W203 Lab 3

Armand Kok, Adam Yang, James De La Torre

Introduction

Our team has been hired by a local political campaign to research North Carolina crime statistics and generate suggestions for policies for reducing crime.

The crime statistics data set being analyzed is a subset of the data used by Cornwell and W. Trumball in their 1994 study. The data set contains the output variable, `crmrate`, which is crimes committed per capita, and it also contains 24 other variables which will be treated as input variables and potential modulators of the crime rate.

We will attempt to build a linear model that regresses `crmrate` on some key variables in the data set. We hope to identify variables that can be reasonably assessed as causal with respect to the crime rate. From our model fidings, we will produce policy proposals which we believe can influence these variables and result in a decrease in crime.

It is important to note that just because a variable is found to correlate with the crime rate, it does not imply that the variable is useful from the perspective of a political campaign. We may find variables that cannot be influenced by any political policy or action. Such variables may improve the predictive ability of our model, but they will not be targeted for change by any policy proposals.

Exploratory Data Analysis

The data file, `crime_v2.csv` was opened and found to contain 97 rows. Each row represents data for a single county in North Carlolina. Immediate inspection of the data revealed a few data cleanup steps were required.

- The last 6 rows of the data set were blanks. These empty records were deleted.
- One row had values of 1 for both `west` and `central`, placing that county in two regions simultaneously. It is unknown whether this is possible, but currently there has been no reason to delete this particular row so the data will be kept for now, as evaluation of variable importance is still ongoing.
- The `prbconv` variable, representing the "probability of conviction" was read in as a factor (a cateogorical variable) instead of a numeric variable. This variable was converted to numeric.

```
library(car)
library(reshape2)
library(ggplot2)

# Adam's dir
mydir <- "/Users/adamyang/Desktop/w203/Lab3/w203-Lab3/"

# Armand's dir
# mydir<-'C:/Users/ak021523/Documents/GitHub/mids-repos/W203/Homework/w203-Lab3/'

# jim's directory mydir<-
# F:/users/jddel/Documents/DATA_SCIENCE_DEGREE_LAPTOP/W203_Stats/Lab_03/'

# read df
crime_df = read.csv(paste0(mydir, "crime_v2.csv"))
```

```r
# summarize all vars
summary(crime_df)
```

```
##      county           year         crmrte              prbarr
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6
##         prbconv      prbpris           avgsen           polpc
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6
##     density            taxpc             west           central
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6
##      urban            pctmin80          wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd             wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed             wsta            wloc             mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
```

```
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

```
str(crime_df)
```

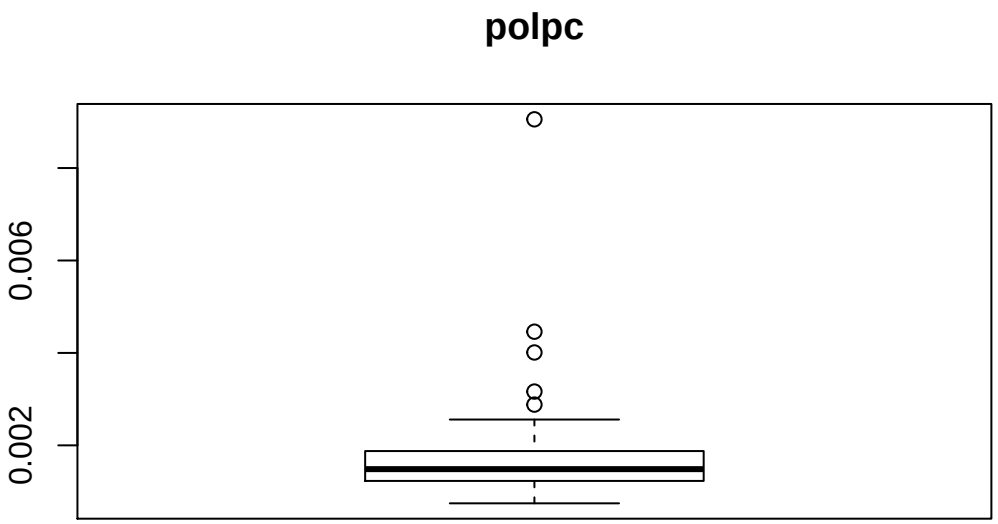
```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

```
# get rid of rows with missing values (this only kills the 6
# blank rows)
crime_df <- crime_df[complete.cases(crime_df), ]

# convert prob of conviction to numeric
crime_df$prbconv <- as.numeric(as.character(crime_df$prbconv))
```

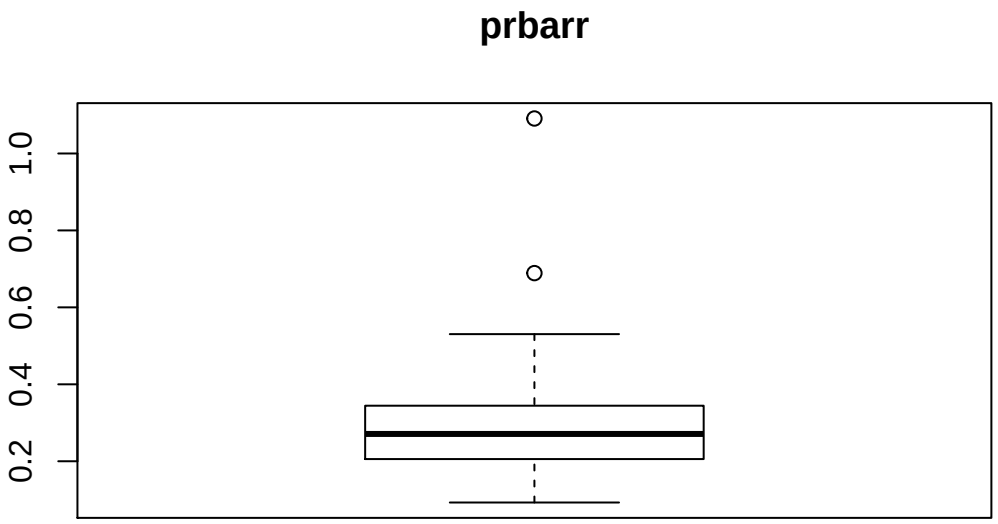
Outlier Identification

After reviewing the distributions of the different variables, there were 4 variables had outliers, which is defined by anything that is more than Q3 + 1.5 IQR or Q1 - 1.5 IQR: - polpc - row 51 - prbarr - row 51 - wser - row 84 - taxpc row 25 After reviewing further, there was no reason for the extreme outliers to be removed from the data set. boxplots of the variables above are shown below.

