Datta_Saurav_Lab3_Draft

Saurav Datta

3/24/2018

```r
#install.packages("sqldf")
# Sys.setenv(JAVA_HOME='/Library/Java/JavaVirtualMachines/jdk1.8.0_151.jdk/Contents/Home')
# install.packages("rJava")
# install.packages("RH2")
#install.packages("gridExtra")
```

```r
library(sqldf)
```

```
## Loading required package: gsubfn

## Warning: package 'gsubfn' was built under R version 3.4.4

## Loading required package: proto

## Loading required package: RSQLite
```

```r
# library(RH2)
library(ggplot2)
library(gridExtra)
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
library(car)
```

```r
getwd()
```

```
## [1] "/Users/sdatta/Documents/1. Personal/MIDS/W203/Course material/Lab3"
```

```r
setwd("/Users/sdatta/Documents/1. Personal/MIDS/W203/Course material/Lab3")
```

```r
#db = dbConnect(SQLite(), dbname="lab3.sqllite")
#sqldf("attach 'lab3.sqllite' as new")
```

```r
#dbRemoveTable(db, "crime0")
```

```r
crime0=read.csv("crime_v2.csv",
                    header = TRUE
                  )
crime1=crime0
```

```r
sqldf("select * from crime1 limit 5")
```

```
##   county year    crmrte   prbarr     prbconv  prbpris avgsen      polpc
## 1      1   87 0.0356036 0.298270 0.527595997 0.436170   6.71 0.00182786
## 2      3   87 0.0152532 0.132029 1.481480002 0.450000   6.35 0.00074588
## 3      5   87 0.0129603 0.444444 0.267856985 0.600000   6.76 0.00123431
## 4      7   87 0.0267532 0.364760 0.525424004 0.435484   7.14 0.00152994
## 5      9   87 0.0106232 0.518219 0.476563007 0.442623   8.22 0.00086018
```

```
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 1 2.4226327 30.99368    0       1     0 20.21870 281.4259 408.7245
## 2 1.0463320 26.89208    0       1     0  7.91632 255.1020 376.2542
## 3 0.4127659 34.81605    1       0     0  3.16053 226.9470 372.2084
## 4 0.4915572 42.94759    0       1     0 47.91610 375.2345 397.6901
## 5 0.5469484 28.05474    1       0     0  1.79619 292.3077 377.3126
##       wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 1 221.2701 453.1722 274.1775 334.54 477.58 292.09 311.91 0.08016878
## 2 196.0101 258.5650 192.3077 300.38 409.83 362.96 301.47 0.03022670
## 3 229.3209 305.9441 209.6972 237.65 358.98 331.53 281.37 0.46511629
## 4 191.1720 281.0651 256.7214 281.80 412.15 328.27 299.03 0.27362204
## 5 206.8215 289.3125 215.1933 290.89 377.35 367.23 342.82 0.06008584
##      pctymle
## 1 0.07787097
## 2 0.08260694
## 3 0.07211538
## 4 0.07353726
## 5 0.07069755
```

Converting prbconv from factor to numeric

We see that column prbconv is factor datatype

```r
crime1$prbconv_cast=as.numeric(as.matrix(crime1$prbconv))
```

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

```r
crime_tmp=sqldf("SELECT * FROM crime1 WHERE NOT (prbconv_cast >1 OR prbarr >1 OR prbpris >1 OR prbconv_c
crime1=crime_tmp
sqldf("SELECT count(*) from crime1")
```

```
##   count(*)
## 1       81
```

Defining common function

```r
f_check_null <- function(in_field_name ){
  sql=sprintf("SELECT COUNT(1) as COUNT_NULL_OR_NA FROM crime1 WHERE (%s IS \"NA\" or %s IS NULL)", in_f
  sqldf(sql)
}
```

```r
f_plot_one <- function(in_db_field_name,in_main_title ){

  title_log=paste("log of",in_main_title, sep = " ")

  par(mfrow=c(2,2))
  hist(in_db_field_name, main=in_main_title)
  hist(log(in_db_field_name), main=title_log)
  boxplot(in_db_field_name, main=in_main_title)

}
```

```r
f_plot_two <- function(in_field_name1,in_xlabel,in_field_name2,in_y_label, in_main_title ){


  theme_update(plot.title = element_text(hjust = 0.5))
```

```r
  p1<-ggplot(crime1, aes_string(in_field_name1,in_field_name2)) +
         geom_point() +
         geom_smooth(na.rm = FALSE, method = loess)
  p1 + ggtitle(in_main_title) +xlab(in_xlabel) + ylab(in_y_label)
}
```

```r
f_plot_three  <- function(in_field_x,in_xlabel,in_field_y,in_y_label){

corr_val=round(cor(in_field_y, in_field_x),4)

main_title=paste(in_xlabel,'v/s',in_y_label, sep = ' ')
plot(in_field_x, in_field_y,
     main = main_title,
      sub=paste("Corr. coefficient:",corr_val),
      xlab=in_xlabel,
      ylab=in_y_label)
m = lm( in_field_y ~ in_field_x)
abline(m)

}
```

Analyzing regions

```r
crime_tmp = sqldf("SELECT *, CASE WHEN west=1 THEN \'WEST\'
                                WHEN central=1 THEN \'CENTRAL\'
                                WHEN urban=1 THEN \'URBAN\'
                                ELSE \'UNKNOWN\'
                          END regionofcrime
                FROM crime1"
)
```

```r
crime1=crime_tmp
```

```r
sqldf("SELECT regionofcrime as regionofcrime, count(8) as countofcrimes from crime1 GROUP BY regionofcri
     )
```

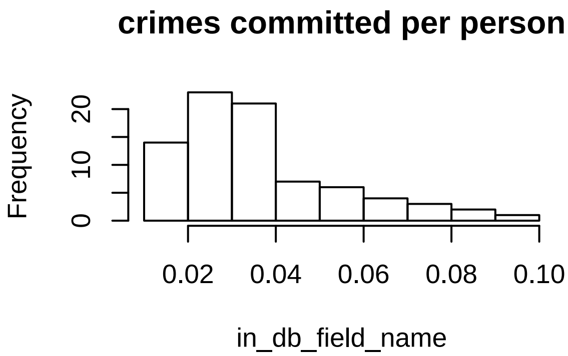
```
##   regionofcrime countofcrimes
## 1       CENTRAL            31
## 2       UNKNOWN            29
## 3         URBAN             2
## 4          WEST            19
```

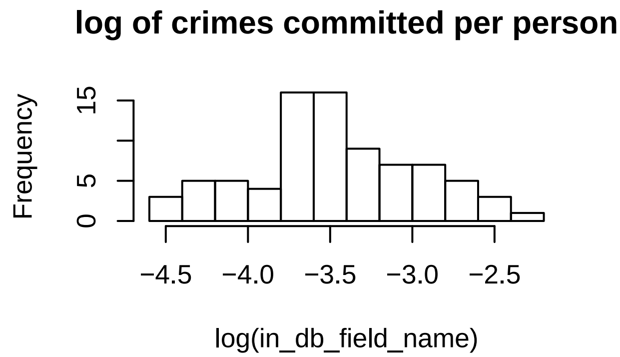
Analyzing crmrte

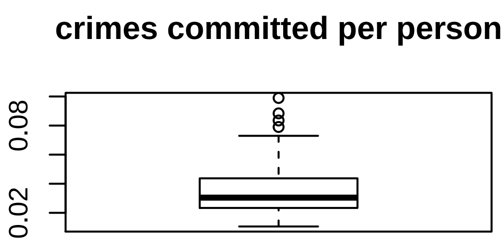
```r
f_check_null("crmrte")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```r
f_plot_one(crime1$crmrte,"crimes committed per person")
crime1$logcrmrte=log(crime1$crmrte)
```







Analyzing the 6 records with missing crmrte values

```
sqldf("SELECT * FROM crime1 WHERE (crmrte IS \"NA\" or crmrte IS NULL) ")
```

```
##  [1] county        year          crmrte        prbarr        prbconv
##  [6] prbpris       avgsen        polpc         density       taxpc
## [11] west          central       urban         pctmin80      wcon
## [16] wtuc          wtrd          wfir          wser          wmfg
## [21] wfed          wsta          wloc          mix           pctymle
## [26] prbconv_cast  regionofcrime logcrmrte
## <0 rows> (or 0-length row.names)
```

We see that all relevant columns of these 6 records are NA. So we can safely delete them

```
crime_tmp=sqldf( c("DELETE FROM crime1 WHERE crmrte IS NULL",
         "SELECT * FROM crime1"
         )
)
```

```
## Warning in rsqlite_fetch(res@ptr, n = n): Don't need to call dbFetch() for
## statements, only for queries
```

```
crime1=crime_tmp
sqldf("SELECT count(*) FROM crime1 ")
```

```
##   count(*)
## 1       81
```

Reanalyzing regions after deleting NAs

```
sqldf("SELECT regionofcrime as regionofcrime, count(8) as countofcrimes from crime1 GROUP BY regionofcr
    )
```

```
##   regionofcrime countofcrimes
## 1       CENTRAL            31
## 2       UNKNOWN            29
```

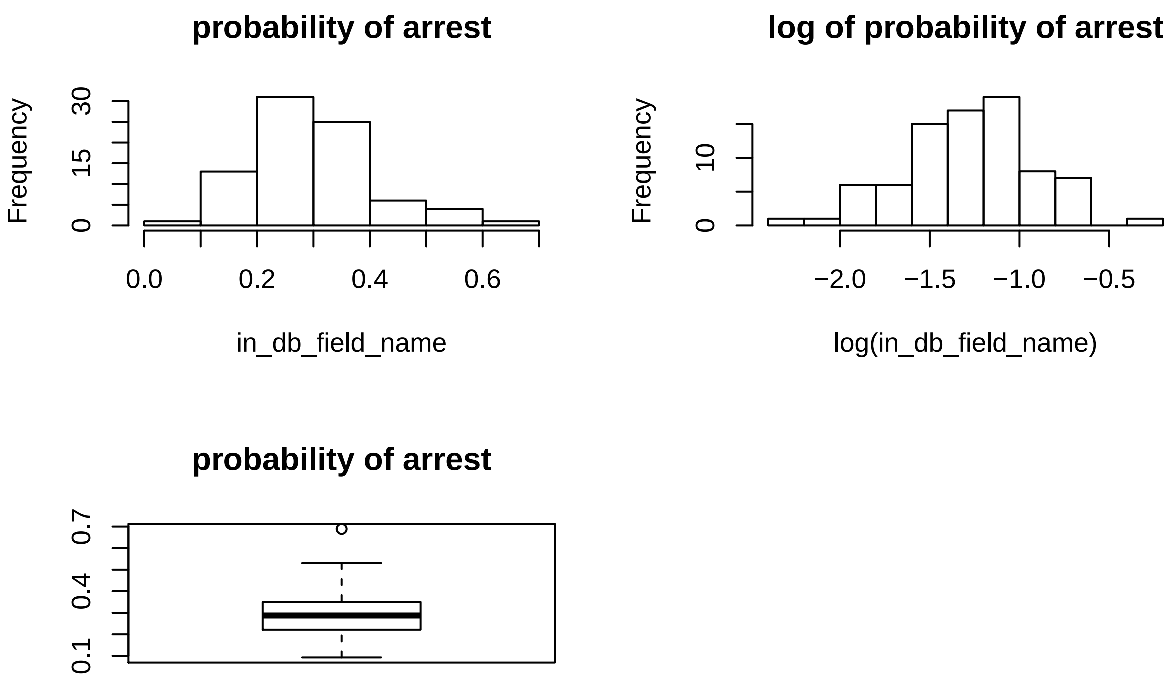
```
## 3          URBAN               2
## 4           WEST              19
```

Analyzing prbarr

```r
f_check_null("prbarr")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```r
f_plot_one(crime1$prbarr,"probability of arrest")
```

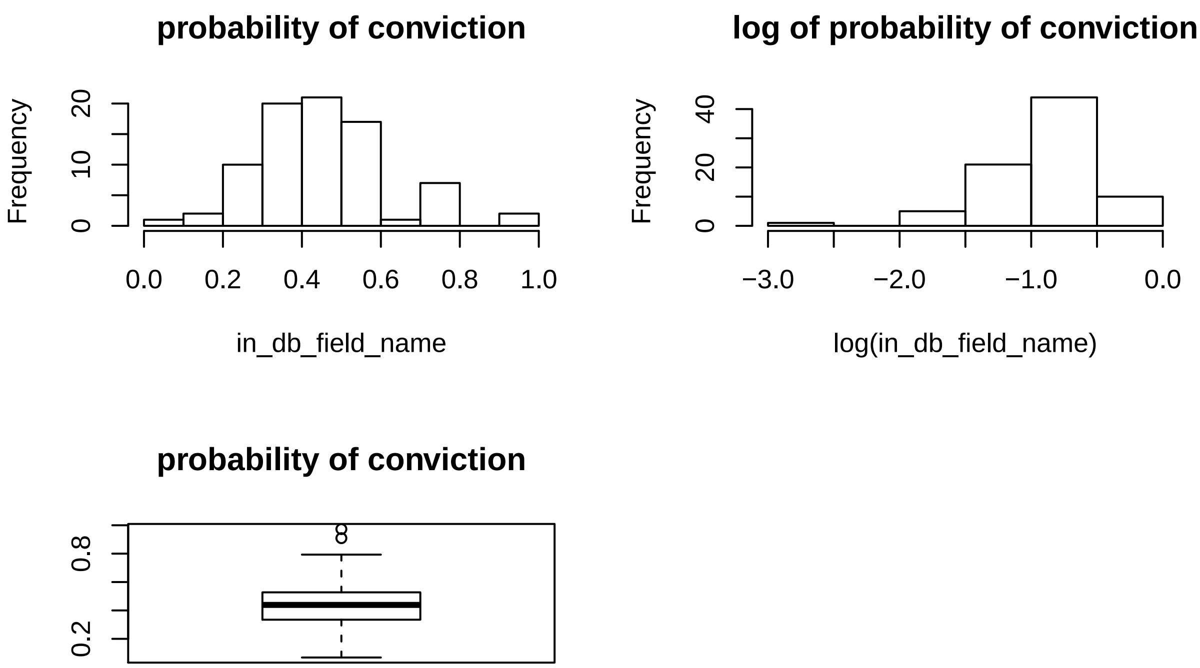


Analyzing prbconv_cast

```r
f_check_null("prbconv_cast")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```r
f_plot_one(crime1$prbconv_cast,"probability of conviction")
```

