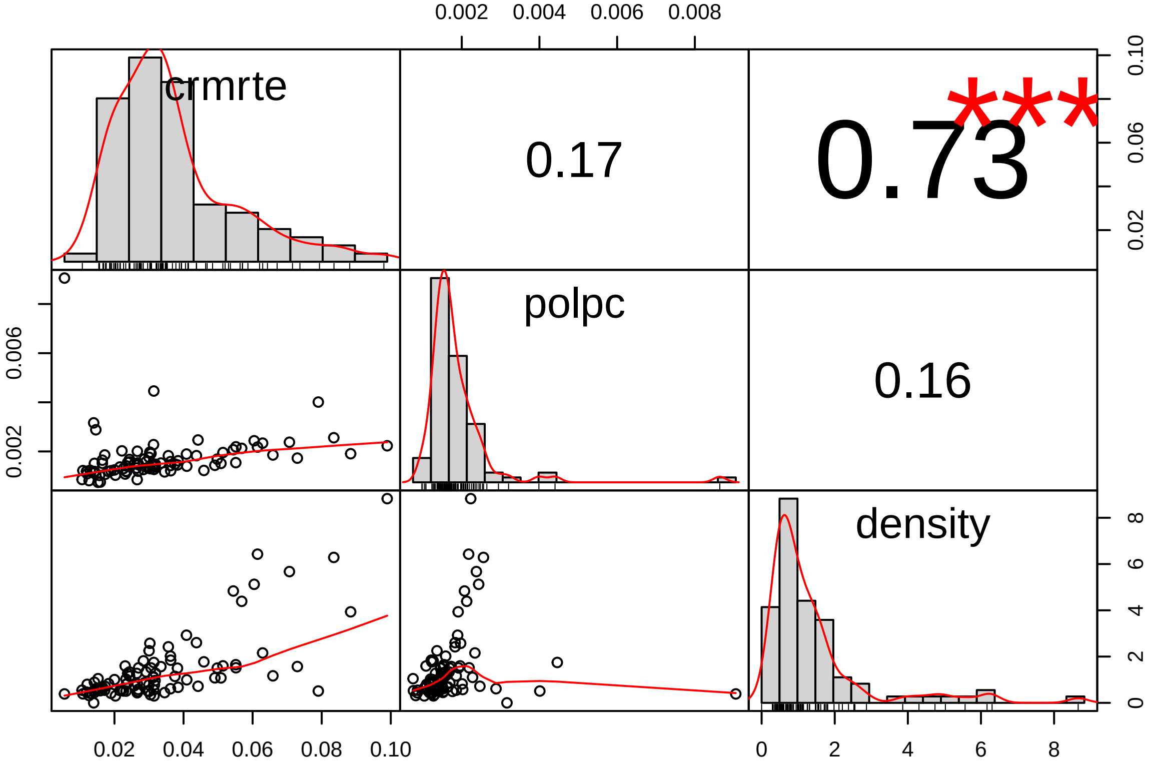
- `polpc` police per capita
- `density` people per sq.mile

Note

Our preliminary analysis on other variables such as weekly wages in different sectors such as transportations, utility, manufacturing, federal employess did not have any convincing effect on the crime rate yet. Therefore we presented our focus on two key explanatory variables here first: `polpc` and `density`.