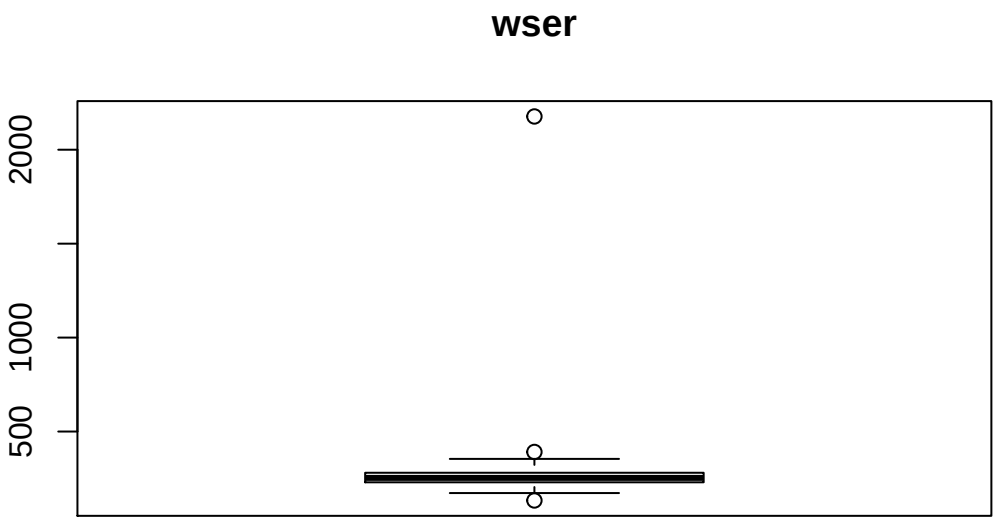
```r
boxplot(crime_df$polpc, main = "polpc")
```



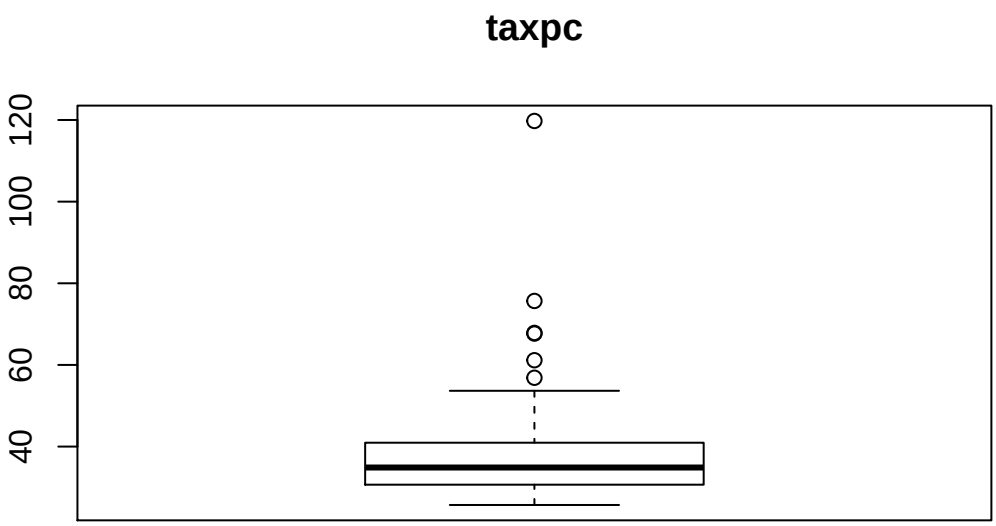

```r
boxplot(crime_df$prbarr, main = "prbarr")
```




```r
boxplot(crime_df$wser, main = "wser")
```




```
boxplot(crime_df$taxpc, main = "taxpc")
```



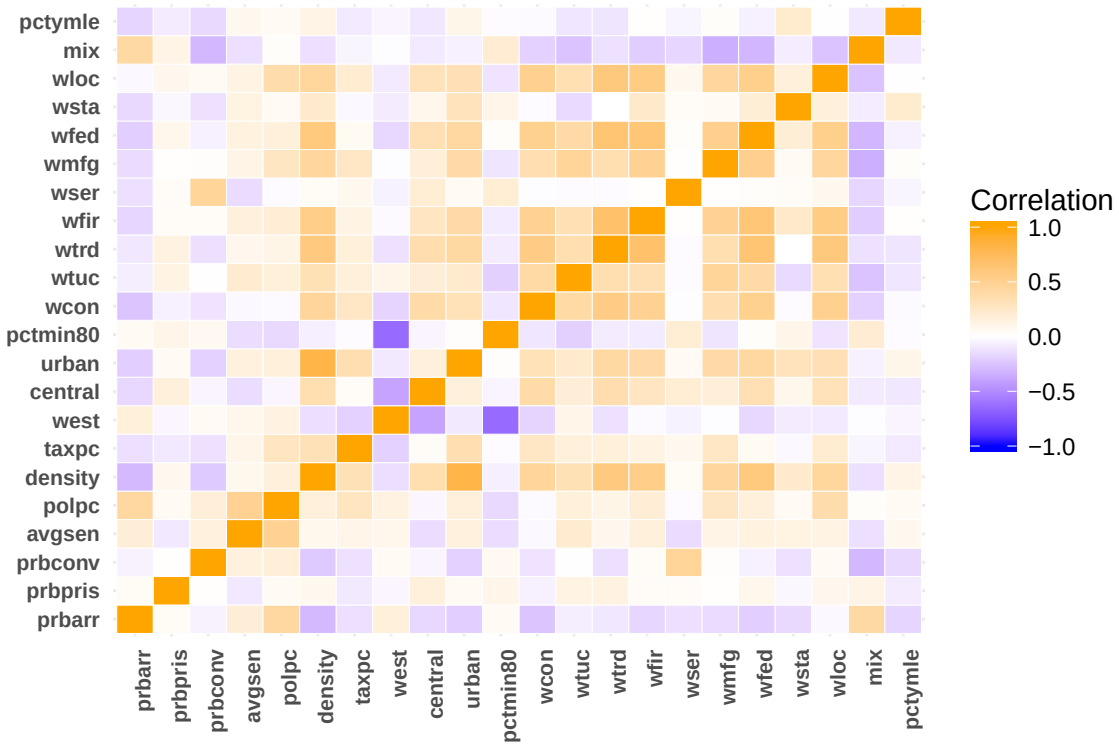

```
# 1.5 IQR from the Q3 = outlier but we can decide which to
# eliminate
```

Check for multicolinearity

Build a correlation matrix. Identify input variables that correlate with one another. Choose only one variable from each correlated pair to include in model-building.

```
# TODO - fix matrix sizing

# correlation matrix for top 4 correlation and bottom 4
# correlation
cor_dr = cor(crime_df[c("prbarr", "prbpris", "prbconv", "avgsen",
    "polpc", "density", "taxpc", "west", "central", "urban",
    "pctmin80", "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg",
    "wfed", "wsta", "wloc", "mix", "pctymle")], use = "complete.obs")

# Heatmap
ggplot(data = melt(cor_dr, na.rm = TRUE), aes(Var2, Var1, fill = value)) +
```