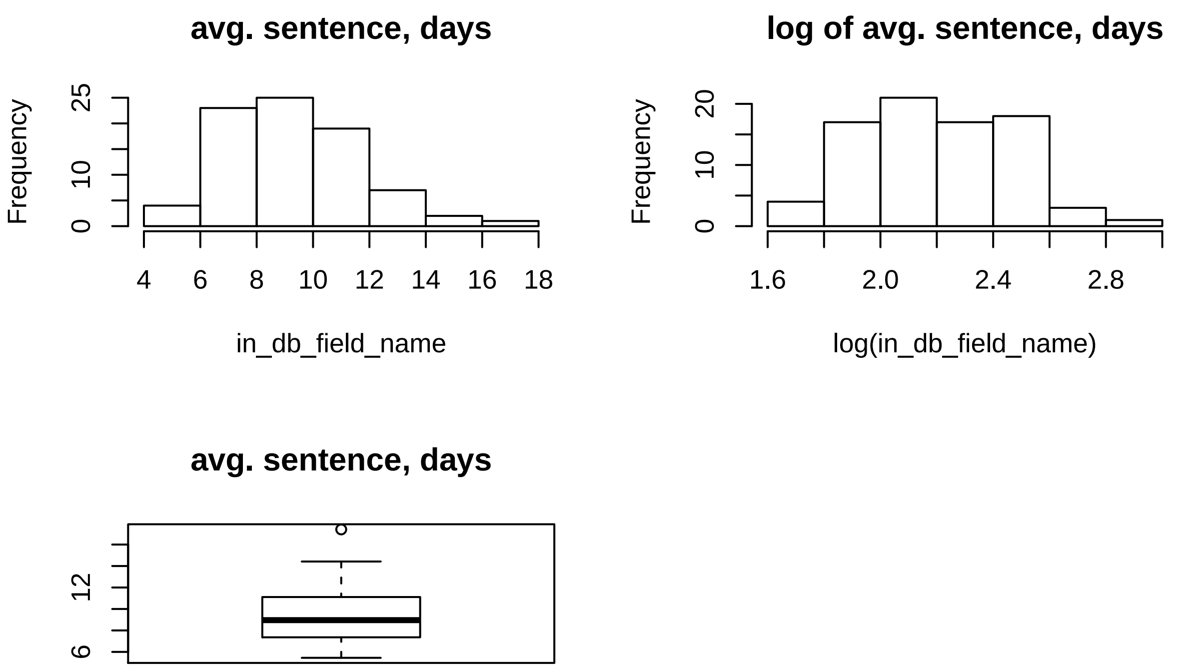


Analyzing avgsen

```
f_check_null("avgsen")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```
f_plot_one(crime1$avgsen,"avg. sentence, days")
crime1$logavgsen =  log(crime1$avgsen)
```

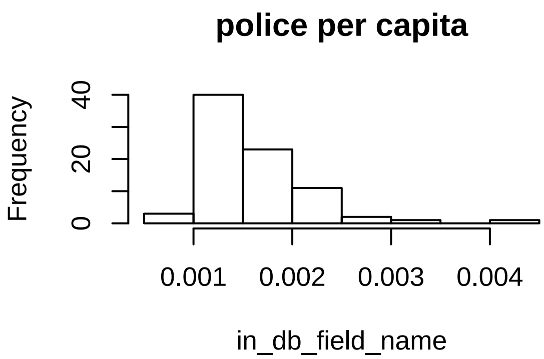


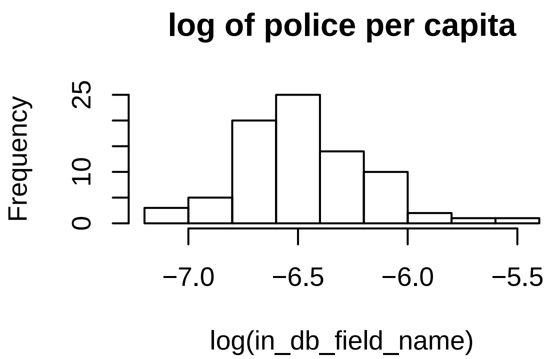
Analyzing polpc

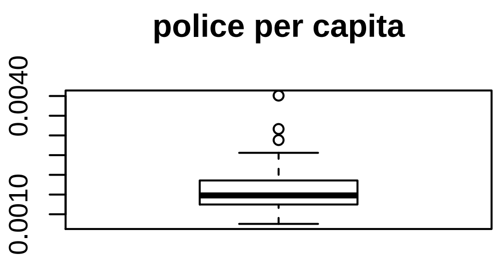
```r
f_check_null("polpc")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```r
f_plot_one(crime1$polpc,"police per capita")
crime1$logpolpc =  log(crime1$polpc)
```







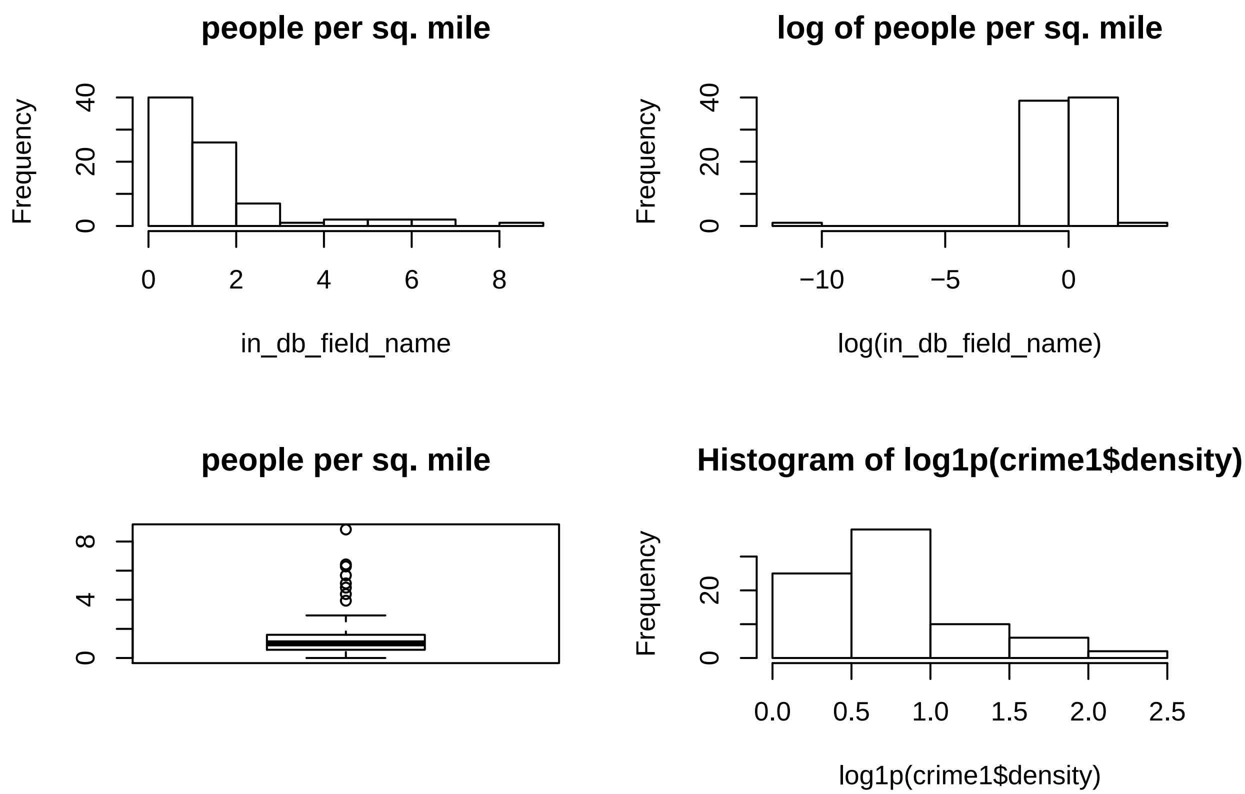
Analyzing density

We see that log1p of density is closer to normal distribution than either log or exp (tried it offline).

```r
f_check_null("density")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```r
f_plot_one(crime1$density,"people per sq. mile")

hist(log1p(crime1$density))
```



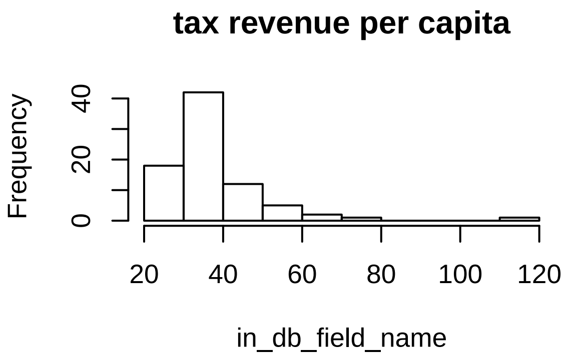

```
crime1$log1pdensity=log1p(crime1$density)
```

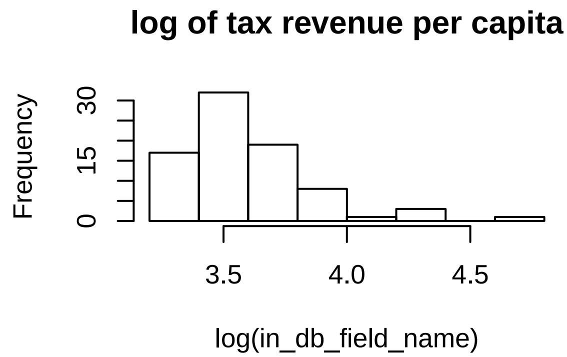
Analyzing taxpc

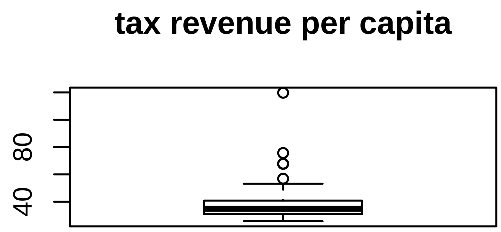
```
f_check_null("taxpc")
```

```
##   COUNT_NULL_OR_NA
## 1                0
```

```
f_plot_one(crime1$taxpc,"tax revenue per capita")
```







Outlier of taxpc=120

Analyzing region

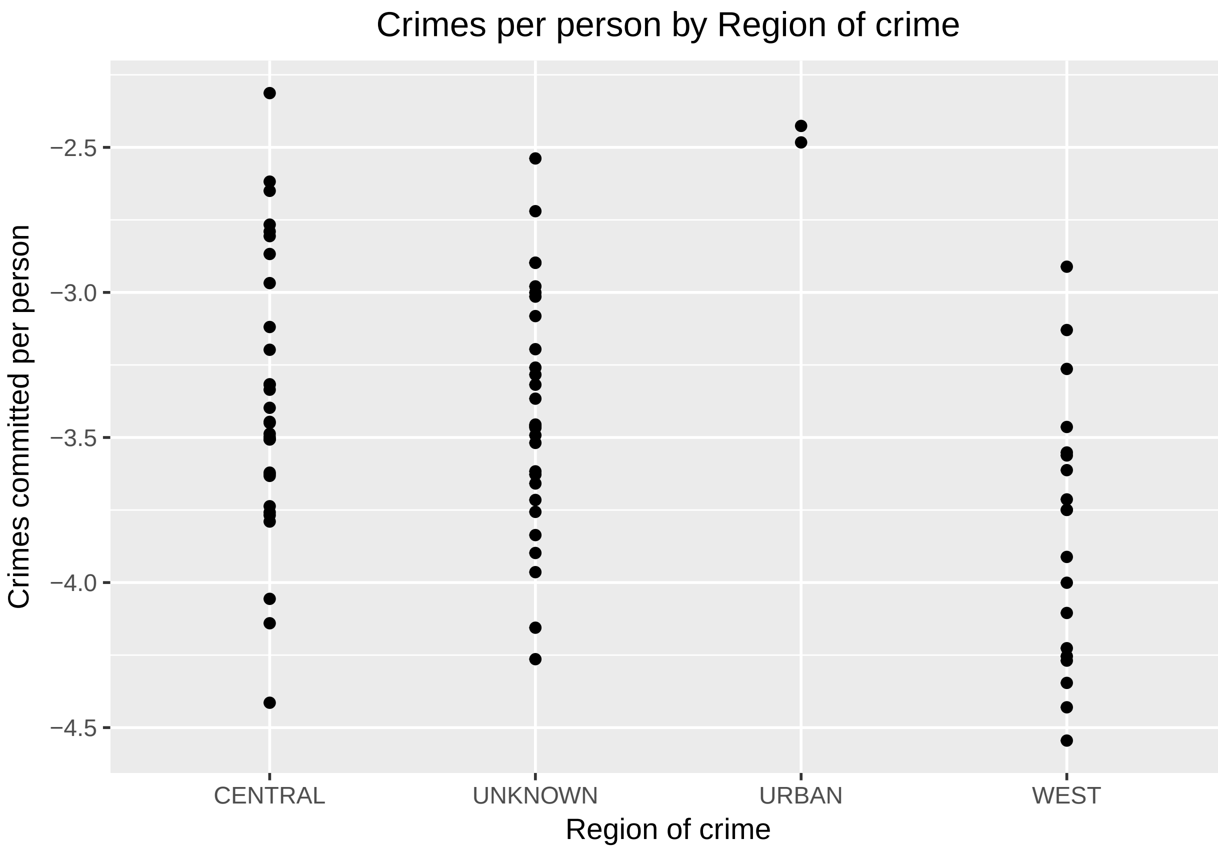
We see that there are 58 records for which we have the region, whereas there are 97 records in the dataset. So there are crimes with unknown region.

```
sqldf("select \'1. WEST\' as REGION, count(8) as COUNT from crime1 where west=1
      UNION
      select \'2. CENTRAL\' as REGION, count(8) as COUNT from crime1 where central=1
       UNION
      select \'3. URBAN\' as REGION, count(8) as COUNT from crime1 where urban=1
      UNION
      select \'4. TOTAL\' as REGION, count(8) as COUNT from crime1 where (west=1 or central=1 or urban =
     )
```

```
##       REGION COUNT
## 1    1. WEST    19
## 2 2. CENTRAL    32
## 3   3. URBAN     8
## 4   4. TOTAL    52
```

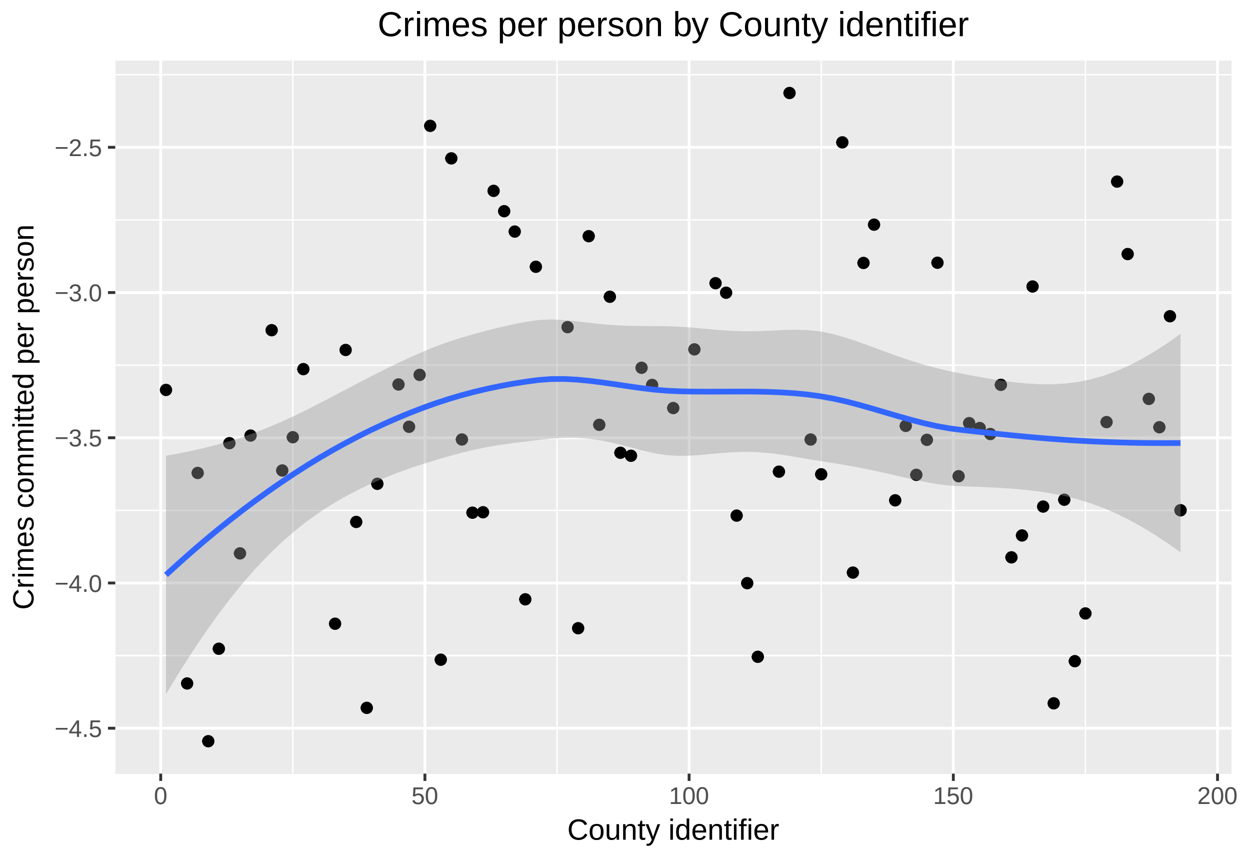
Analyzing Crimes committed per person by region

```
#
f_plot_two("regionofcrime","Region of crime","logcrmrte","Crimes committed per person","Crimes per perso
```