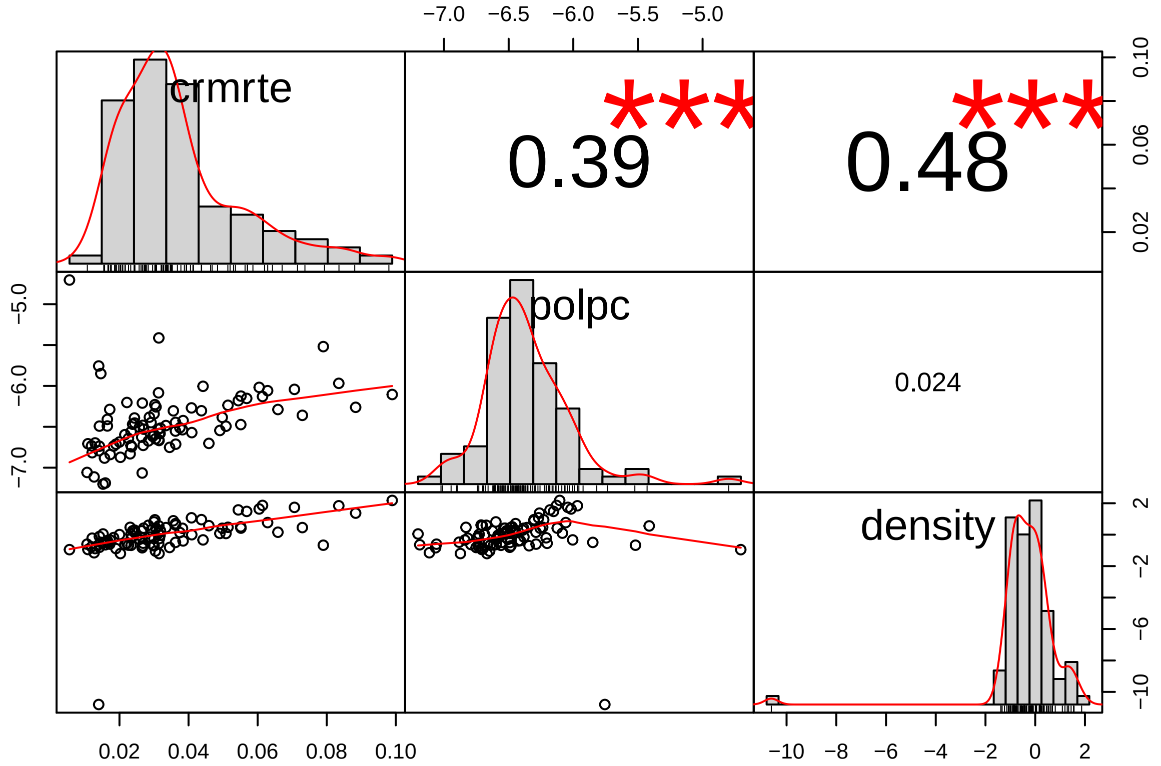
Correlation matrix between 3 variables (dependent and independents)

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



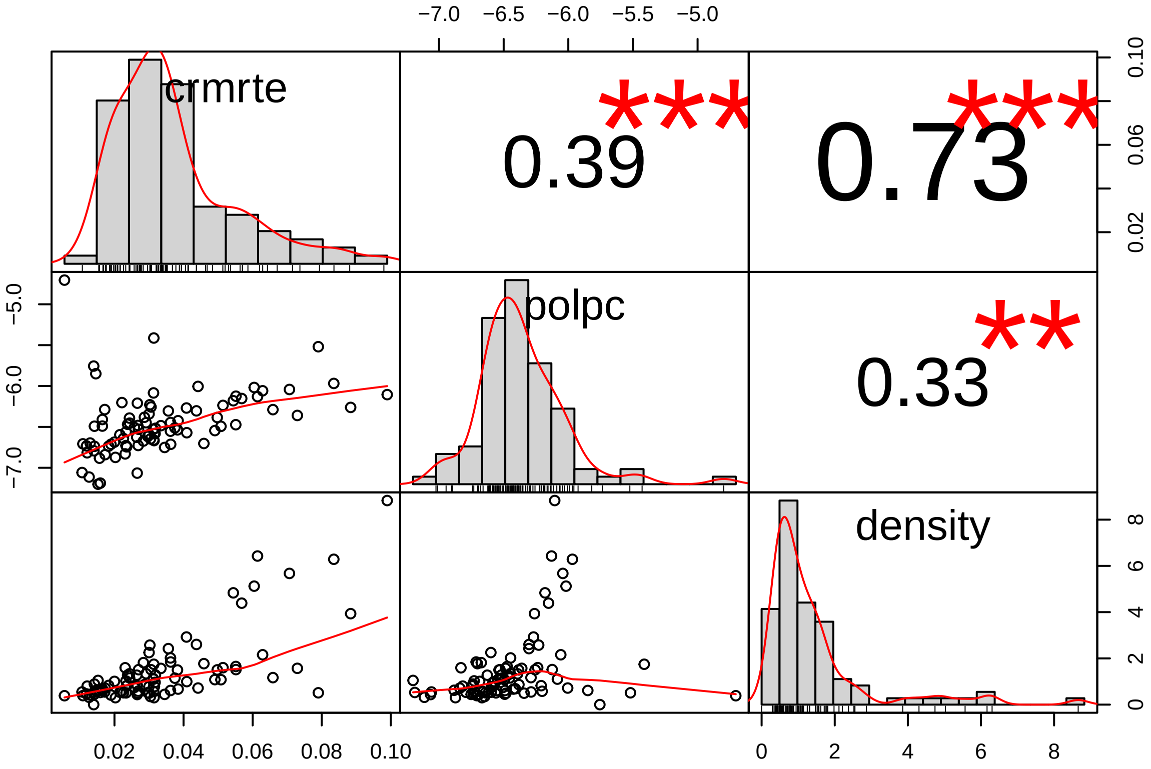
We observed that there is a high positive correlation between `crmrte` and `density` (`0.73` with high significance). There is also a minimal correlation between crime rate and the police per capita. We further checked the correlation after transforming independent variables into log.