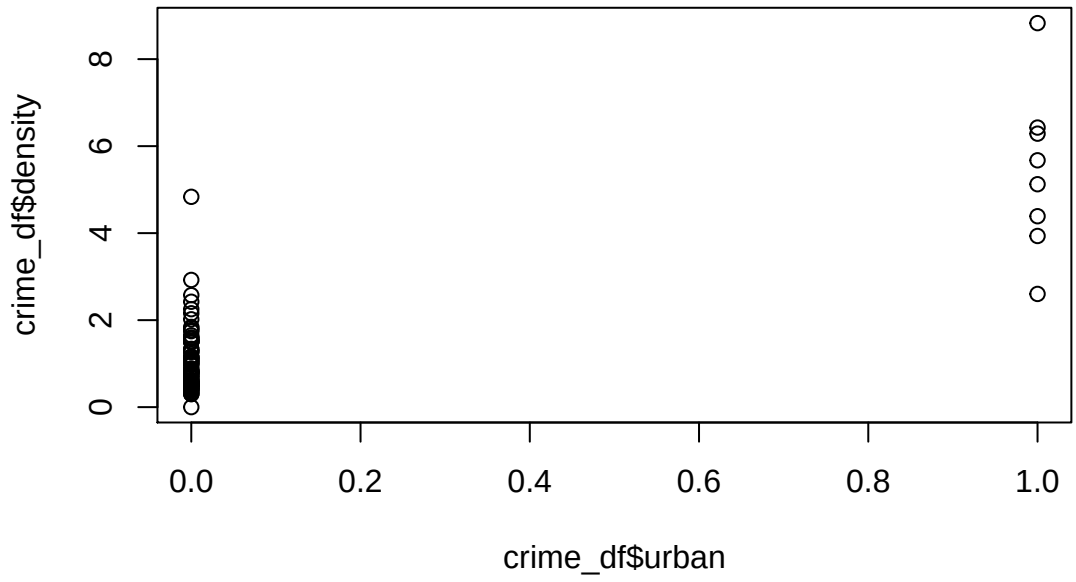
```
    theme_minimal() + geom_tile(color = "white") + scale_fill_gradient2(low = "blue",
    high = "orange", mid = "white", midpoint = 0, limit = c(-1,
        1), name = "Correlation") + theme(axis.text.x = element_text(face = "bold",
    angle = 90, vjust = 1, size = 8, hjust = 1), axis.text.y = element_text(face = "bold",
    size = 8), axis.title.x = element_blank(), axis.title.y = element_blank())
```



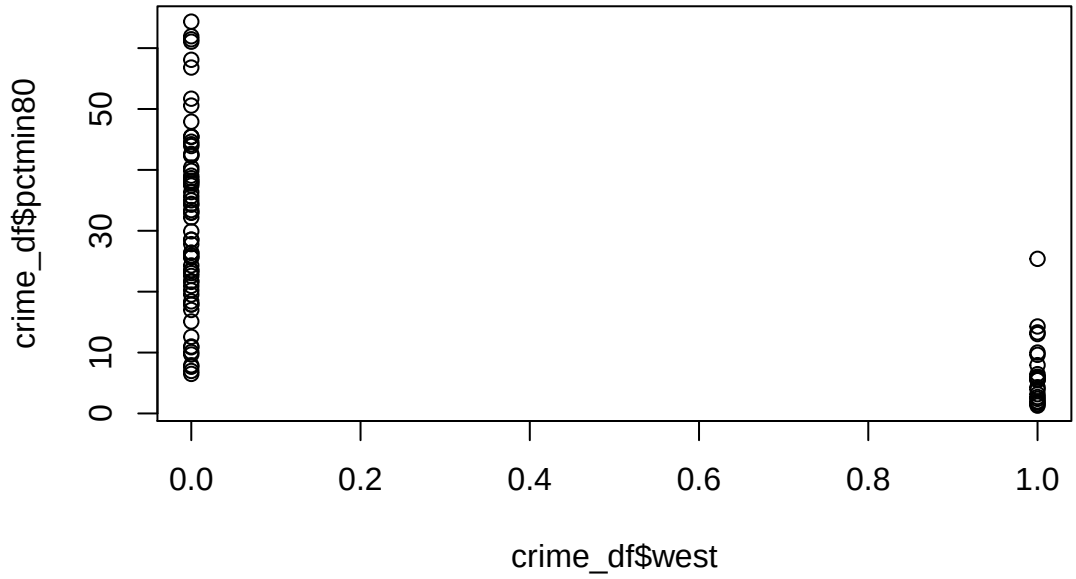
One of the assumptions for multiple OLS regression is to avoid perfect multicollinearity between independent variables. This, however, is not common in practical cases. Less than perfect multicollinearity is a more common problem that will not cause bias in the OLS, but would introduce large variances and covariances. As a result, precise estimation would become difficult so it can be beneficial to remove certain imperfect multicollinearity variables.

After reviewing the correlation matrix in detail, there were 5 pairs of variables that have a somewhat strong correlation to each other (i.e. has correlation > 0.6), which are plotted below. Based on the plots, then the following variables were removed from the final model: - urban - this is somewhat redundant with density. - west - west was removed because it is a dummy variable, and pctmin80 is a continuous one which may contain more information for the regression model. - wtrd, wfed, wfir - wages tend to be higher with density, so density was kept as it can succinctly represent the same information. Below are the scatterplots of the different correlated variables

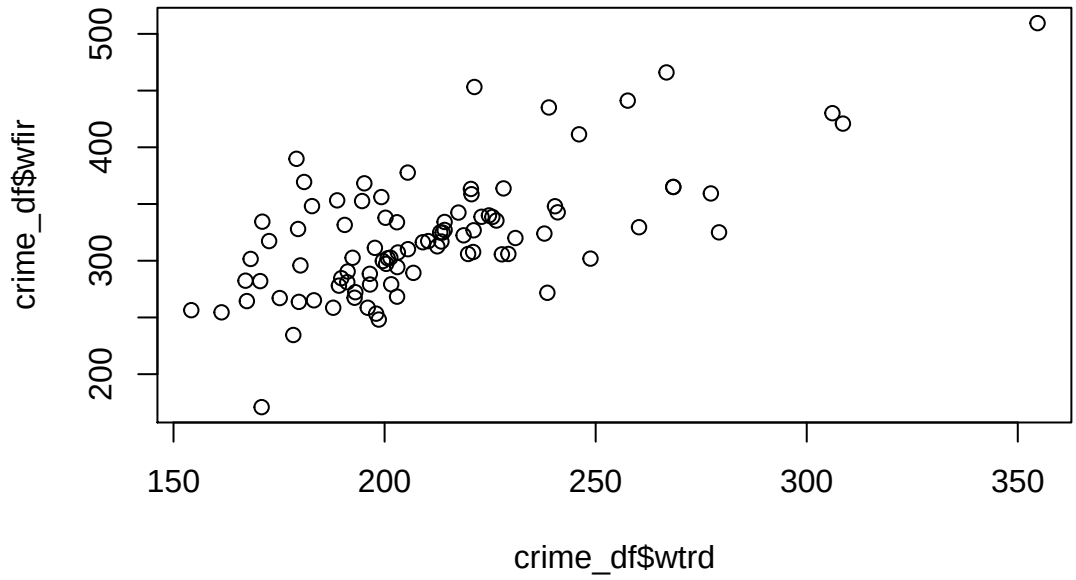
```
plot(crime_df$urban, crime_df$density)
```