Analyzing Crimes committed per person by region

```
f_plot_two("county","County identifier","logcrmrte","Crimes committed per person","Crimes per person by
```




```
sqldf("SELECT county,crmrte FROM crime1 WHERE crmrte>=0.09  ")
```

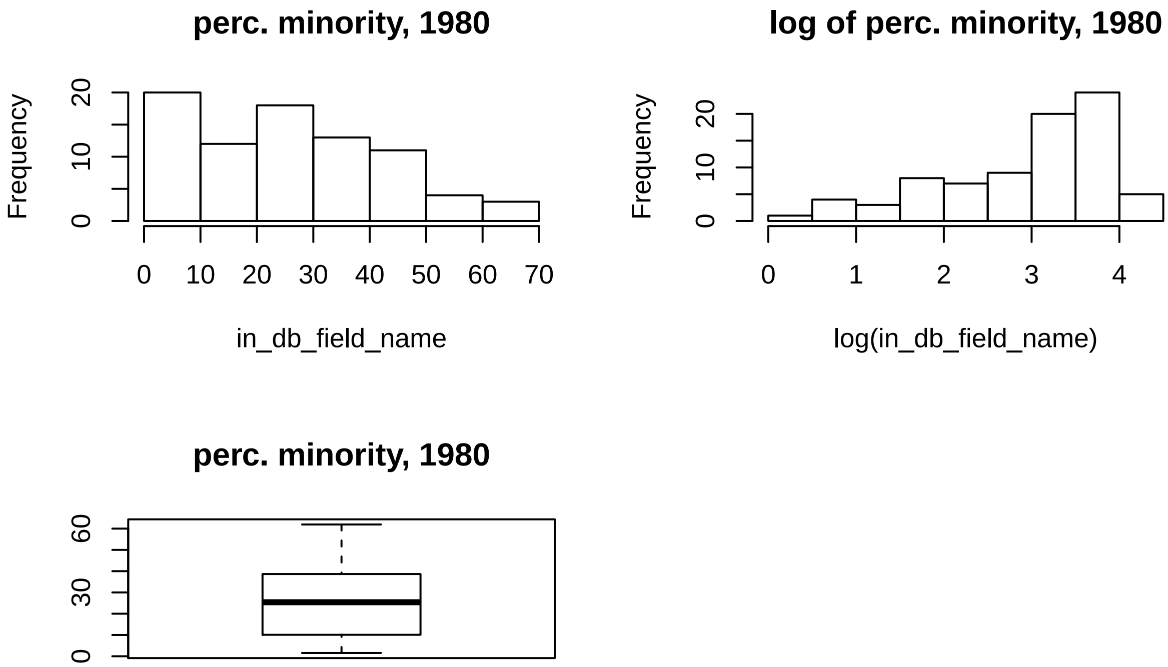
```
##   county    crmrte
## 1    119 0.0989659
```

Analyzing percent minority

```
f_check_null("pctmin80")
```

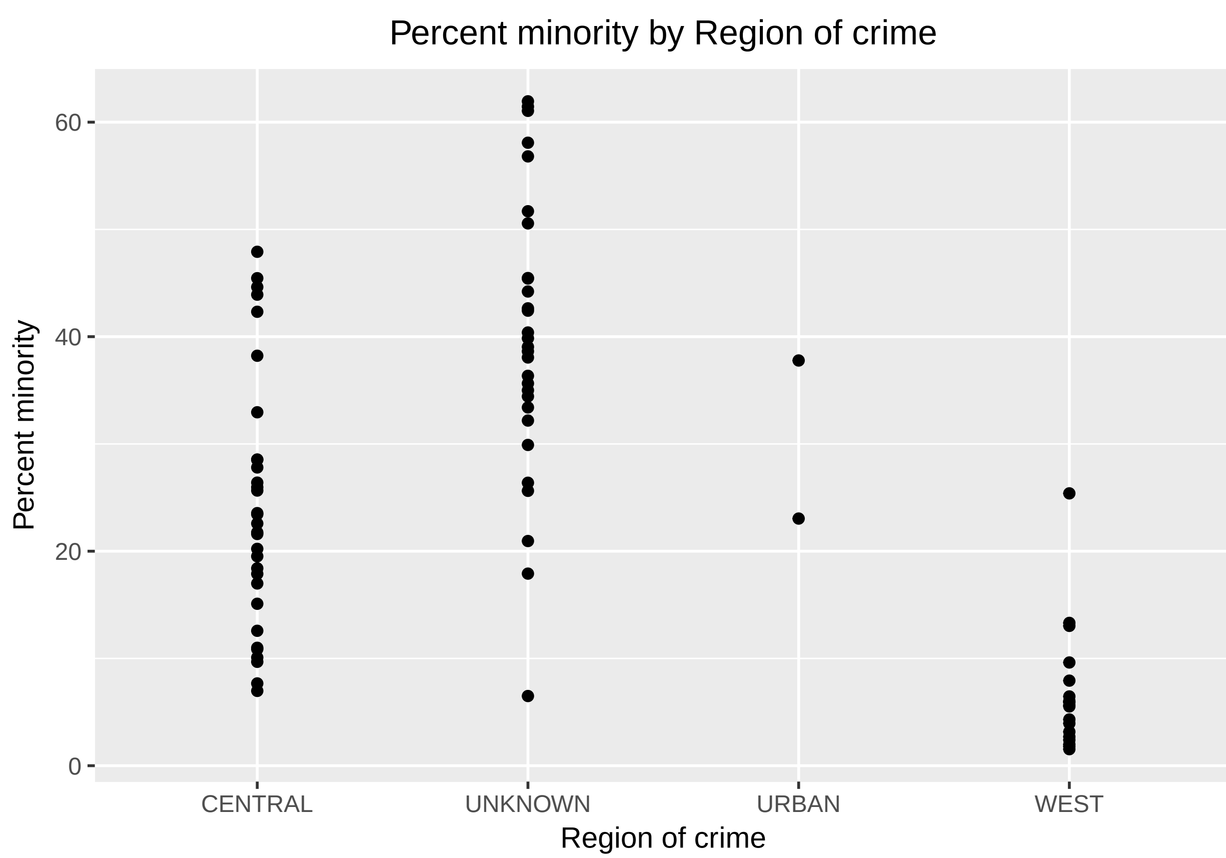
```
##   COUNT_NULL_OR_NA
## 1                0
```

```
f_plot_one(crime1$pctmin80,"perc. minority, 1980")
```



Analyzing percent minority by region

```
f_plot_two("regionofcrime","Region of crime","pctmin80","Percent minority","Percent minority by Region
```

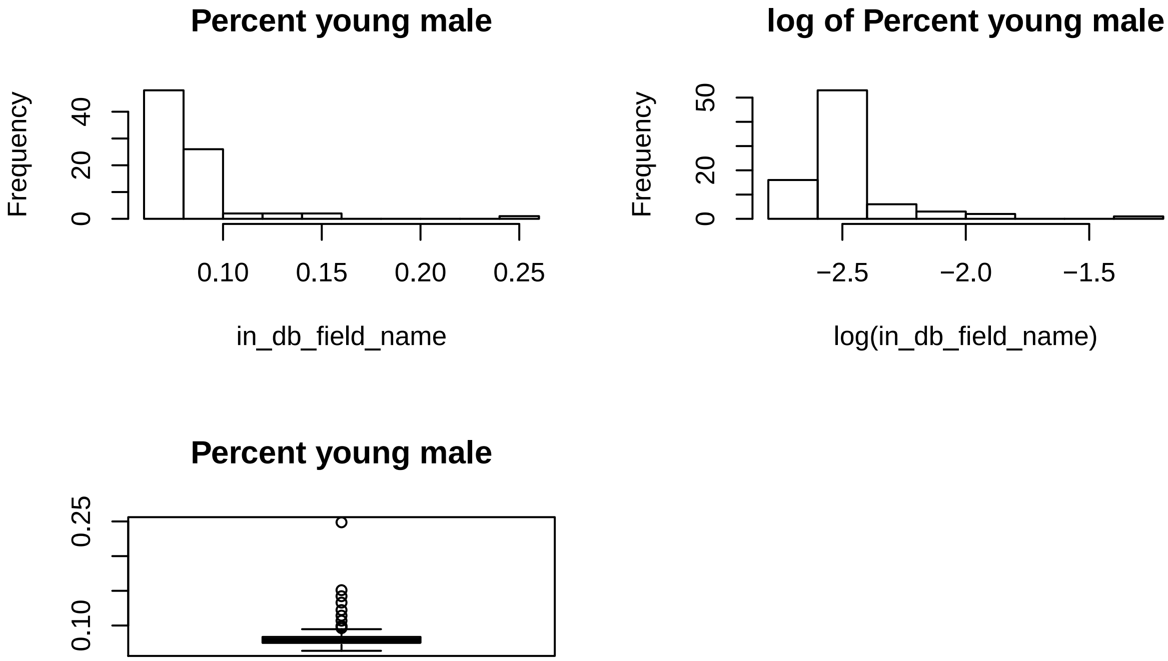


Analyzing pctymle

```
f_check_null("pctymle")
```

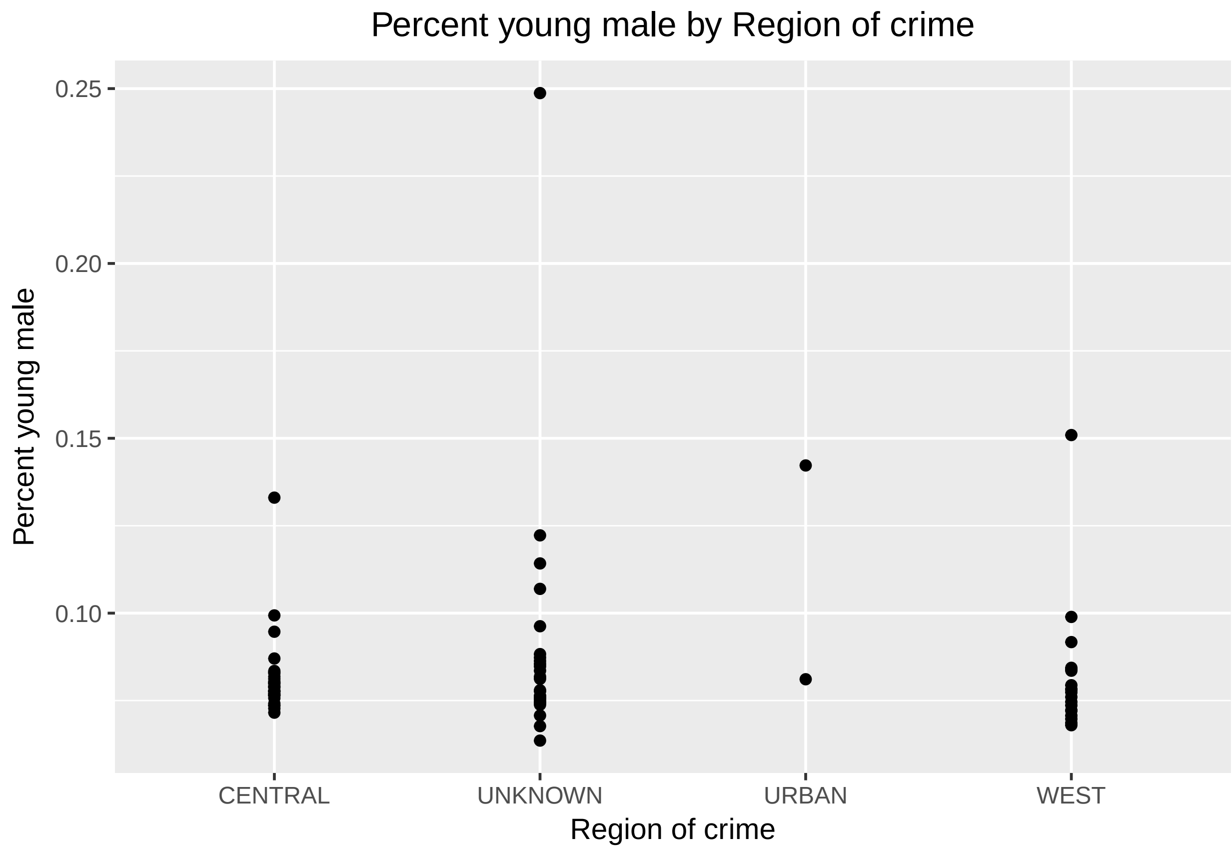
```
##   COUNT_NULL_OR_NA
## 1                0
```

```
f_plot_one(crime1$pctymle,"Percent young male")
```



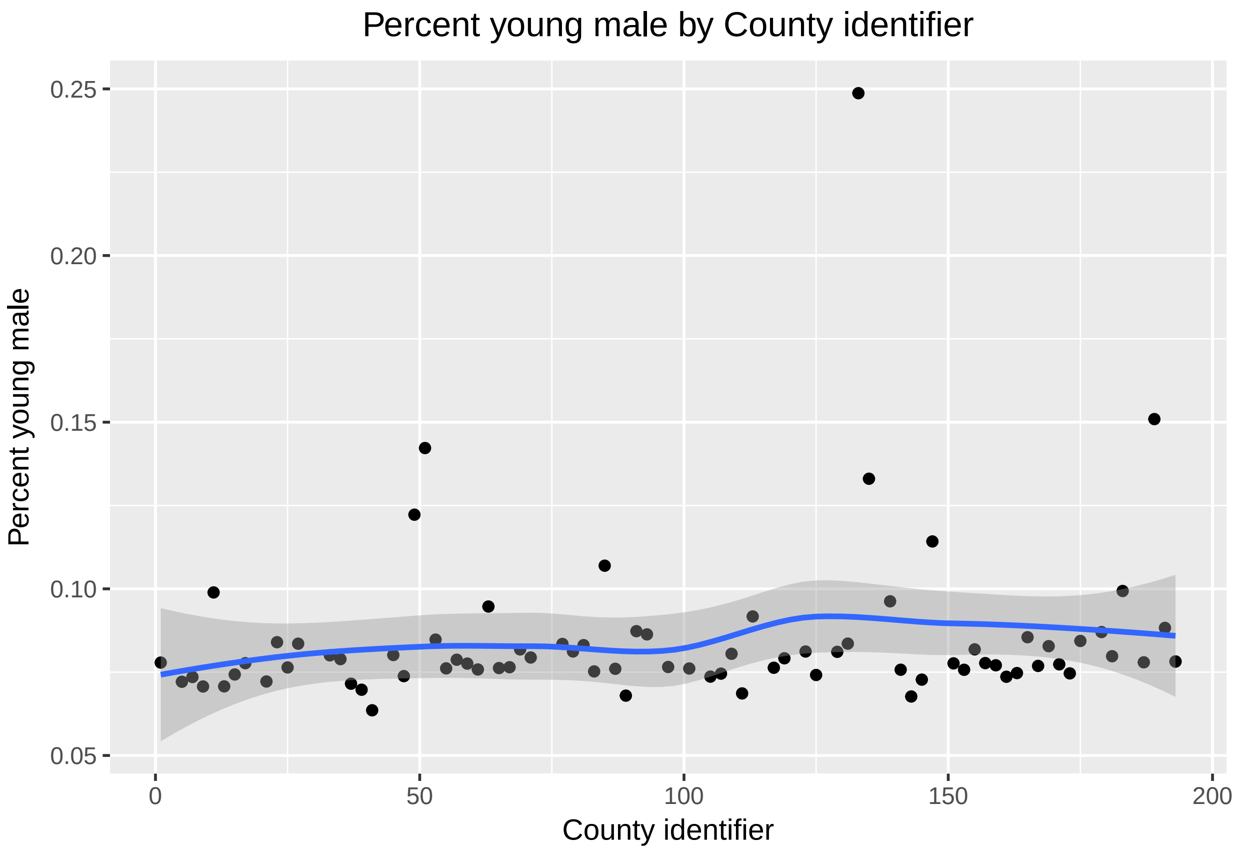
Analyzing pctymle by region

```
f_plot_two("regionofcrime","Region of crime","pctymle","Percent young male","Percent young male by Regi
```