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



Two key variables after log transformation show that there is a high correlation between `crmrte` and `polpc` with `0.39` with a clear scatterplot showing a linear relationship. However the initial correlation between `crmrte` and `density` dropped from `0.79` to `0.48` with the scatter plot going upwards. We therefore assumed that the `log(polpc)` variable gave us a better linear relationship to our dependent variable whereas ordinary data on `density` is more linear relationship with `crmrte`. We will further checked upon our assumption below.

```
table3 = cbind(crime_num[3], log(crime_num[8]), crime_num[9])
chart.Correlation(table3, histogram = TRUE, pch = 19)
```



Regression on log(polpc) and density

```
regress1 = lm(crmrte ~ log(polpc) + density, data = crime)
regress1$coefficients
```

```
## (Intercept)  log(polpc)     density
##  0.07722137  0.00863000  0.00835829
```

```
summary(regress1)$r.squared
```

```
## [1] 0.5575867
```

Developing Base Model (Model I) with TWO key variables

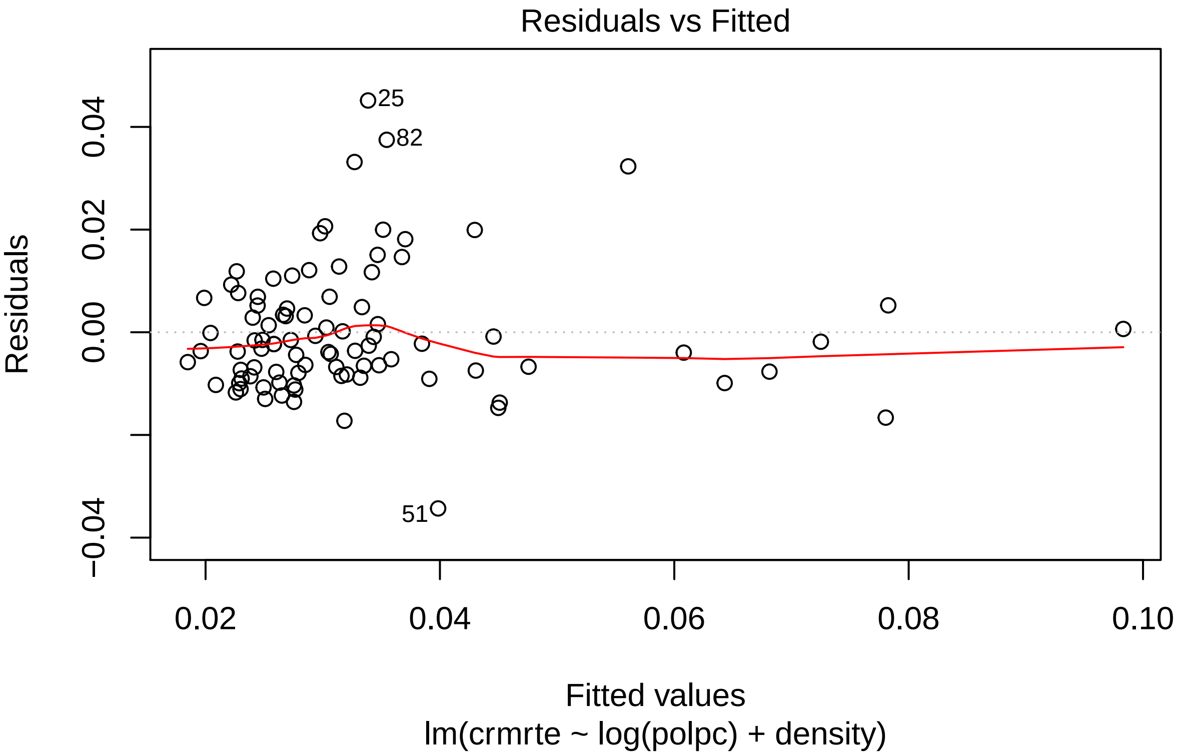
Our two key variables have $R^2 = 0.56$.