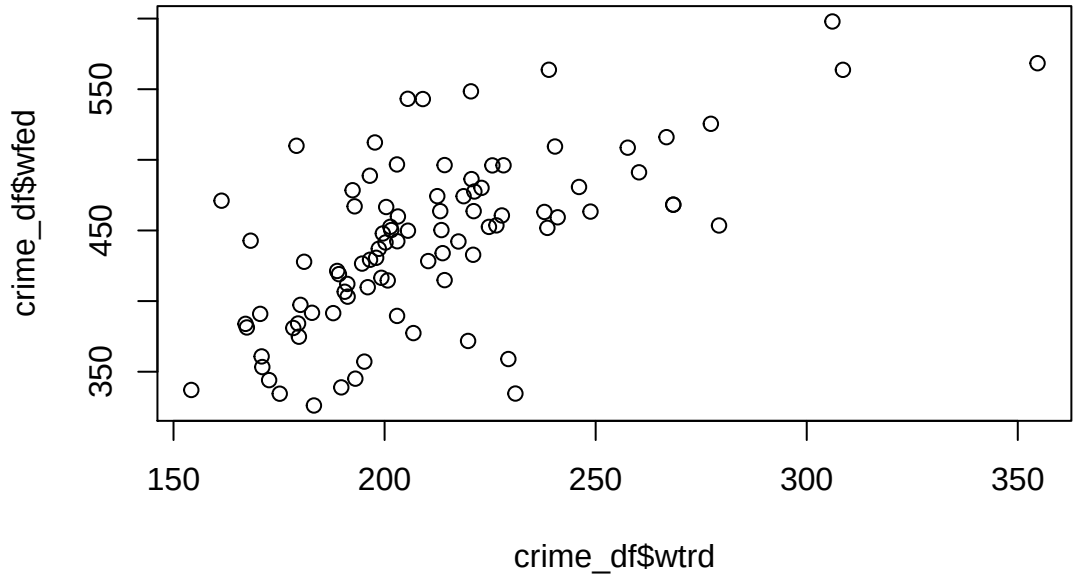

```
plot(crime_df$west, crime_df$pctmin80)
```



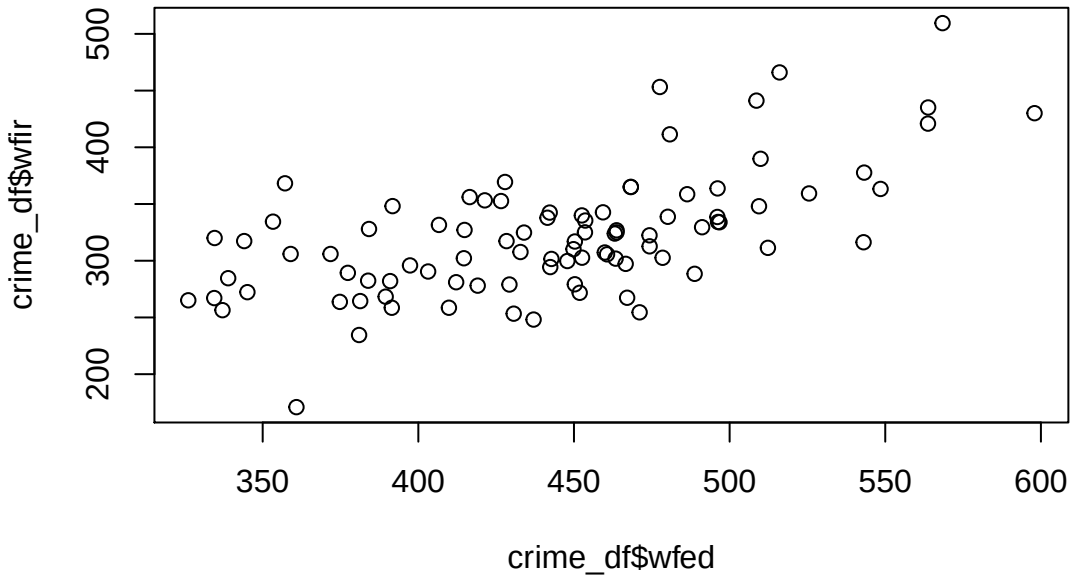

```
plot(crime_df$wtrd, crime_df$wfir)
```



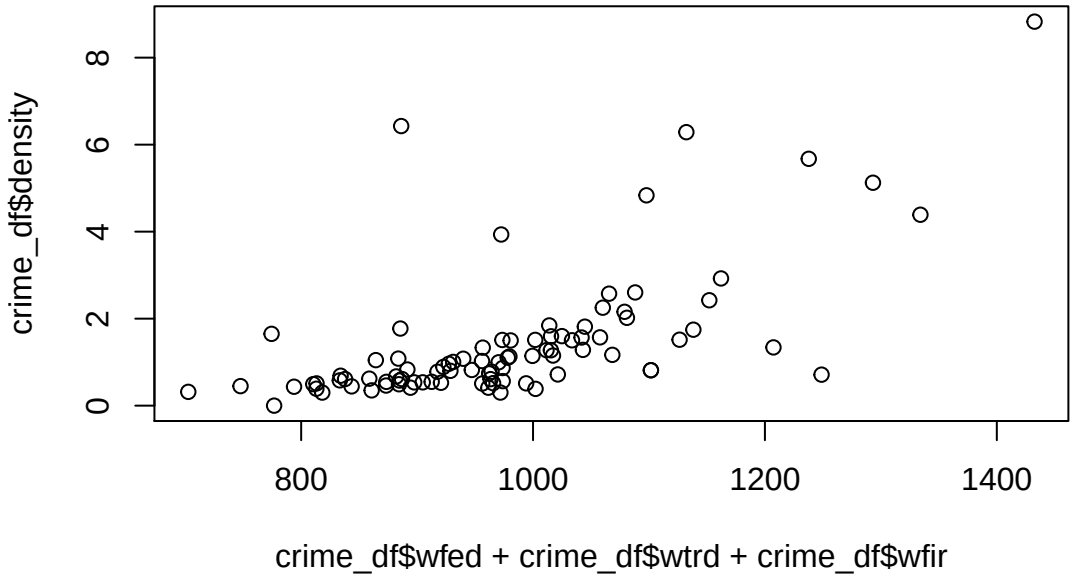

```
plot(crime_df$wtrd, crime_df$wfed)
```




```
plot(crime_df$wfed, crime_df$wfir)
```




```
plot(crime_df$wfed + crime_df$wtrd + crime_df$wfir, crime_df$density)
```