We see that the UKNOWN region has the highest percent of young male

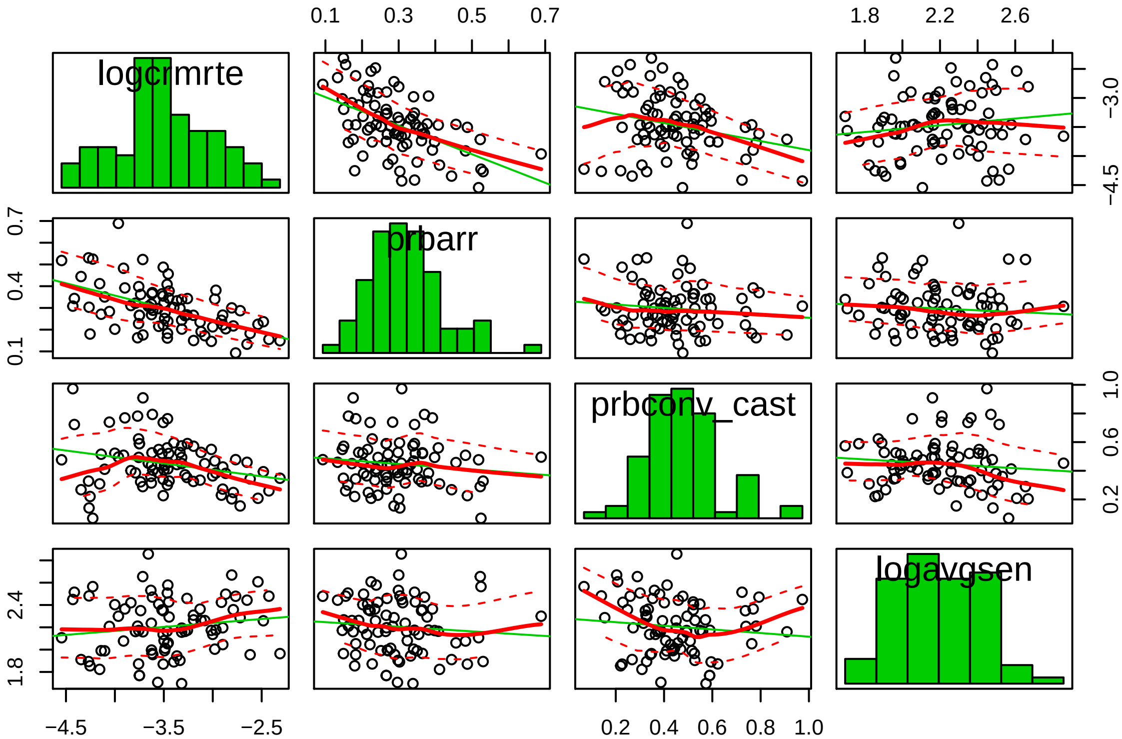
```
f_plot_two("county","County identifier","pctymle","Percent young male","Percent young male by County ide
```



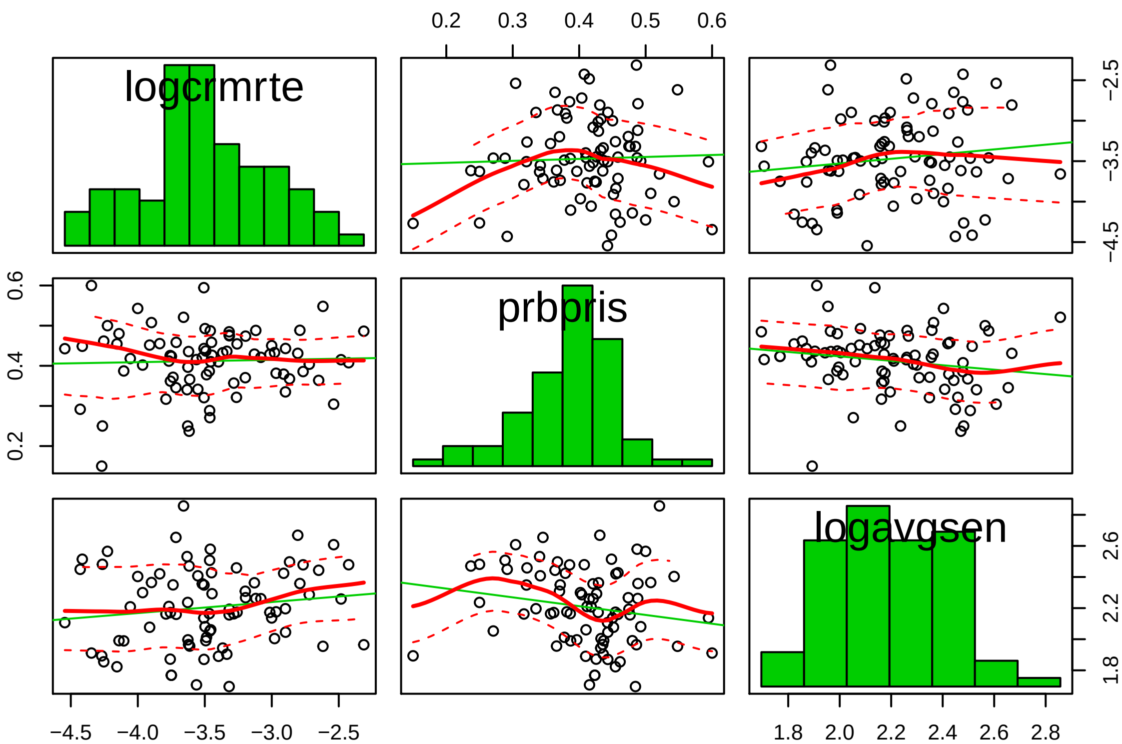

```
sqldf("SELECT county,pctymle from crime1 WHERE pctymle>=0.20")
```

```
##   county   pctymle
## 1    133 0.2487116
```

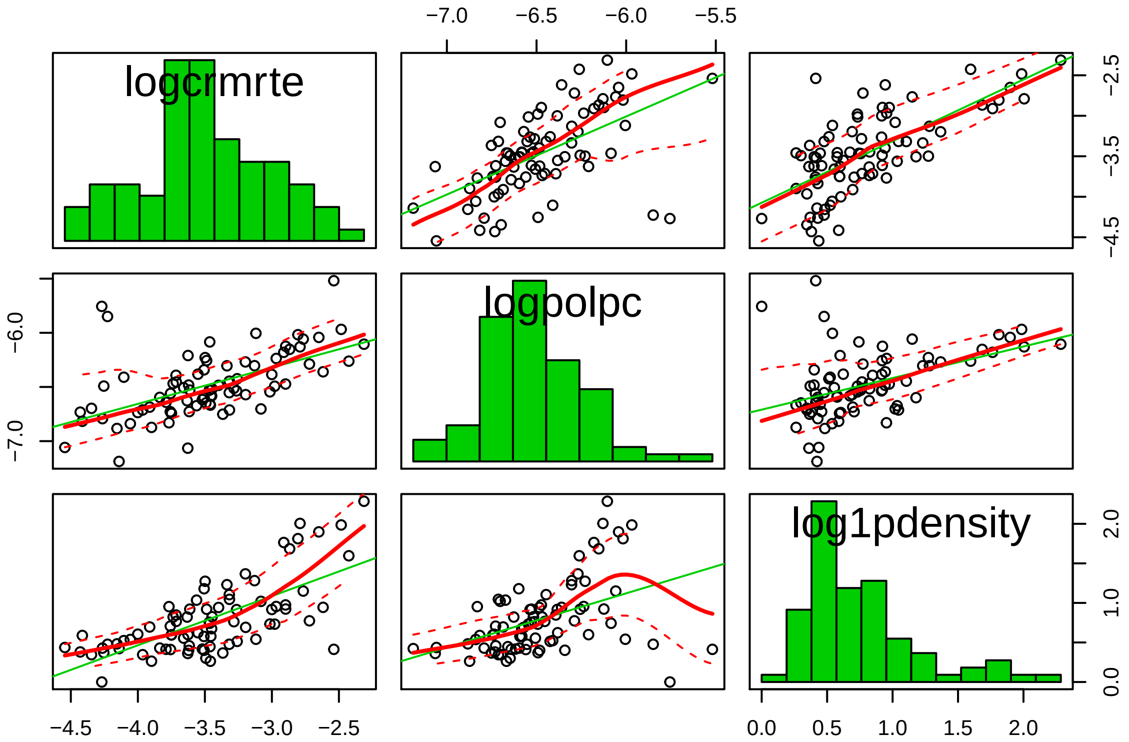
```
scatterplotMatrix(crime1[,c("logcrmrte","prbarr",  "prbconv_cast", "logavgsen")], diagonal = "histogram
```



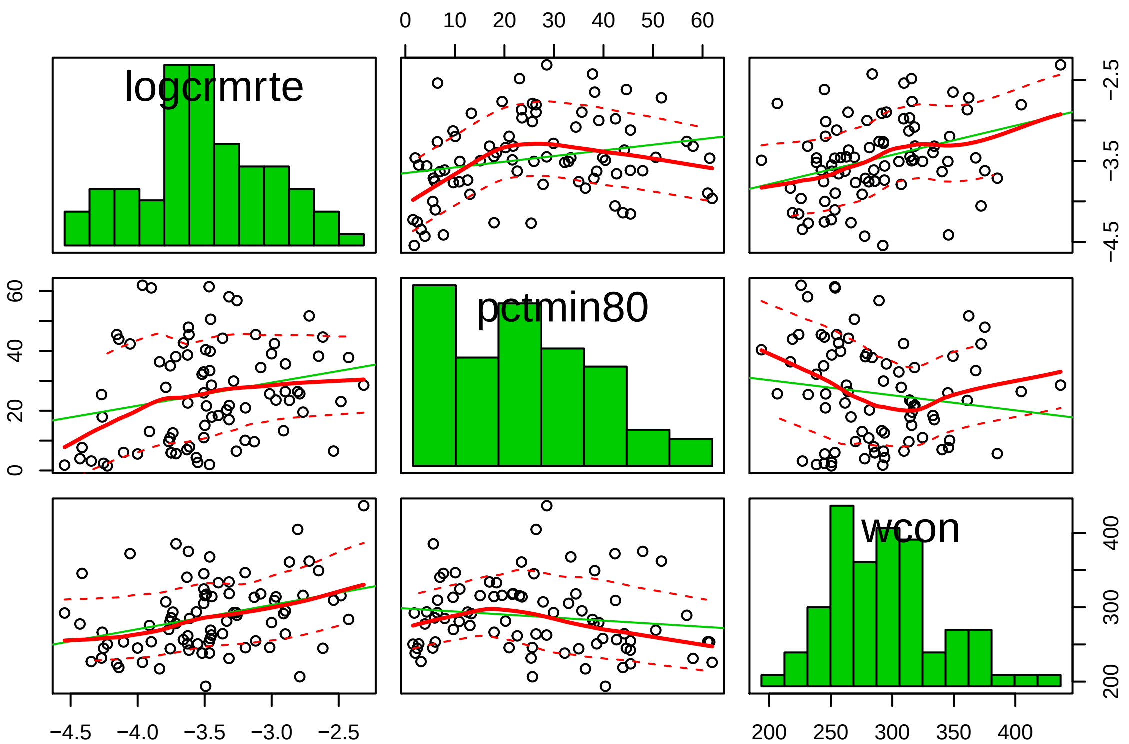

```
scatterplotMatrix(crime1[,c("logcrmrte","prbpris",  "logavgsen")], diagonal = "histogram")
```



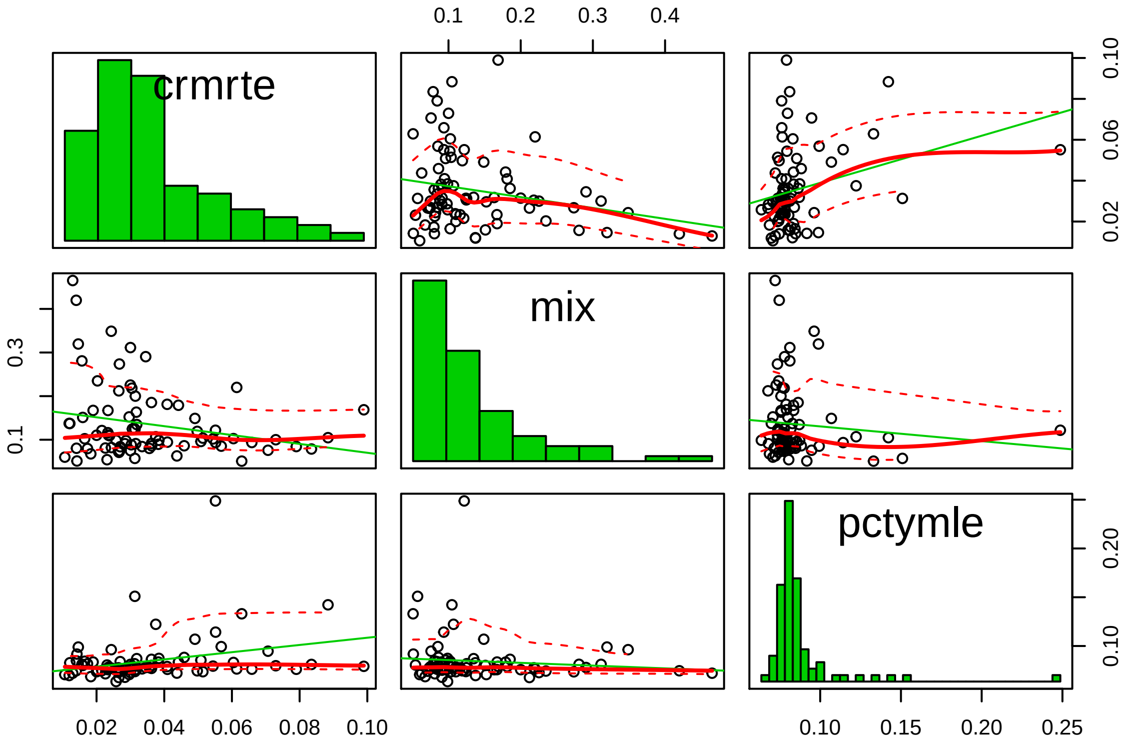

```
scatterplotMatrix(crime1[,c("logcrmrte","logpolpc",  "log1pdensity")], diagonal = "histogram")
```