$\widehat{\text{crmrte}} = \beta_0 + \beta_1 \cdot \log(\text{polpc}) + \beta_2 \cdot \text{density}$
$\widehat{\text{crmrte}} = 0.08 + 0.01 \cdot \log(\text{polpc}) + 0.01 \cdot \text{density}$

Interestingly, we also observed that after log transformation, there is a postive correlation `0.33` between `polpc` and `density`, which reflects **`multicollinearity`**. We will explore multicollinearity in our next data analysis.

Checking if the coefficients are unbiased by the redisuals and fitted parameters

```r
plot(regress1, which = 1)
```



We found that our coefficients are unbiased because the residuals and fitted plot shows that all observations are equally spread out above and below our fitted line (`red line`).

Modifying our base model with additional variable

Omitted Variable Bias (OVB)

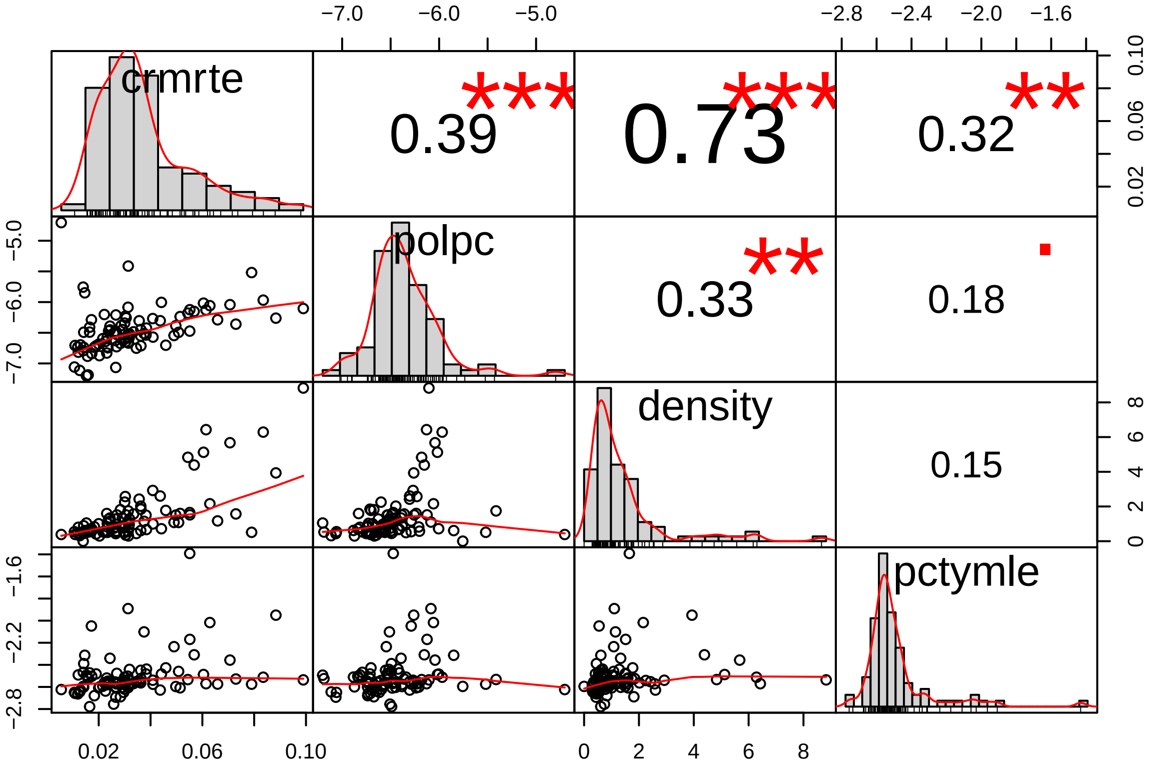
We are concerned that the key variables we are currently interested, `polpc` and `density` have other variables that are highly corrrelated to each other such as the location in North Carolia, and if there's a multicollinearity between `polpc` and `density` which indicates the population distribution.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

We first explored if there's any OVB in one of our key variables `density` because the `density` variable represents the entire population per squared mile. However we do not have any information on if the majority of crime were committed by any person in the given population. Since there is one variable in our dataset `pctymle` **percent of young males**, we can now explore another key variable in our model (Model II).