Standardize Independent Variables

In order to compare the impacts of the different independent variables, the values of those variables needed to be standardized so that the slope coefficients are similar in scale (e.g. if the range of a variable is between 0 and 1, then the coefficient may be larger than that of a variable that ranges between 0-200). For the standardization, the variables were all scaled to range between 0 and 1, based on the min and max values.

```
# make a copy of crime_df for standardizing values
std_crime_df <- cbind(crime_df)

# a function to standardize values (fraction of range)
standardize_values <- function(x) {
    (x - min(x))/(max(x) - min(x))
}

# for all columns other than county number, year, and crime
```

```r
# rate, standardize between 0 and 1
for (col in 4:ncol(std_crime_df)) {
    std_crime_df[, col] <- standardize_values(std_crime_df[,
        col])
}

summary(std_crime_df)
```

```
##      county           year         crmrte              prbarr
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.0000
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.1131
##  Median :105.0   Median :87   Median :0.029986   Median :0.1785
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.2025
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.2521
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.0000
##     prbconv          prbpris           avgsen            polpc
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.1350   1st Qu.:0.4773   1st Qu.:0.1279   1st Qu.:0.05837
##  Median :0.1873   Median :0.6076   Median :0.2428   Median :0.08900
##  Mean   :0.2352   Mean   :0.5795   Mean   :0.2785   Mean   :0.11510
##  3rd Qu.:0.2535   3rd Qu.:0.6817   3rd Qu.:0.3943   3rd Qu.:0.13611
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##     density            taxpc              west           central
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.06201   1st Qu.:0.05283   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.10900   Median :0.09756   Median :0.0000   Median :0.0000
##  Mean   :0.16186   Mean   :0.13142   Mean   :0.2527   Mean   :0.3736
##  3rd Qu.:0.17765   3rd Qu.:0.16217   3rd Qu.:0.5000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##      urban           pctmin80           wcon             wtuc
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.1358   1st Qu.:0.2350   1st Qu.:0.4394
##  Median :0.00000   Median :0.3652   Median :0.3611   Median :0.5143
##  Mean   :0.08791   Mean   :0.3839   Mean   :0.3772   Mean   :0.5264
##  3rd Qu.:0.00000   3rd Qu.:0.5845   3rd Qu.:0.4983   3rd Qu.:0.6011
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       wtrd             wfir             wser             wmfg
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.1828   1st Qu.:0.3414   1st Qu.:0.04727   1st Qu.:0.2686
##  Median :0.2435   Median :0.4324   Median :0.05880   Median :0.3326
##  Mean   :0.2861   Mean   :0.4465   Mean   :0.06973   Mean   :0.3640
##  3rd Qu.:0.3538   3rd Qu.:0.5152   3rd Qu.:0.07216   3rd Qu.:0.4131
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##       wfed             wsta             wloc             mix
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2727   1st Qu.:0.2943   1st Qu.:0.3901   1st Qu.:0.1372
##  Median :0.4552   Median :0.4118   Median :0.4625   Median :0.1846
##  Mean   :0.4297   Mean   :0.4111   Mean   :0.4936   Mean   :0.2452
##  3rd Qu.:0.5589   3rd Qu.:0.5150   3rd Qu.:0.6049   3rd Qu.:0.2966
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     pctymle
##  Min.   :0.00000
##  1st Qu.:0.06579
##  Median :0.08338
```

```
##  Mean   :0.11688
##  3rd Qu.:0.11439
##  Max.   :1.00000
```

Now that we have standardized the units of all input variables, we can compute model slope coefficients that will be in comparable units.

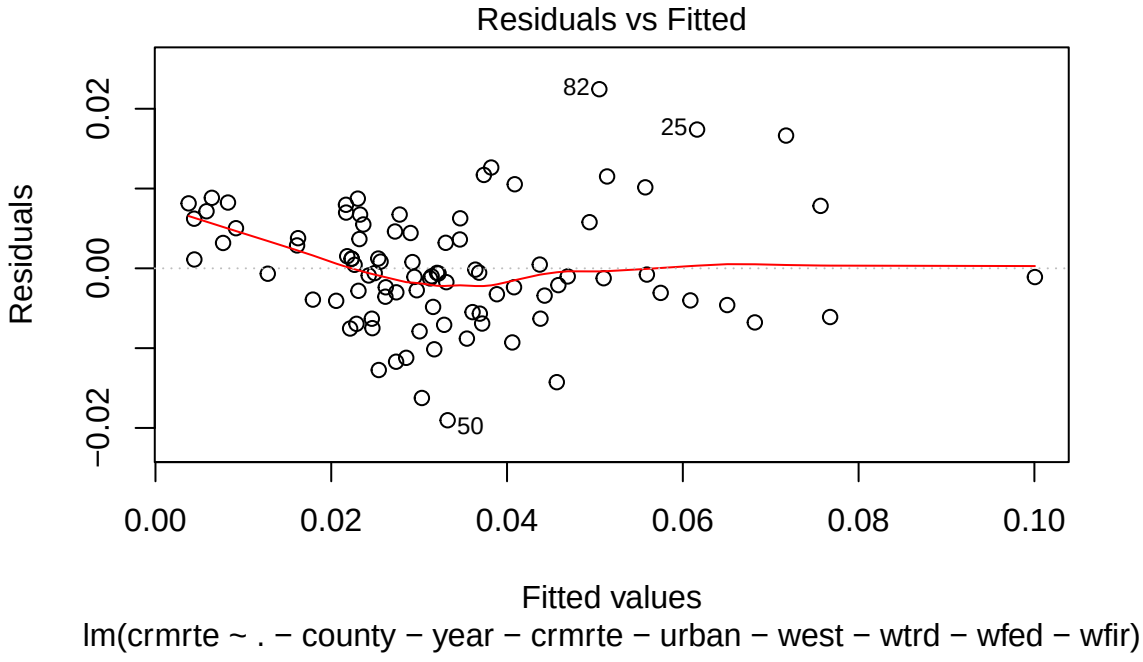
Standardized Regression Model

A multi variable regression model was created using the data set that has been standardized above.

Then the model was evaluated for potential high leverage/influence data points as well as potential biases.

In review the following findings were noted: - row 84 and 25 have a high Cook's distance and high standardized residuals, which means the data point can be problematic for the regression model. - row 25 and 84 were also noted earlier to be an extreme outier for the wser variable. Thus based on this finding the point will be removed and the regression will be redone. - Judging from the residuals vs. fitted plot the model may have some bias when the predicted value crmrte is between 0 to 0.04. Particularly the model tend to underpredict lower crmrates, and overpredict medium crmrte. - From the Normal Q-Q line, it looks like that majority of predictions follow the line, indicating a normal and independent distribution.

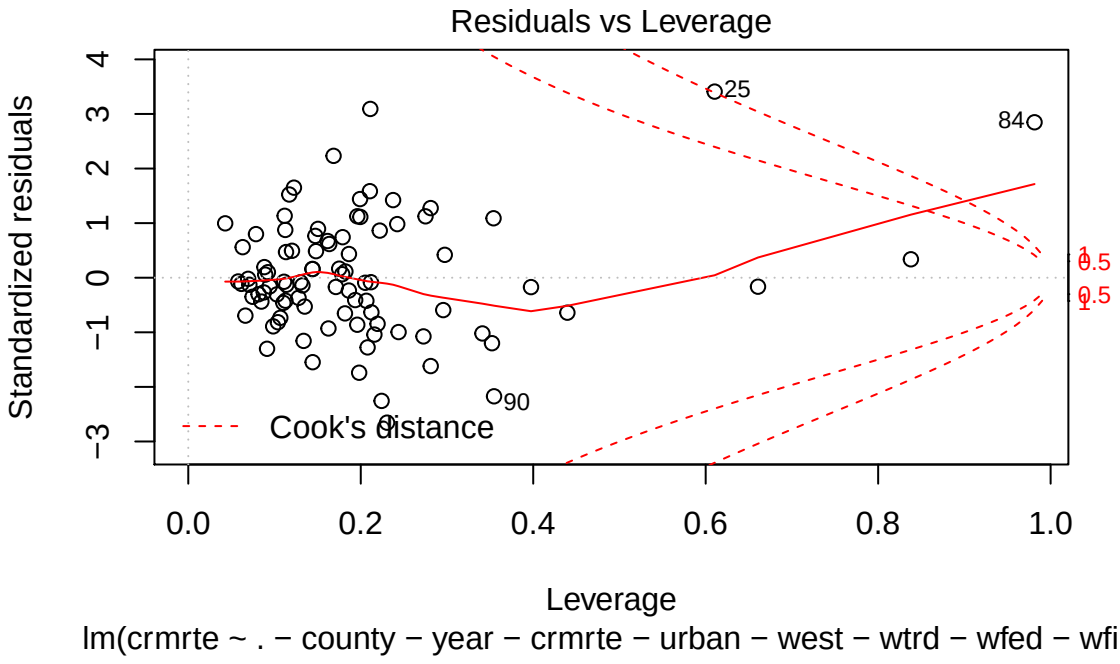
```
# TODO clean out the warning
std_model <- lm(crmrte ~ . - county - year - crmrte - urban -
    west - wtrd - wfed - wfir, data = std_crime_df)

plot(std_model, 1)
```