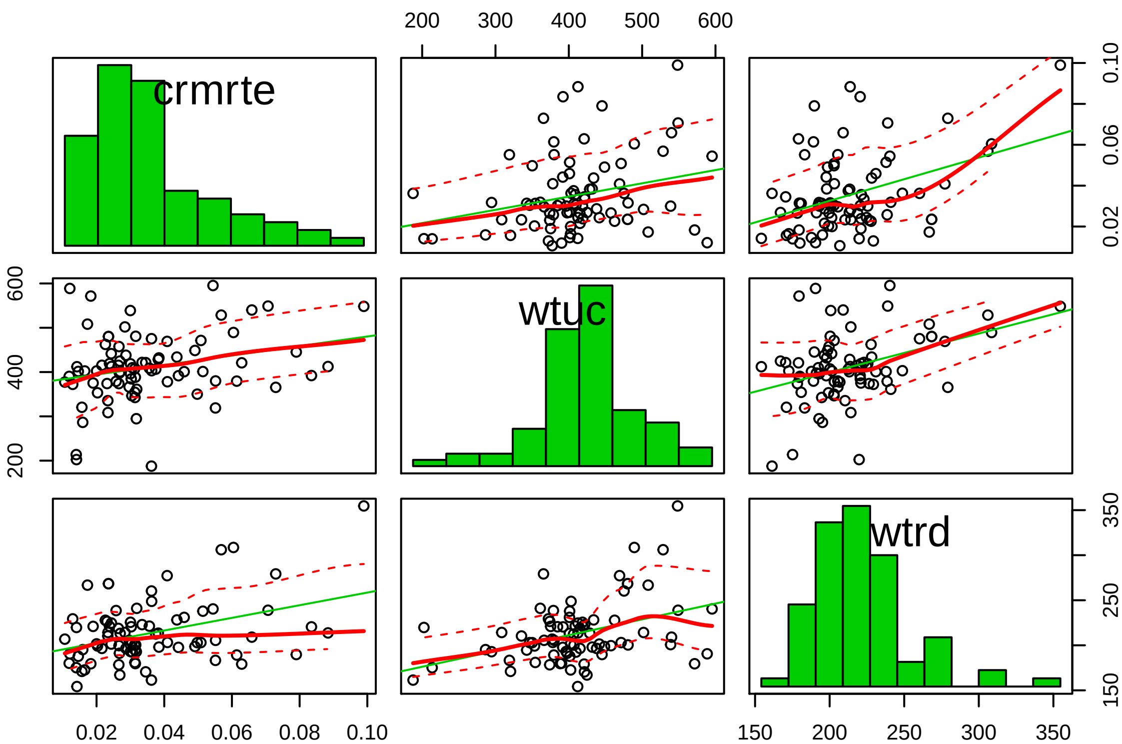

```
scatterplotMatrix(crime1[,c("logcrmrte","pctmin80",  "wcon")], diagonal = "histogram")
```



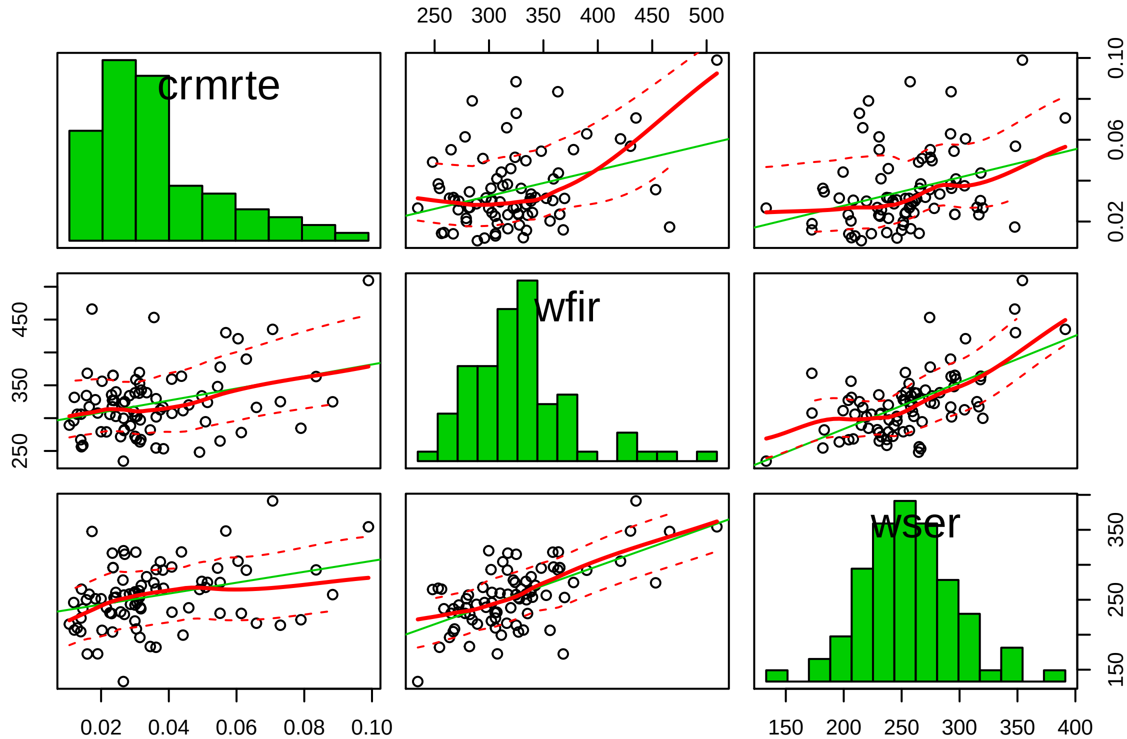

```
scatterplotMatrix(crime1[,c("crmrte","mix","pctymle")], diagonal = "histogram")
```



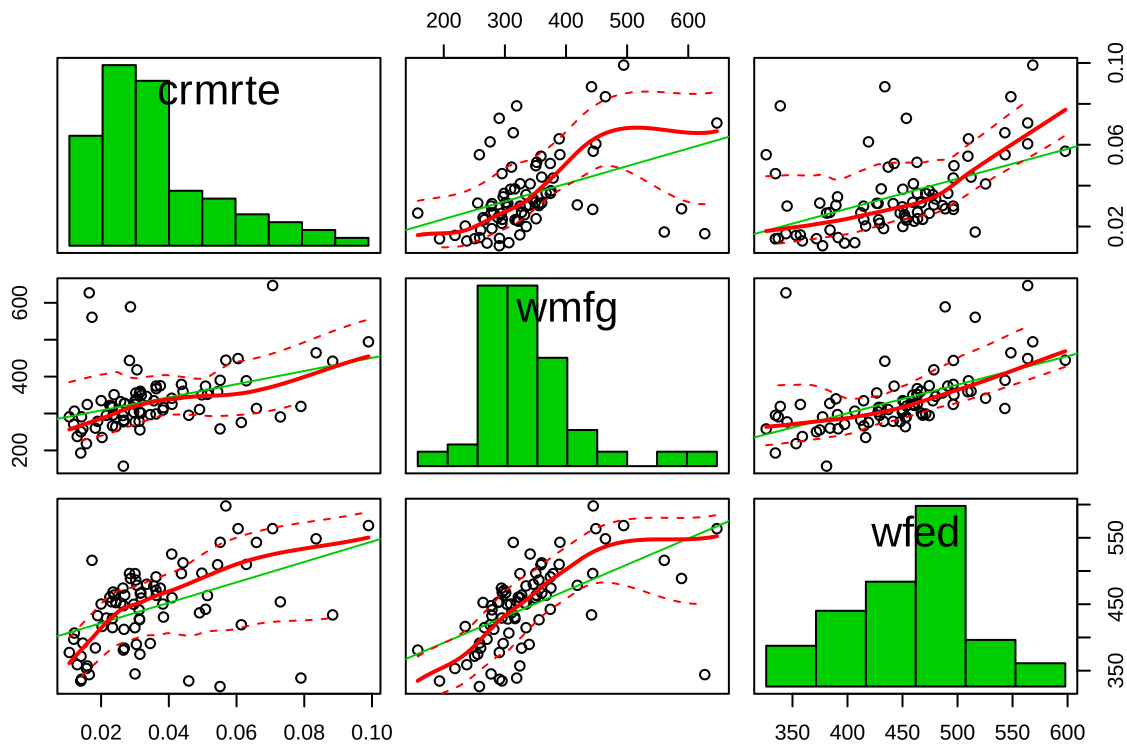

```r
scatterplotMatrix(crime1[,c("crmrte","wtuc",  "wtrd")], diagonal = "histogram")
```



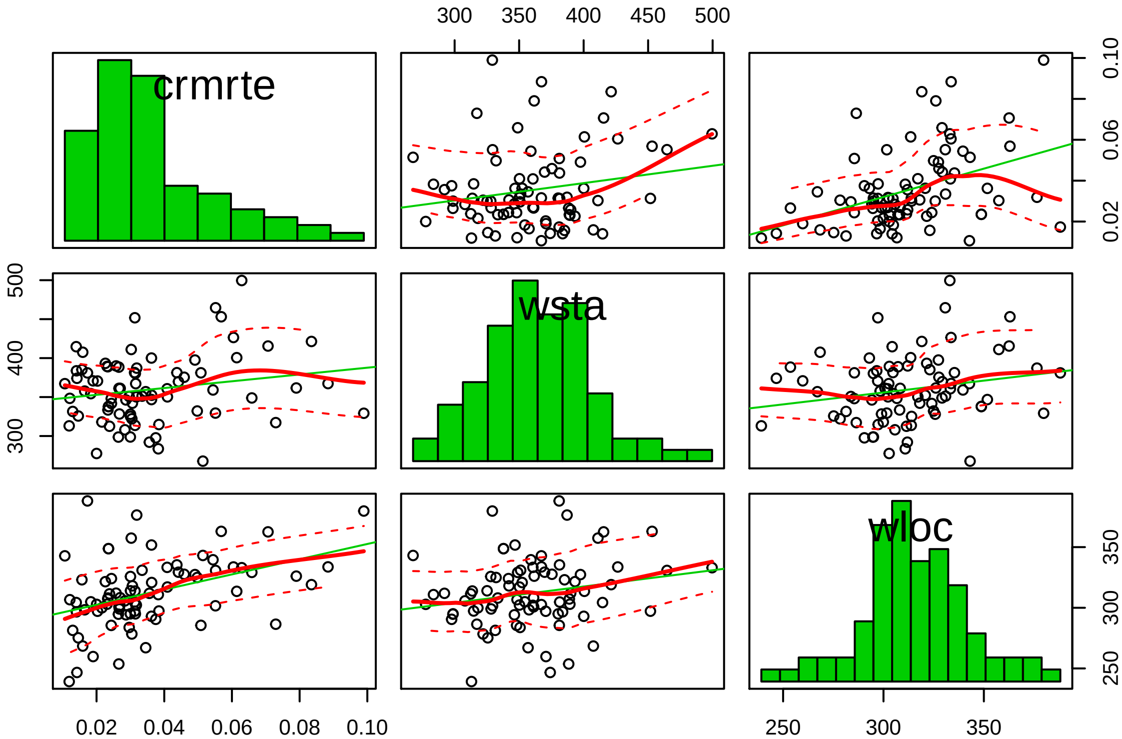

```r
scatterplotMatrix(crime1[,c("crmrte","wfir",  "wser")], diagonal = "histogram")
```



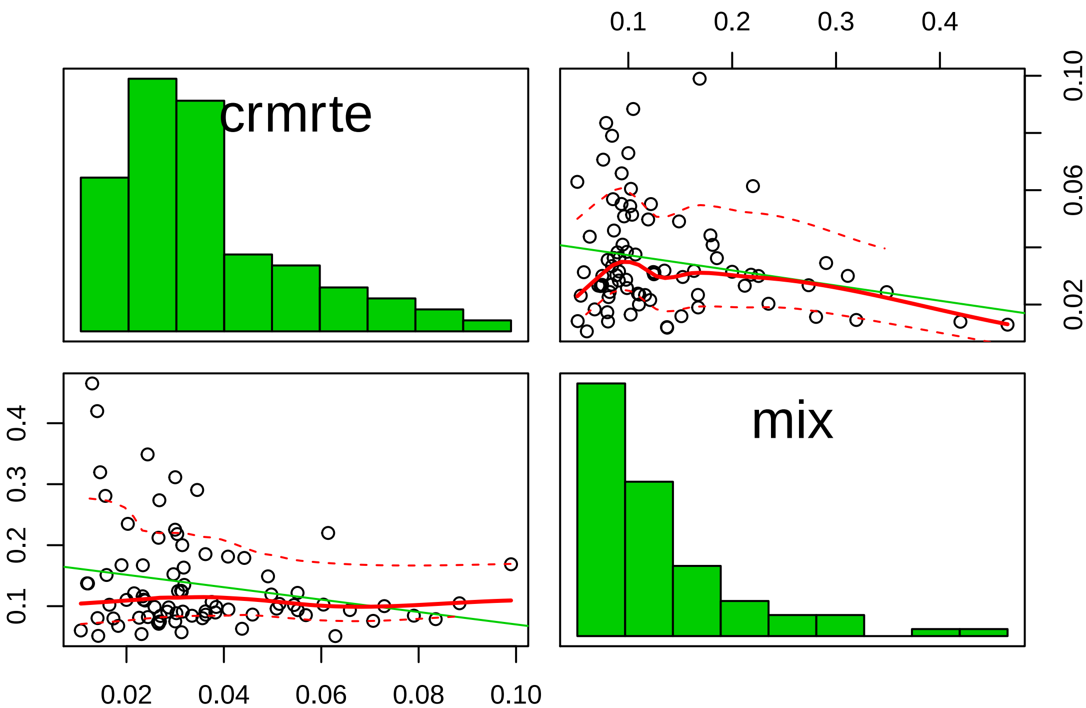

```
scatterplotMatrix(crime1[,c("crmrte","wmfg",  "wfed")], diagonal = "histogram")
```



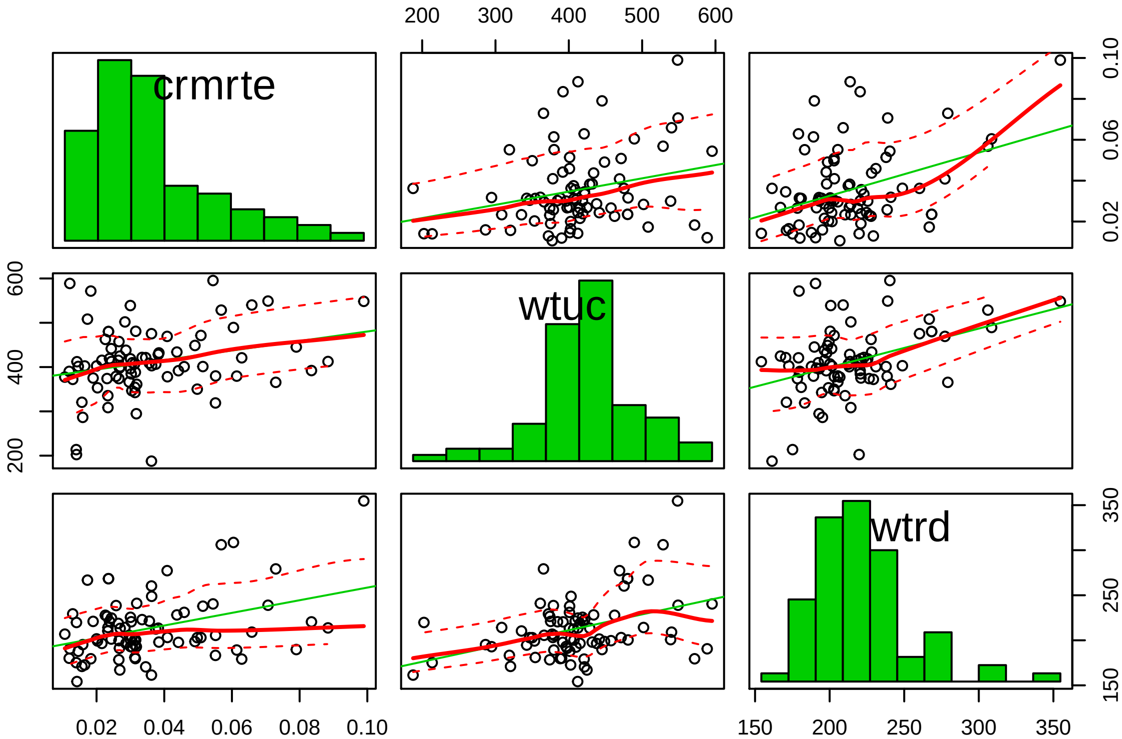

```
scatterplotMatrix(crime1[,c("crmrte","wsta",  "wloc")], diagonal = "histogram")
```