Developing a more accurate model

We now added another variable `pctymle` in our regression model. We observed that transforming `pctymle` into log shows a clear normal distribution. We show our analysis here

```
table4 = cbind(crime_num[3], log(crime_num[8]), crime_num[9],
    log(crime_num[25]))
chart.Correlation(table4, histogram = TRUE, pch = 19)
```



Regression on THREE key variables

```
regress2 = lm(crmrte ~ log(polpc) + density + log(pctymle), data = crime)
regress2$coefficients
```

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

```
summary(regress2)$r.squared
```

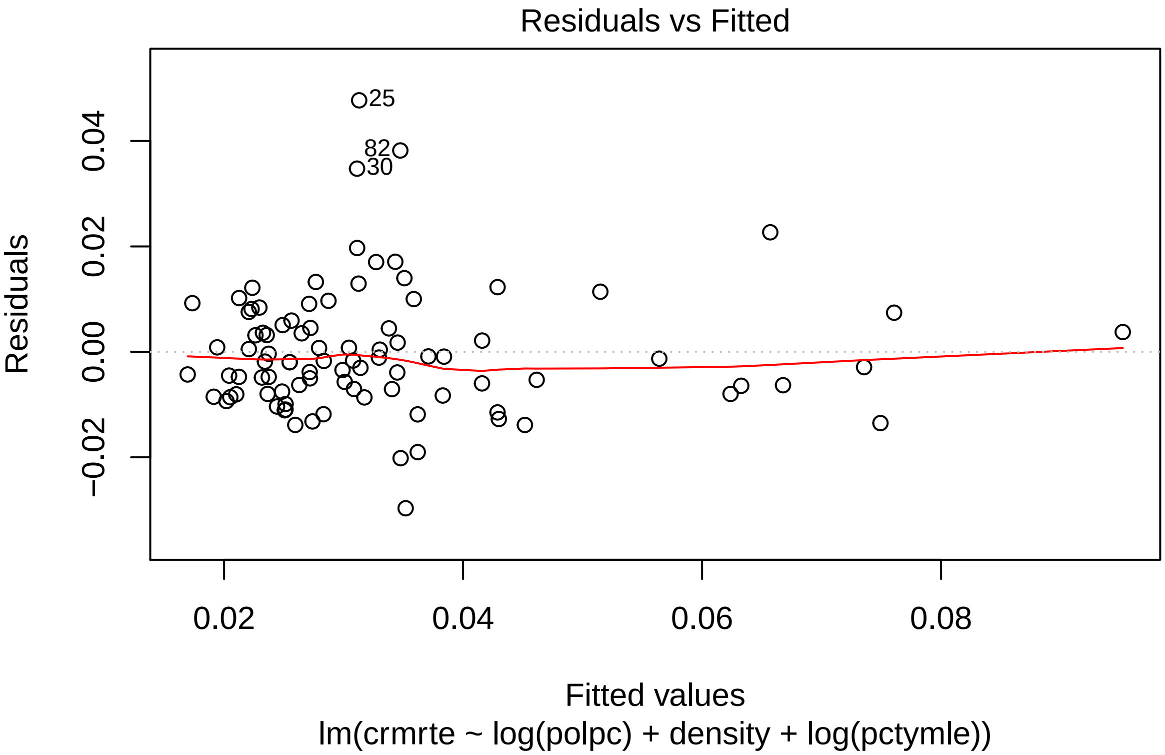
```
## [1] 0.5965125
```

Developing more accurate model (Model II) with THREE key variables

With three key variables, our model shows an improvement in R^2. Our three key variables have $R^2 = 0.60$.
$\widehat{\text{crmrte}} = \beta_0 + \beta_1 \cdot \log(\text{polpc}) + \beta_2 \cdot \text{density} + \beta_3 \cdot \text{logpctymle}$
$\widehat{\text{crmrte}} = 0.08 + 0.01 \cdot \log(\text{polpc}) + 0.01 \cdot \text{density} + 0.02 \cdot \log(\text{pctymle})$

Checking if the coefficients are unbiased by the redisuals and fitted parameters

```
plot(regress2, which = 1)
```



Our modified model indicates that three coefficients for three key variables are `unbiased`.