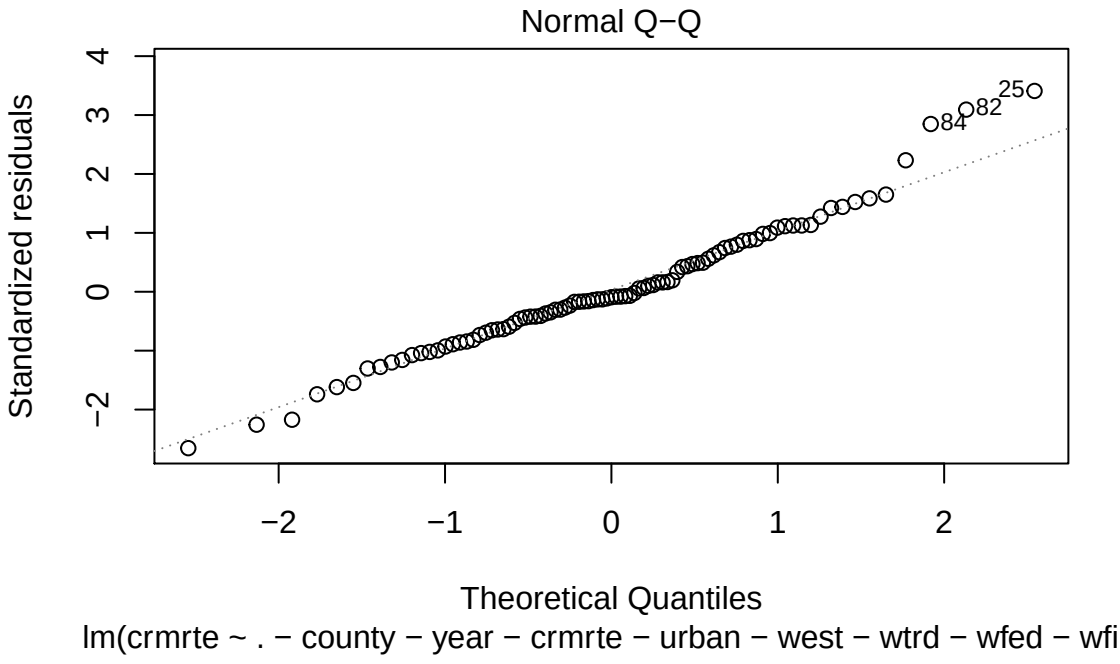

```
plot(std_model, 5)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

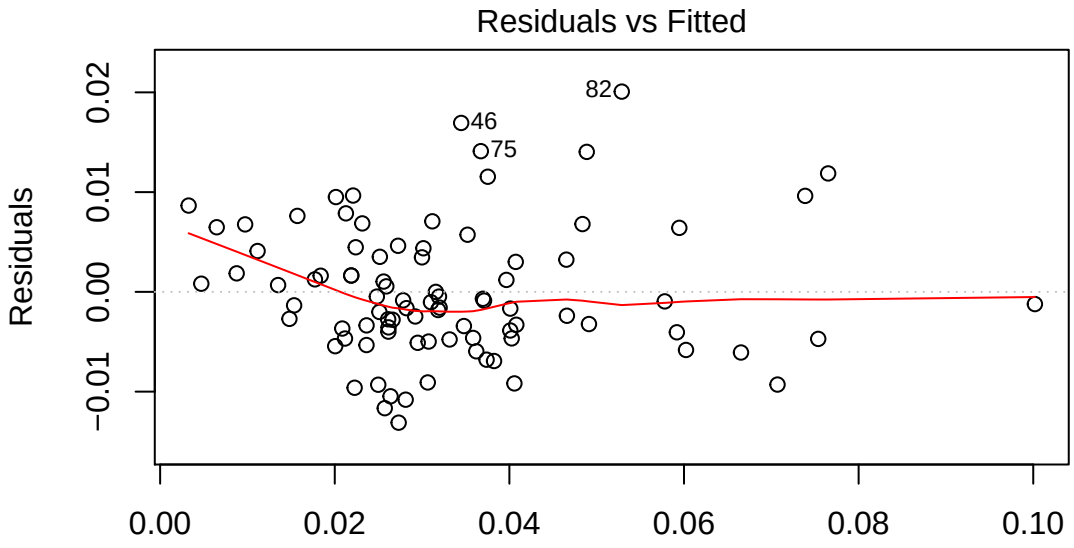