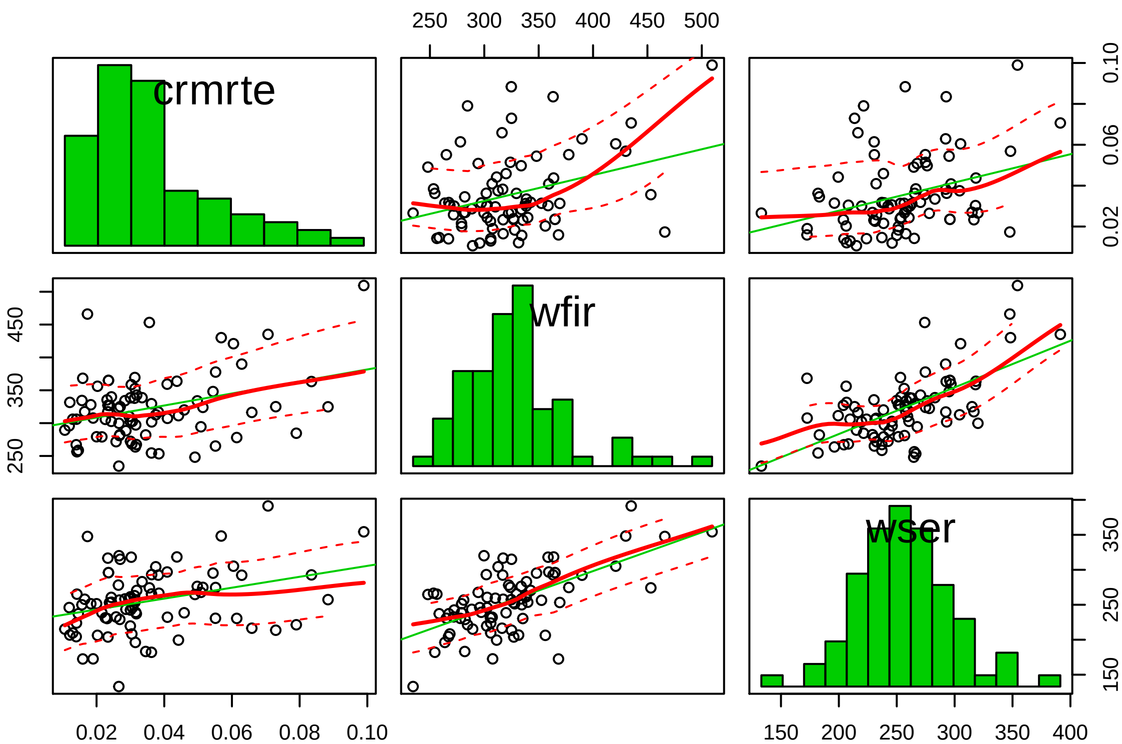

```r
scatterplotMatrix(crime1[,c("crmrte","mix")], diagonal = "histogram")
```




```r
scatterplotMatrix(crime1[,c("crmrte","wtuc",  "wtrd")], diagonal = "histogram")
```



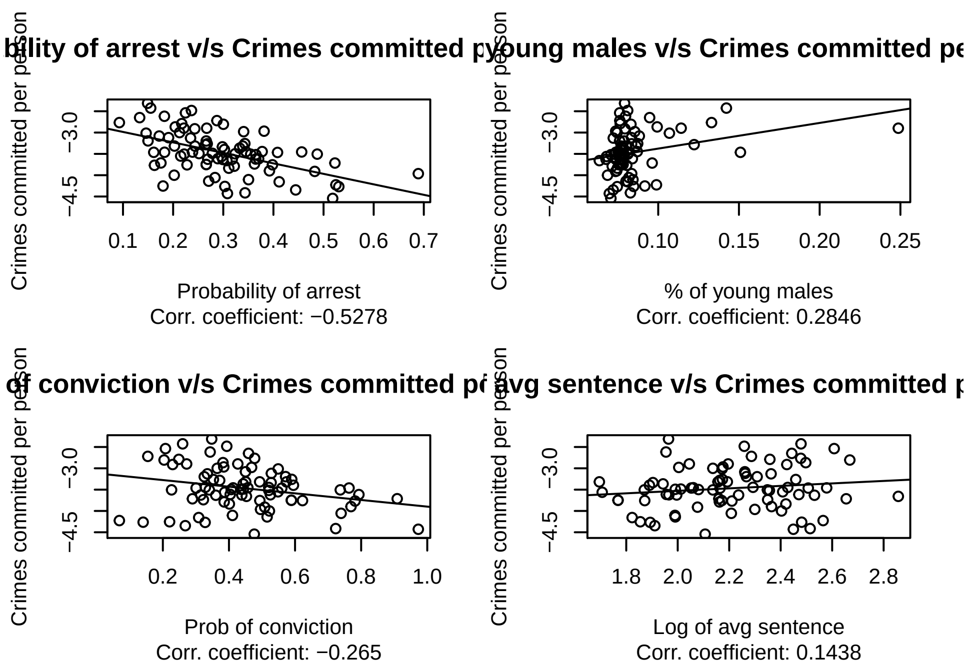

```r
scatterplotMatrix(crime1[,c("crmrte","wfir",  "wser")], diagonal = "histogram")
```



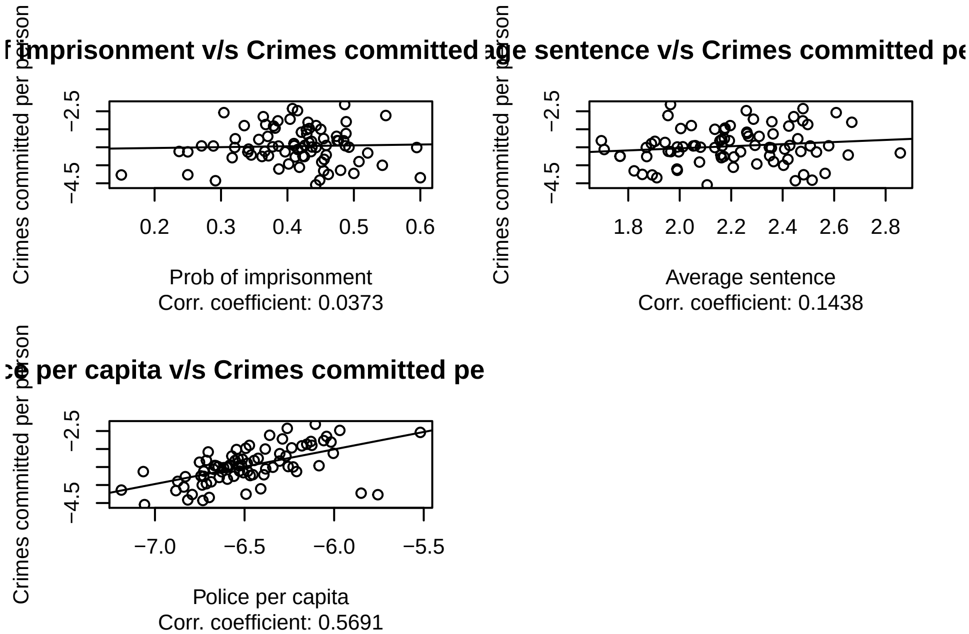

```r
# corr_val=round(cor(crime1$logcrmrte, crime1$prbarr),4)
#
# main_title=paste(in_xlabel,'v/s',in_y_label, sep = ' ')
# plot(crime1$prbarr, crime1$logcrmrte,
#       main = main_title,
```

```r
#         sub=paste("Corr. coefficient:",corr_val),
#         xlab=in_xlabel,
#         ylab=in_y_label)
# m = lm(in_field_x ~ in_field_y)
# abline(m)
```

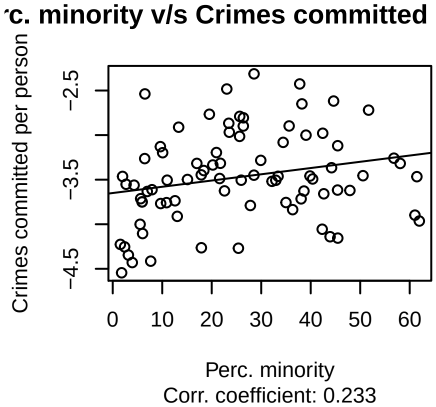
```r
par(mfrow = c(4, 2))

f_plot_three(crime1$prbarr, "Probability of arrest",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$pctymle, "% of young males",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$prbconv_cast, "Prob of conviction",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$logavgsen, "Log of avg sentence",crime1$logcrmrte,"Crimes committed per person" )
```

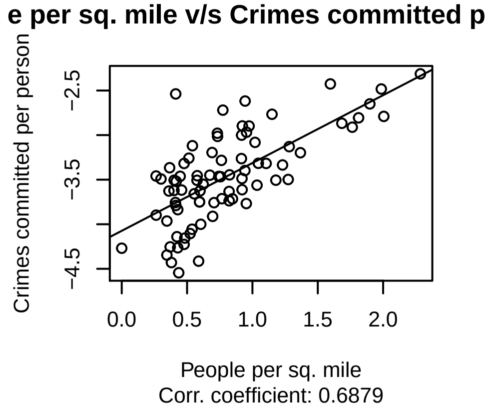


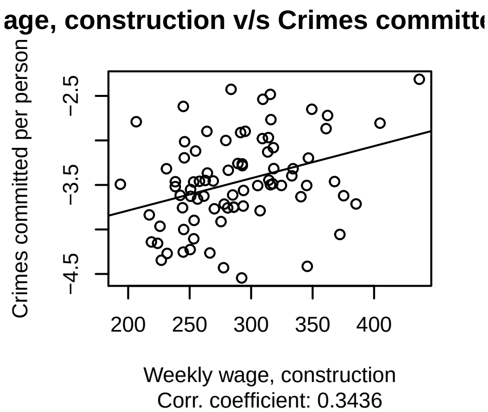

```r
par(mfrow = c(3, 2))

f_plot_three(crime1$prbpris, "Prob of imprisonment",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$logavgsen, "Average sentence",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$logpolpc, "Police per capita",crime1$logcrmrte,"Crimes committed per person" )
```



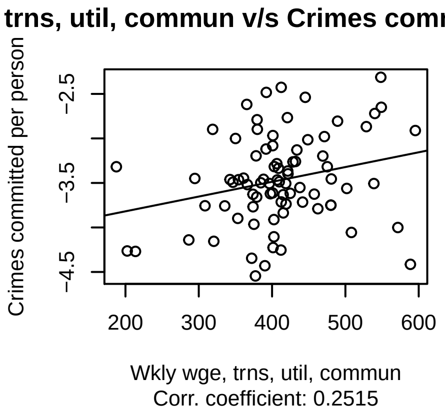

```r
par(mfrow = c(3, 2))

f_plot_three(crime1$pctmin80, "Perc. minority",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$log1pdensity, "People per sq. mile",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$wcon, "Weekly wage, construction",crime1$logcrmrte,"Crimes committed per person" )
```

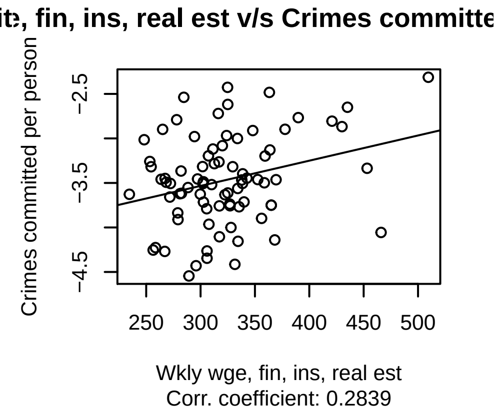


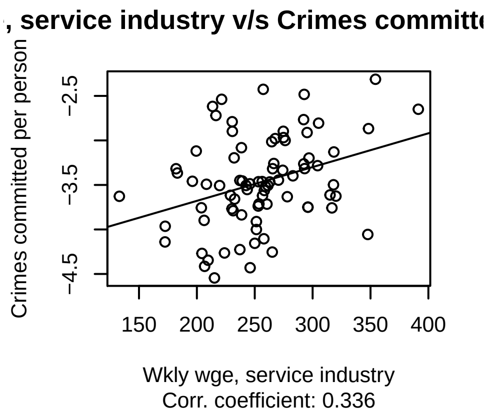




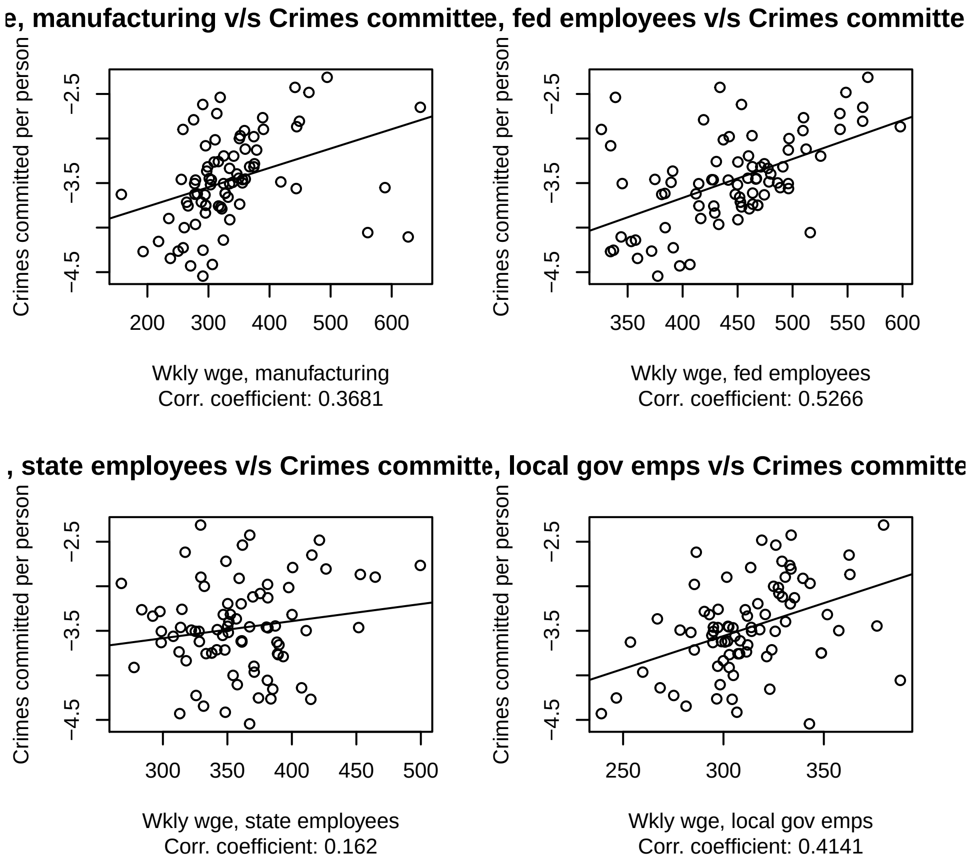

```
par(mfrow = c(3, 2))

f_plot_three(crime1$wtuc, "Wkly wge, trns, util, commun",crime1$logcrmrte,"Crimes committed per person"

f_plot_three(crime1$wfir, "Wkly wge, fin, ins, real est",crime1$logcrmrte,"Crimes committed per person"

f_plot_three(crime1$wser, "Wkly wge, service industry",crime1$logcrmrte,"Crimes committed per person" )
```








```r
par(mfrow = c(3, 2))

f_plot_three(crime1$wmfg, "Wkly wge, manufacturing",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$wfed, "Wkly wge, fed employees",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$wsta, "Wkly wge, state employees",crime1$logcrmrte,"Crimes committed per person" )

f_plot_three(crime1$wloc, "Wkly wge, local gov emps",crime1$logcrmrte,"Crimes committed per person" )
```



Strong positive correlation:

crmrte v/s polpc crmrte v/s density crmrte v/s wcon crmrte v/s wser crmrte v/s wmfg crmrte v/s wfed crmrte v/s wloc

Weak positive correlation:

crmrte v/s pctymle crmrte v/s percent of minority crmrte v/s wkly wge, fin, ins, real est crmrte v/s wtuc

Strong negative correlation:

crmrte v/s prbarr

Weak negative correlation:

crmrte v/s prbconv

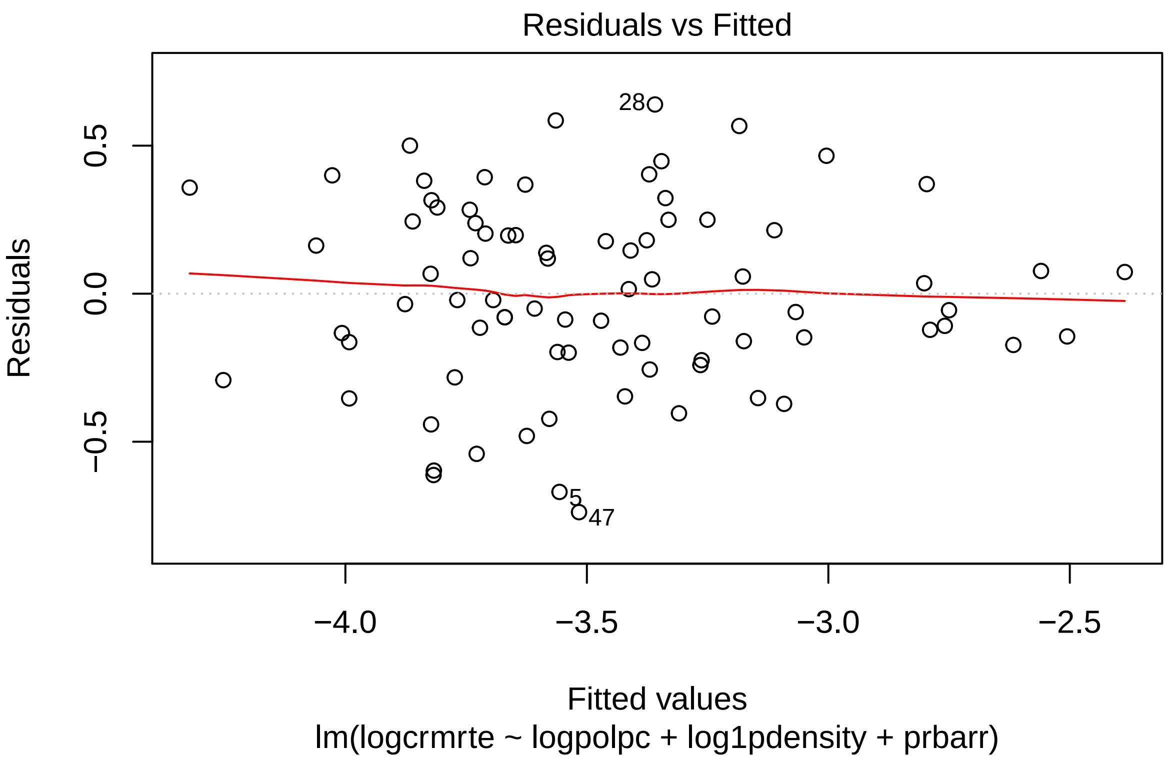
Model 1: Crime related variables

log(crime rate) = log(beta0 + beta1*polpc* + *beta2*density)

```
logcrmrte.lm1 = lm( logcrmrte ~ logpolpc + log1pdensity + prbarr , data=crime1)
logcrmrte.lm1
```

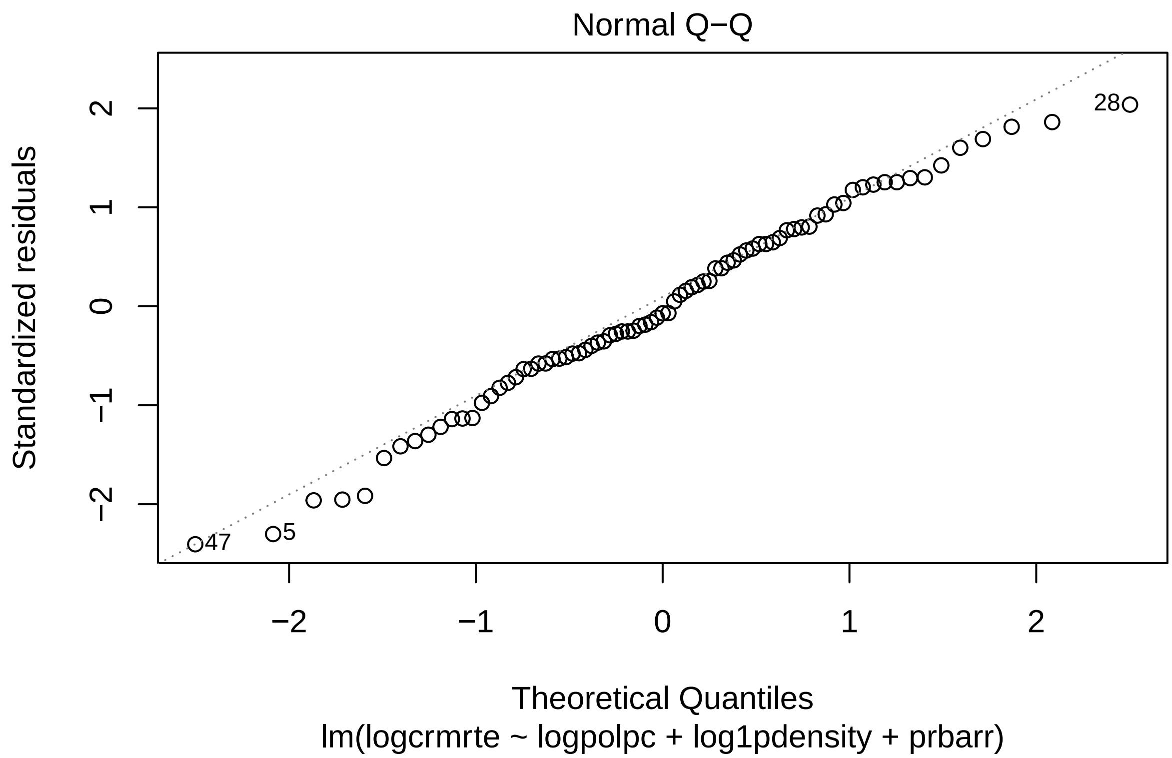
```
##
## Call:
## lm(formula = logcrmrte ~ logpolpc + log1pdensity + prbarr, data = crime1)
##
## Coefficients:
##  (Intercept)      logpolpc  log1pdensity        prbarr
##       0.2349        0.5674        0.4545       -1.3148
```

```r
plot(logcrmrte.lm1, which =1)
```



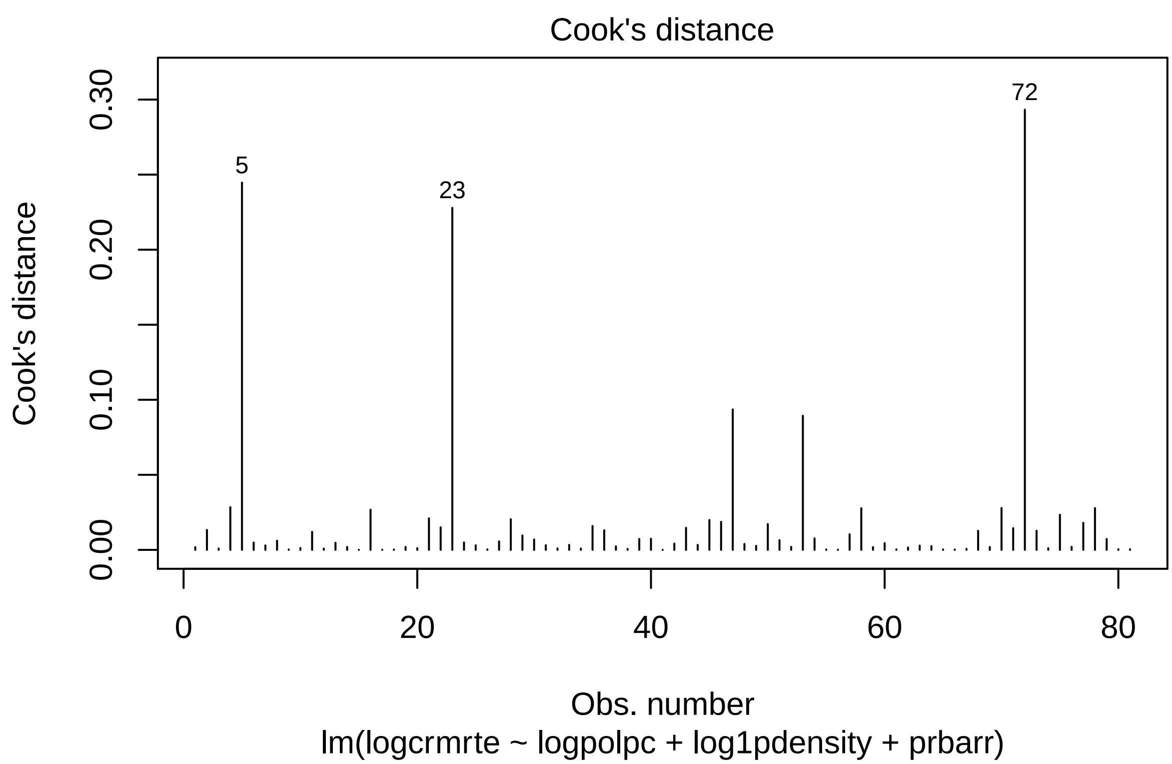
The regression lines is close to 0 residual. This indicates there is a linear relationship among logcrmrte with logpolpc + log1pdensity.

```r
plot(logcrmrte.lm1, which =2)
```



The Normal QQ plot is roughly on a straight line. This indicates that our data has been sourced from a normal distribution

```r
plot(logcrmrte.lm1, which =4)
```



None of the points have Cook's distance greater than 1. However, Points 5, 23 and 72 have higher leverage than all other points. Let us observe the points:

```r
crime1[5,]
```

```
##   county year    crmrte   prbarr     prbconv prbpris avgsen      polpc
## 5     11   87 0.0146067 0.524664 0.068376102     0.5     13 0.00288203
##     density    taxpc west central urban pctmin80     wcon     wtuc
## 5 0.6113361 35.22974    1       0     0   1.5407 250.4006 401.3378
##       wtrd    wfir     wser  wmfg   wfed   wsta   wloc       mix   pctymle
## 5 187.8255 258.565 237.1507 258.6 391.48 325.71 275.22 0.3195266 0.0989192
##   prbconv_cast regionofcrime logcrmrte logavgsen logpolpc log1pdensity
## 5    0.0683761          WEST -4.226275  2.564949 -5.84926    0.4770637
```

```r
crime1[23,]
```

```
##    county year    crmrte   prbarr     prbconv  prbpris avgsen      polpc
## 23     55   87 0.0790163 0.224628 0.207830995 0.304348  13.57 0.00400962
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 23 0.5115089 119.7615    0       0     0  6.49622 309.5238 445.2762
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc        mix
## 23 189.7436 284.5933 221.3903 319.21 338.91 361.68 326.08 0.08437271
##       pctymle prbconv_cast regionofcrime logcrmrte logavgsen  logpolpc
## 23 0.07613807     0.207831       UNKNOWN -2.538101  2.607861 -5.519059
##    log1pdensity
## 23    0.4131085
```

```r
crime1[72,]
```

```
##    county year    crmrte   prbarr     prbconv prbpris avgsen      polpc
## 72    173   87 0.0139937 0.530435 0.327868998    0.15   6.64 0.00316379
##        density    taxpc west central urban pctmin80    wcon     wtuc
## 72 2.03422e-05 37.72702    1       0     0  25.3914 231.696 213.6752
##        wtrd    wfir     wser   wmfg   wfed   wsta   wloc       mix
## 72 175.1604 267.094 204.3792 193.01 334.44 414.68 304.32 0.4197531
##       pctymle prbconv_cast regionofcrime logcrmrte logavgsen  logpolpc
## 72 0.07462687     0.327869          WEST -4.269148  1.893112 -5.755985
##    log1pdensity
## 72 2.034199e-05
```

Checking for multicollinearity

```r
## Reference: https://www.r-bloggers.com/collinearity-and-stepwise-vif-selection/
## https://datascienceplus.com/multicollinearity-in-r/
vif(logcrmrte.lm1)
```

```
##     logpolpc log1pdensity       prbarr
##     1.249780     1.518756     1.256574
```

Since vif output for both th variable sis less than 10, there does not exist a multicollinearity between the variables.

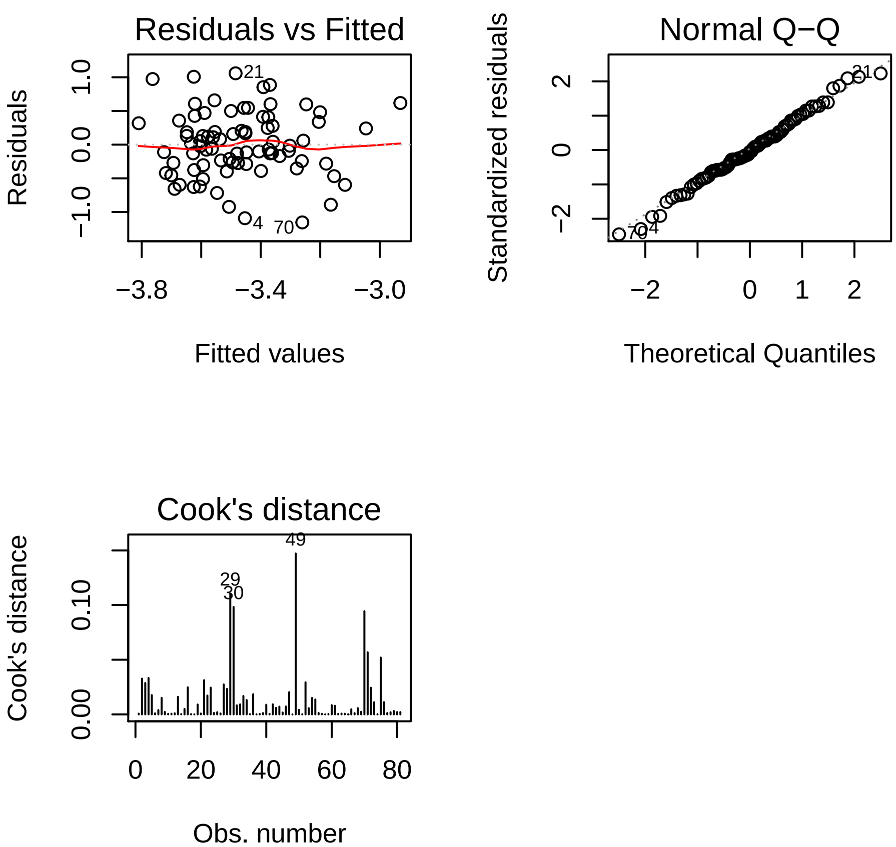
Model 2: Wage variable wcon

log(crime rate) = log(beta0 + beta1*wcon + error)

```r
logcrmrte.lm2 = lm( logcrmrte ~ wcon, data=crime1)
logcrmrte.lm2
```

```
##
## Call:
## lm(formula = logcrmrte ~ wcon, data = crime1)
##
## Coefficients:
## (Intercept)         wcon
##   -4.510458     0.003617
```

```r
par(mfrow =c(2,2))
plot(logcrmrte.lm2,which=c(1,2,4))
par(mfrow =c(1,1))
```



The regression lines is close to 0 residual. This indicates there is a linear relationship among logcrmrte with wcon. The Normal QQ plot is roughly on a straight line. This indicates that our data has been sourced from a normal distribution None of the points have Cook's distance greater than 1.