```
plot(std_model, 2)
```

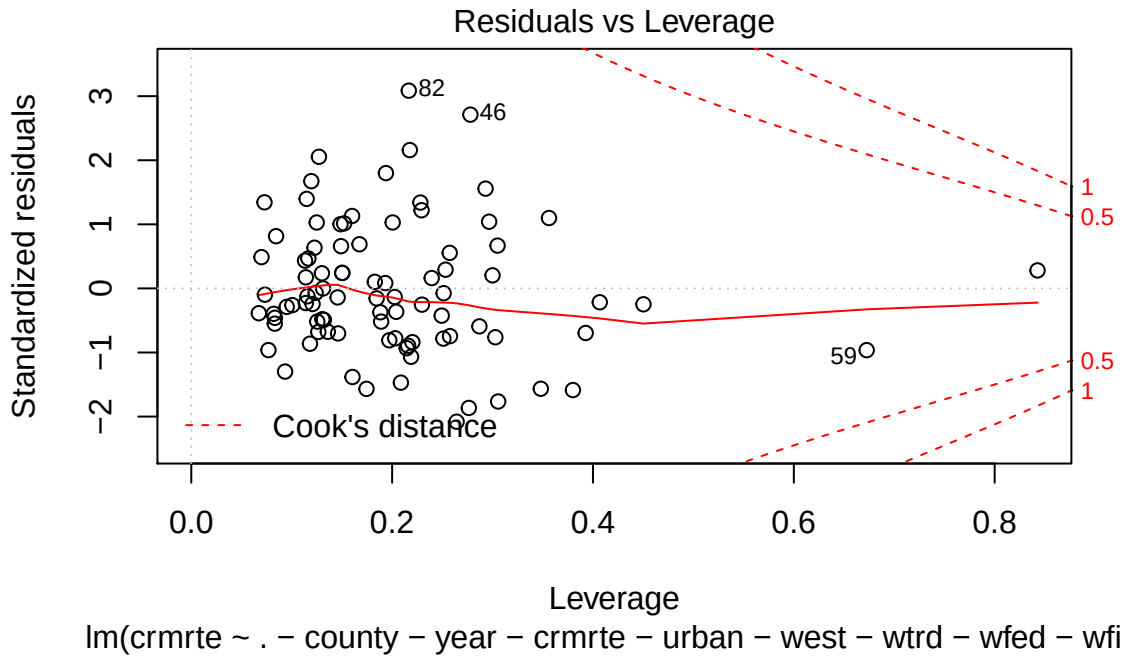



```
# summary(std_model)$r.squared
```

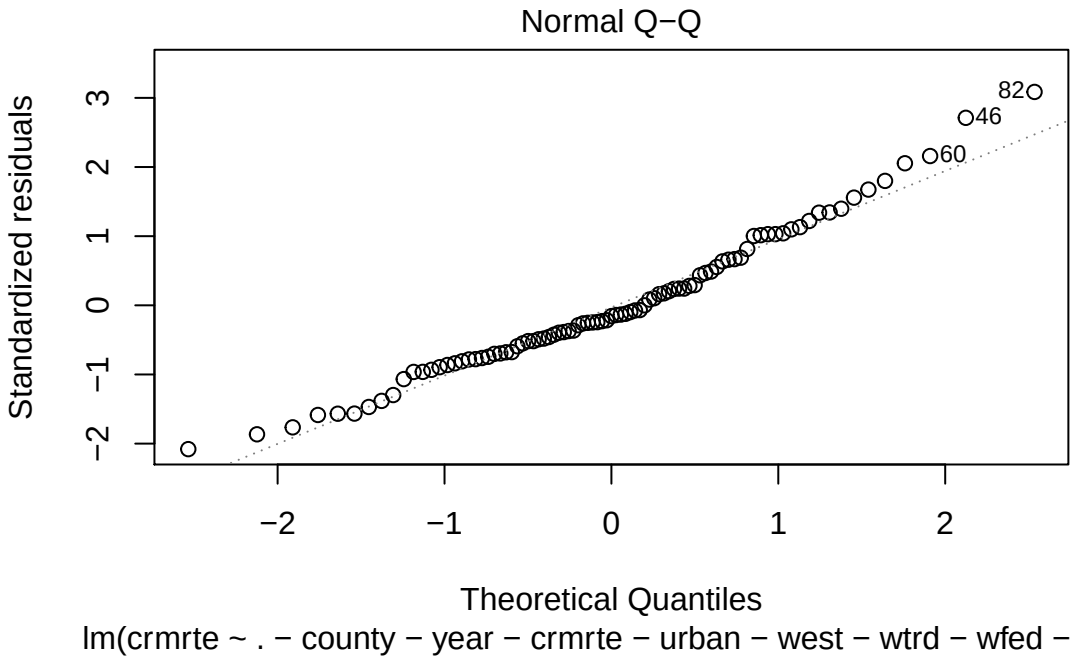
```
std_crime_df2 <- std_crime_df[-c(84, 25), ]

std_model2 <- lm(crmrte ~ . - county - year - crmrte - urban -
    west - wtrd - wfed - wfir, data = std_crime_df2)

plot(std_model2, 1)
```




```
plot(std_model2, 5)
```



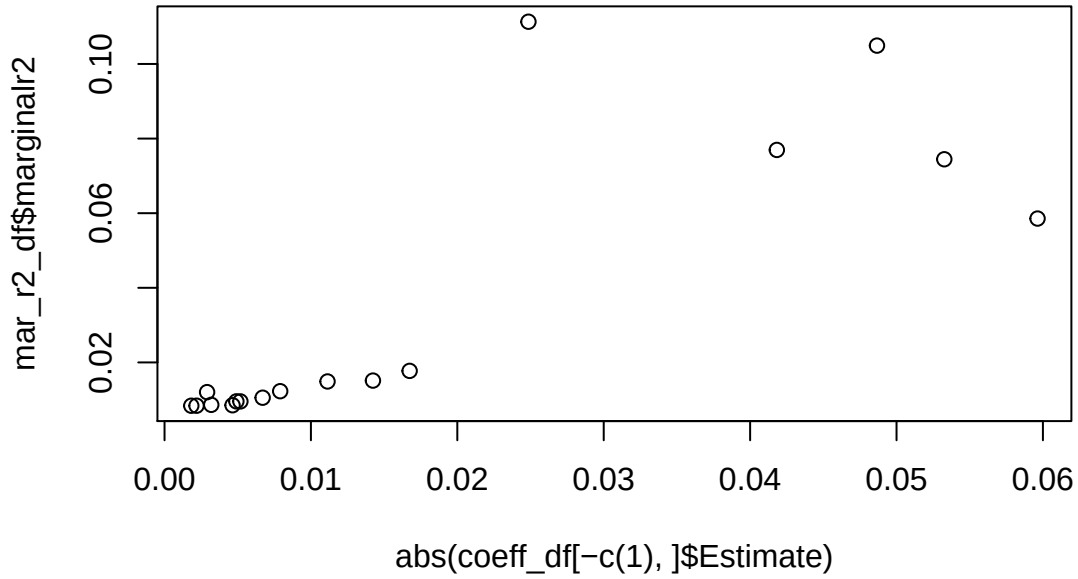

```
plot(std_model2, 2)
```



In order to find which variables are most impactful to crmrte, the marginal R-squared against the standardized coefficients were reviewed. Based on the plots, the following variables were found to have the highest marginal R-squared and absolute slope coefficient: -prbarr -prbconv -polpc -density -pctmin80

```
coeff_df = data.frame(summary(std_model)$coefficients)
# summary(std_model)$r.squared

# base R-Squared
base_model <- lm(crmrte ~ . - county - year - crmrte, data = std_crime_df)
base_r2 <- summary(base_model)$r.squared

# create list of variables for the for-loop
var_names <- colnames(std_crime_df)
remove <- c("county", "year", "crmrte", "urban", "west", "wtrd",
    "wfed", "wfir")
var_names <- var_names[!var_names %in% remove]

# initiate an empty vector to store the marginal R-Squared
var_r2_delta = c()

# loop through the variable names and store the marginal
# R-Squared
for (i in var_names) {
    fmla <- as.formula(paste("crmrte ~ - crmrte +", paste(var_names[!var_names %in%
        i], collapse = "+")))
    delta_model <- lm(fmla, data = crime_df)
    r2_delta <- base_r2 - summary(delta_model)$r.squared
    var_r2_delta <- c(var_r2_delta, r2_delta)
}

# put the variable and marginal R-squared in a dataframe
mar_r2_df <- data.frame(v1 = var_names, v2 = var_r2_delta)
colnames(mar_r2_df) <- c("variable", "marginalr2")
```

```r
# sort dataframe by marginal R-squared in a descending order
# mar_r2_df <- mar_r2_df[rev(order(mar_r2_df$marginalr2)),]

plot(abs(coeff_df[-c(1), ]$Estimate), mar_r2_df$marginalr2)
```




```r
subset(mar_r2_df, marginalr2 > 0.04)
```

```
##   variable marginalr2
## 1   prbarr 0.07445392
## 2  prbconv 0.07695649
## 5    polpc 0.05856302
## 6  density 0.10492397
## 9 pctmin80 0.11132549
```

Non-Standardized Regressions

The following is the model that contains almost all available variables as explanatory variables with the exception of variables we excluded due to potential multi-collinearity.