Model 2: Wage variable wmfg

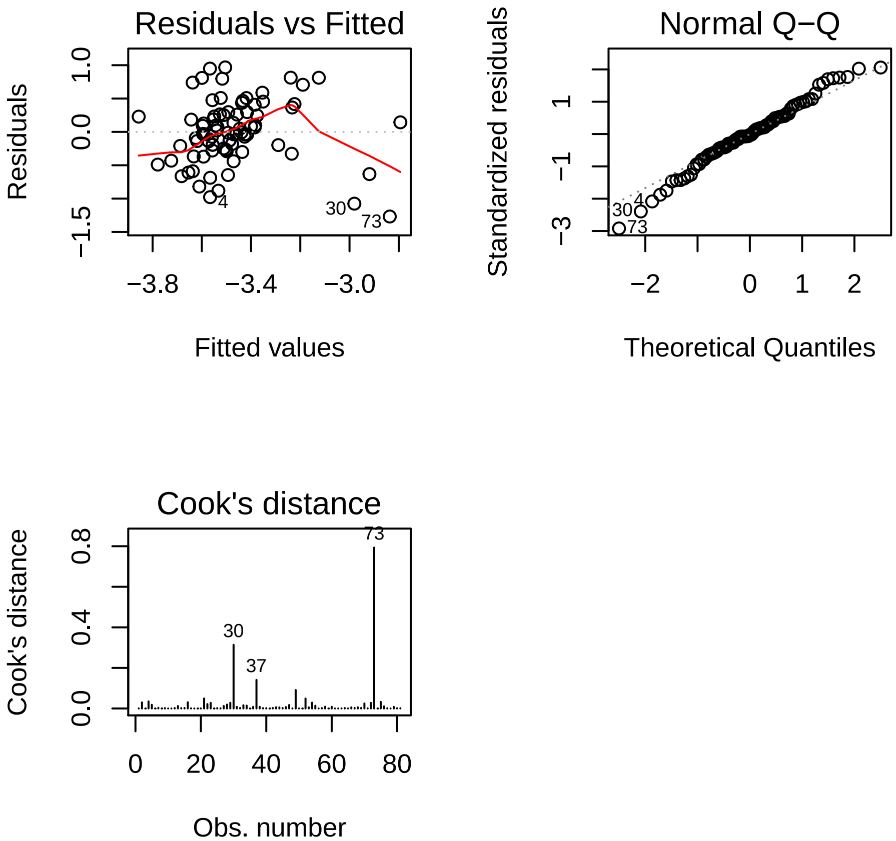
log(crime rate) = log(beta0 + beta1*wmfg + error)

```r
logcrmrte.lm3 = lm( logcrmrte ~ wmfg, data=crime1)
logcrmrte.lm3
```

```
##
## Call:
## lm(formula = logcrmrte ~ wmfg, data = crime1)
##
## Coefficients:
```

```
## (Intercept)         wmfg
##   -4.198123     0.002172
```

```
par(mfrow =c(2,2))
plot(logcrmrte.lm3,which=c(1,2,4))
par(mfrow =c(1,1))
```



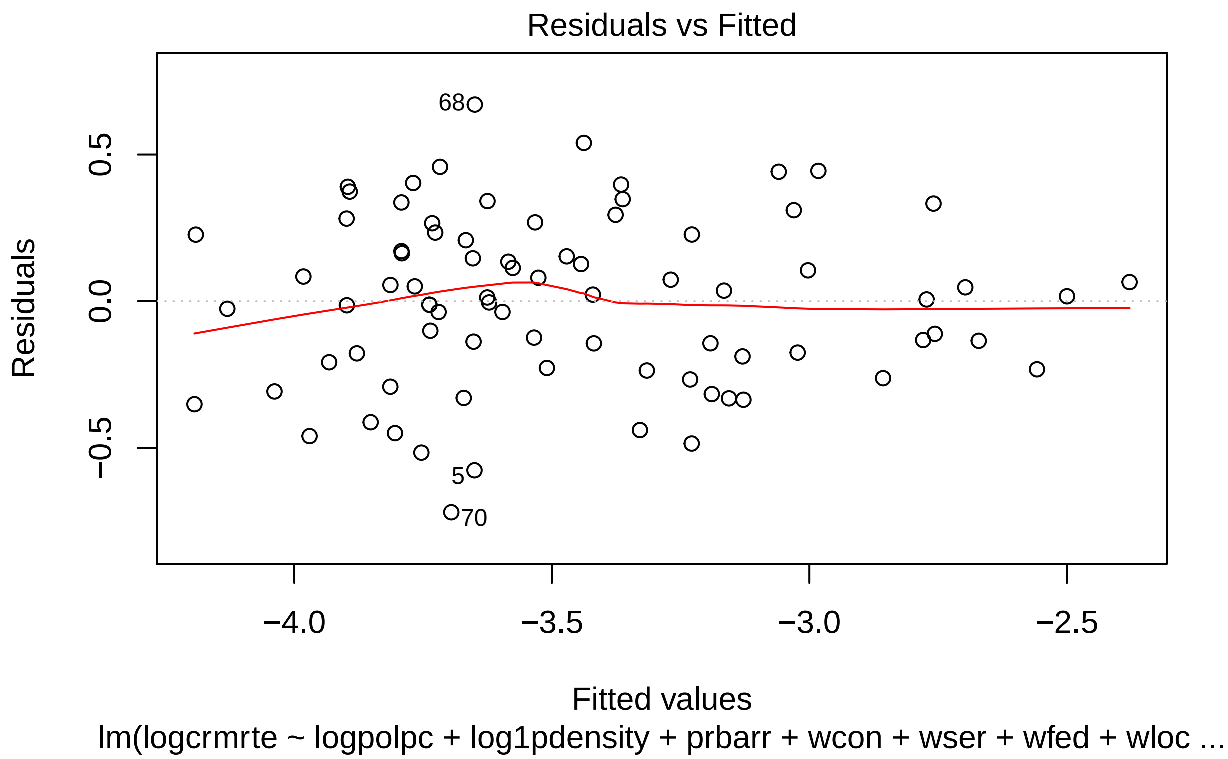
The regression lines is varying from residual 0. This indicates there is no linear relationship between logcrmrte and wmfg. The Normal QQ plot is roughly on a straight line. This indicates that our data has been sourced from a normal distribution None of the points have Cook's distance greater than 1.

Model 3: All variables with strong positive or negative correlation

```
logcrmrte.lm3 = lm( logcrmrte ~ logpolpc + log1pdensity +prbarr + wcon + wser  + wfed + wloc, data=crim
logcrmrte.lm3
```

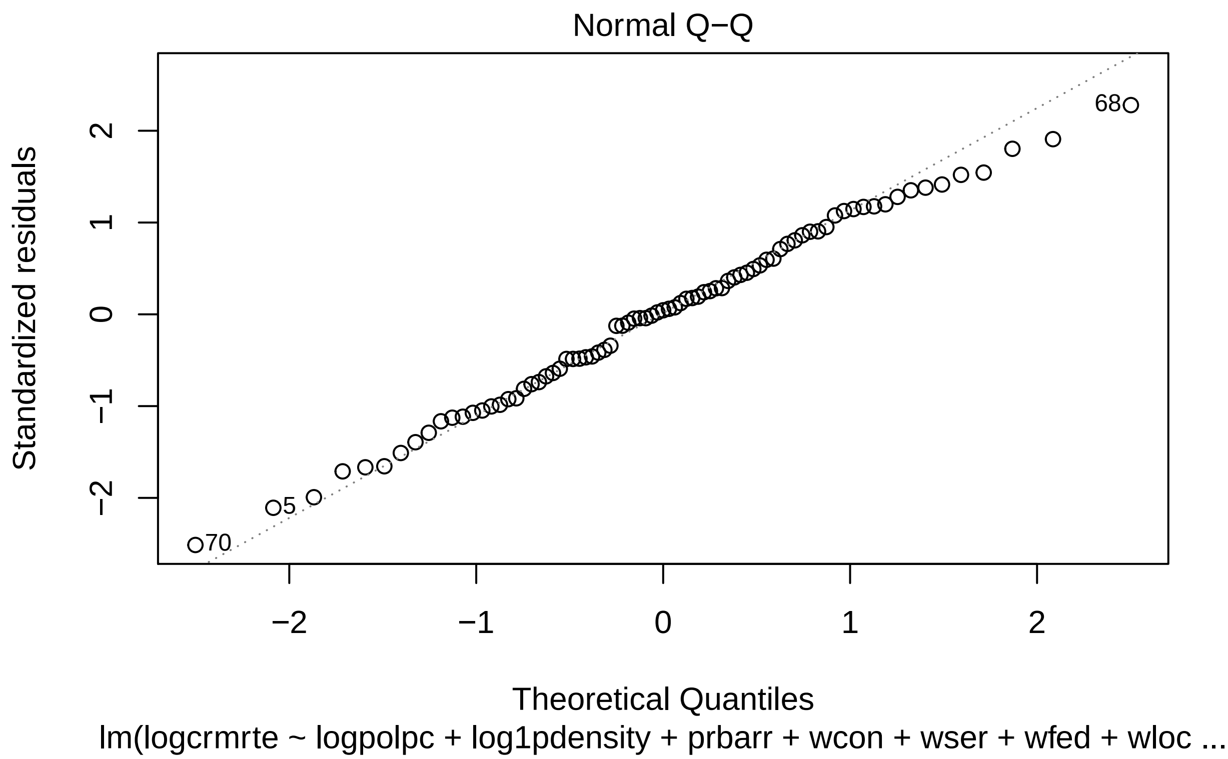
```
##
## Call:
## lm(formula = logcrmrte ~ logpolpc + log1pdensity + prbarr + wcon +
##     wser + wfed + wloc, data = crime1)
##
## Coefficients:
##  (Intercept)      logpolpc  log1pdensity        prbarr          wcon
##    0.1163312     0.5837347     0.4703757    -1.4184433     0.0003637
##         wser          wfed          wloc
##   -0.0036034     0.0013687     0.0014368
```

```r
plot(logcrmrte.lm3, which =1)
```



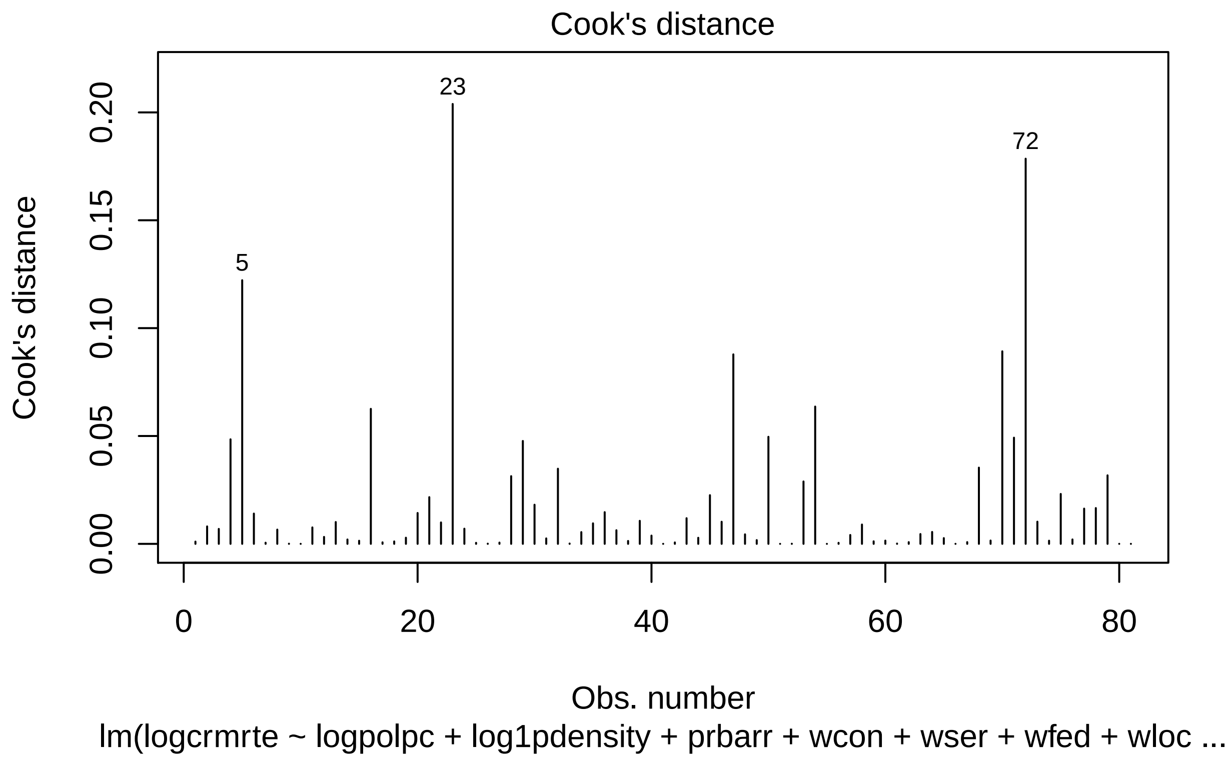
The regression lines is close to 0 residual. This indicates there is a linear relationship between logcrmrte and (logpolpc + log1pdensity +prbarr + wcon + wser + wfed + wloc)

```r
plot(logcrmrte.lm3, which =2)
```



The Normal QQ plot is roughly on a straight line. This indicates that our data has been sourced from a normal distribution

```r
plot(logcrmrte.lm3, which =4)
```



None of the points have Cook's distance greater than 1. However, Points 5, 23 and 72 have higher leverage than all other points.

```r
vif(logcrmrte.lm3)
```

```
##     logpolpc log1pdensity       prbarr         wcon         wser 
##     1.264598     2.420224     1.277392     1.769968     2.189740 
##         wfed         wloc 
##     2.108466     1.989238
```

There is no multicollinearity since none of the vif values are more than 10.

Comparing the models

```r
summary(logcrmrte.lm1)$r.squared
```

```
## [1] 0.6223734
```

```r
summary(logcrmrte.lm2)$r.squared
```

```
## [1] 0.1180323
```

```r
summary(logcrmrte.lm3)$r.squared
```

```
## [1] 0.6744848
```

```r
## The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models
AIC(logcrmrte.lm1, logcrmrte.lm2, logcrmrte.lm3)
```

```
##               df       AIC
```

```
## logcrmrte.lm1  5  49.47324
## logcrmrte.lm2  3 114.18145
## logcrmrte.lm3  9  45.44501
```

Based on the AIC output logcrmrte.lm1 or logcrmrte.lm3 is preferred.

```r
stargazer(logcrmrte.lm1, logcrmrte.lm2,logcrmrte.lm3, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting Crime rate per persone",
          keep.stat = c("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: Wed, Mar 28, 2018 - 19:01:14

Table 1: Linear Models Predicting Crime rate per persone

	Dependent variable:		
	logcrmrte		
	(1)	(2)	(3)
logpolpc	0.567		0.584
log1pdensity	0.455		0.470
prbarr	−1.315		−1.418
wcon		0.004	0.0004
wser			−0.004
wfed			0.001
wloc			0.001
Constant	0.235	−4.510	0.116
Observations	81	81	81
R^2	0.622	0.118	0.674