```r
crime_df2 <- crime_df[-c(84, 25), ]

model1 <- lm(crmrte ~ . - county - year - crmrte - urban - west -
    wtrd - wfed - wfir, data = crime_df2)

summary(model1)$r.squared
```

```
## [1] 0.8688977
```

```r
summary(model1)$coefficients
```

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  3.097640e-02 1.561081e-02  1.9842914 5.108937e-02
## prbarr      -5.078247e-02 8.704418e-03 -5.8341022 1.478721e-07
## prbconv     -1.962352e-02 3.293124e-03 -5.9589361 8.903946e-08
## prbpris      4.754774e-03 1.048442e-02  0.4535087 6.515656e-01
## avgsen      -3.961756e-04 3.497705e-04 -1.1326727 2.611624e-01
```

```
## polpc        6.460940e+00 1.346144e+00  4.7995886 8.534766e-06
## density      6.845704e-03 7.973078e-04  8.5860246 1.369694e-12
## taxpc       -7.103721e-05 1.021711e-04 -0.6952767 4.891513e-01
## central     -3.517708e-03 1.926533e-03 -1.8259265 7.206619e-02
## pctmin80     3.898315e-04 5.176540e-05  7.5307361 1.236934e-10
## wcon         4.081808e-05 2.373695e-05  1.7196007 8.986184e-02
## wtuc         4.373842e-06 1.324754e-05  0.3301626 7.422493e-01
## wser        -6.293562e-05 2.794740e-05 -2.2519307 2.742112e-02
## wmfg         4.568252e-06 1.208881e-05  0.3778909 7.066390e-01
## wsta        -4.273992e-05 2.105298e-05 -2.0301130 4.609284e-02
## wloc         4.531803e-05 4.143979e-05  1.0935875 2.778324e-01
## mix         -2.294321e-02 1.269191e-02 -1.8077035 7.488831e-02
## pctymle      9.580106e-02 3.779334e-02  2.5348663 1.345432e-02
```

The following is the model that contains a transformed explanatory variable.

```
model_transform <- lm(crmrte ~ prbarr + log(prbconv) + density,
    data = crime_df2)

summary(model_transform)$r.squared
```

```
## [1] 0.6570935
```

```
summary(model_transform)$coefficients
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)   0.025420503 0.0035106022  7.241066 1.857260e-10
## prbarr       -0.028710438 0.0089889944 -3.193954 1.969045e-03
## log(prbconv) -0.006276946 0.0022761837 -2.757662 7.125235e-03
## density       0.007903815 0.0008331222  9.486981 5.580124e-15
```

The following is the model that contains only variables that were identified to be most relevant to crmrte based on their marginal R-squared and standardized slope coefficient values.

```
model_key <- lm(crmrte ~ prbarr + prbconv + polpc + density +
    pctmin80, data = crime_df2)

summary(model_key)$r.squared
```

```
## [1] 0.8204393
```

```
summary(model_key)$coefficients
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  0.0300488820 3.494735e-03  8.598328 4.156915e-13
## prbarr      -0.0555832603 8.317408e-03 -6.682763 2.515871e-09
## prbconv     -0.0179293179 3.139371e-03 -5.711118 1.698543e-07
## polpc        6.1601721055 1.204450e+00  5.114512 1.989594e-06
## density      0.0063705861 6.966292e-04  9.144873 3.349488e-14
## pctmin80     0.0003808799 5.212093e-05  7.307620 1.527153e-10
```

Stargazer Regression Table for Model Specifications

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(model_transform, model_key, model1, title = "Linear Models Parameters Predicting Crime Rate",
    type = "text", report = "vc", keep.stat = c("rsq", "n"),
    omit.table.layout = "n")
```

Linear Models Parameters Predicting Crime Rate

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
prbarr	-0.029	-0.056	-0.051
log(prbconv)	-0.006		
prbconv	-0.018	-0.020	
prbpris	0.005		
avgsen	-0.0004		
polpc	6.160	6.461	
density	0.008	0.006	0.007
taxpc	-0.0001		
central	-0.004		
pctmin80	0.0004	0.0004	
wcon	0.00004		
wtuc	0.00000		
wser	-0.0001		
wmfg	0.00000		
wsta	-0.00004		
wloc	0.00005		
mix	-0.023		
pctymle	0.096		
Constant	0.025	0.030	0.031
Observations	89	89	89
R2	0.657	0.820	0.869

Recommendation

For interpretability purposes, the model was re-done using non-standardized variables: -prbarr -prbconv -polpc -density -pctmin80

Recommendation for political campaign: - police per capita has a positive slope coefficient with crmrte, and this may be due to more police are present in areas with high crmrte. This suggests that purely hiring more police officers may not be an impactful solution. - However probability of arrest and conviction both have a negative slope coefficients. The model suggests that perhaps a zero tolerance policy towards crime is needed to increase arrests and convictions and thus deter crimes from happening. - In terms areas with large minority population and high density, since these variable cannot be changed that much, perhaps a community outreach (e.g. job training program, afterschool programs, tutor/mentor program) to educate areas with a lot of minority can be done, so that crimes can be reduced in those areas.

Omitted Variables

Potential Omitted Variable #1: poverty_rate

$$crmrte = \beta_0 + \beta_1 * density + \beta_2 * poverty\_rate + u$$

$$poverty\_rate = \alpha_0 + \alpha_1 * density + u$$

- One thing that was noticeable in the data is that crmrate was higher in dense areas and large minority population, however this may be due to an omitted variable that is not available in the data set.
- For example: in dense areas the cost of living may be much higher, which can explain why higher wages are correlated with dense areas, but because of the higher cost of living. Because of this, there may be a lot more people living under the poverty line, which would encourage them to commit crimes and hence why dense areas have higher crmrte.
- so the density slope coefficient in this instance is probably higher than it should be β_2 and α_1 would be positive.
- Maybe tax revenue or wages can help proxy this omitted variable.

Potential Omitted Variable #2: discrimination

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * discrimination$$

$$discrimination = \alpha_0 + \alpha_1 * pctmin80$$

- Similarly minorities may be arrested for crimes more often than necessary due to discrimination. - in this scenario β_2 and $alpha_1$ would be a positive value.

Potential Omitted Variable #3: raised_in_oneparent_hh

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * raised\_in\_2parents\_hh$$

$$raised\_in\_2parents\_hh = \alpha_0 + \alpha_1 * pctmin80$$

- In this scenario, minorities may be more likely to be raised in a single parent house hold. Thus making them more likely to commit crimes. - β_2 would be positive and α_1 would be negative.

Potential Omitted Variable #4: unemployment

$$crmrte = \beta_0 + \beta_1 * density + \beta_2 * unemployment$$

$$unemployment = \alpha_0 + \alpha_1 * density$$

- Higher umployment = higher crime rate (beta2 > 0)
- Higher density = higher unemployment (alpha1 > 0)
- beta1 was positive, therefore, it might be higher than it should've been.

Potential Omitted Variable #5: years_of_education

$$crmrte = \beta_0 + \beta_1 * pctmin80 + \beta_2 * years\_of\_education$$

$$years\_of\_education = \alpha_0 + \alpha_1 * pctmin80$$

- Higher avg years of education for a county would result in lower crime rate, beta2 < 0 - Higher percentage of minorities = lower average years of education for a county, alpha1 <0 - $\beta_2 * \alpha_1 > 0$, beta1 >0, therefore, it might be higher than it should've been.