W203 Lab 3

Armand Kok, Adam Yang, James De La Torre

Introduction

Is the introduction clear? Is the research question specific and well defined? Could the research question lead to an actionable policy reccomendation? Does it motivate the analysis? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.

Our team has been hired by a local political campaign to provide research on North Carolina crime statistics and to generate policy suggestions for reducing crime. Our candidate seeks to portray herself as being "pro-cop" and "tough on crime", and she espouses strong policing and enforcement. She also has a strong desire to understand the situations faced by the minority population within the state, and she has expressed a keen interest in understanding how minority communities are impacted by crime.

The crime statistics dataset provided for analysis is a subset of the data used by Cornwell and W. Trumball in their 1994 study. The dependent variable, of our study is the crimes commited per capita, given as `crmrate`, while there are 24 other variables in the dataset, each of which can be potential modulators of the crime rate. We aim to build a linear model that regresses `crmrate` on the key variables in the dataset. In particular, we are interested in examining the potential of the following policies in reducing crime rate: * Policy to increase the police per capita of a county * Policy to implement a more stringent arrest protocol * Policy to enhance community outreach in high density and minority communities

In addition, we aim to identify other factors that may reduce crime and attempt to fully explore other possible political strategies. Not all correlating variables will have an actionable solution, though their inclusion in the regression model will contribute to its accuracy.

2.0 Data Loading and Cleaning

TO DO: Look for any top-coding or bottom coding. TO DO: remove this instructiuon line **Did the team notice any anomalous values? Is there a sufficient justification for any data points that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Overall, does the report demonstrate a thorough understanding of the data?**

The data provided is a sample from 91 counties in North Carolina, containing information from 1987. The variables in the dataset and their meanings are shown below:

Variable	Label	Variable	Label
county	county identifier	**urban**	=1 if in SMSA
year	1987	**pctmin80**	perc. minority, 1980
crmrte	crimes committed per person	**wcon**	weekly wage, construction
prbarr	'probability' of arrest *****	**wtuc**	wkly wge, trns, util, commun
prbconv	'probability' of conviction *****	**wtrd**	wkly wge, whlesle, retail trade
prbpris	'probability' of prison sentence *****	**wfir**	wkly wge, fin, ins, real est
avgsen	avg. sentence, days	**wser**	wkly wge, service industry
polpc	police per capita	**wmfg**	wkly wge, manufacturing
density	people per sq. mile	**wfed**	wkly wge, fed employees
taxpc	tax revenue per capita	**wsta**	wkly wge, state employees
west	=1 if in western N.C.	**wloc**	wkly wge, local gov emps

Variable	Label	Variable	Label
central	=1 if in central N.C.	**mix**	offense mix: face-to-face/other
pctymle	percent young male		

* *These are not true probabilities that are limited between 0 and 1, but are rations instead. For example,* `probconv` *is the ratio of the number of convictions to the number of arrests, which can be larger than 1.*

2.1 Loading the Data

The data file, `crime_v2.csv` was opened and found to contain 97 rows.

```
# Import all libraries that will be used in the lab
library(car)
library(reshape2)
library(ggplot2)
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
```

```
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
# 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"))
```

2.2 Data Cleanup

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.

```
# 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
```

```r
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 ...
```

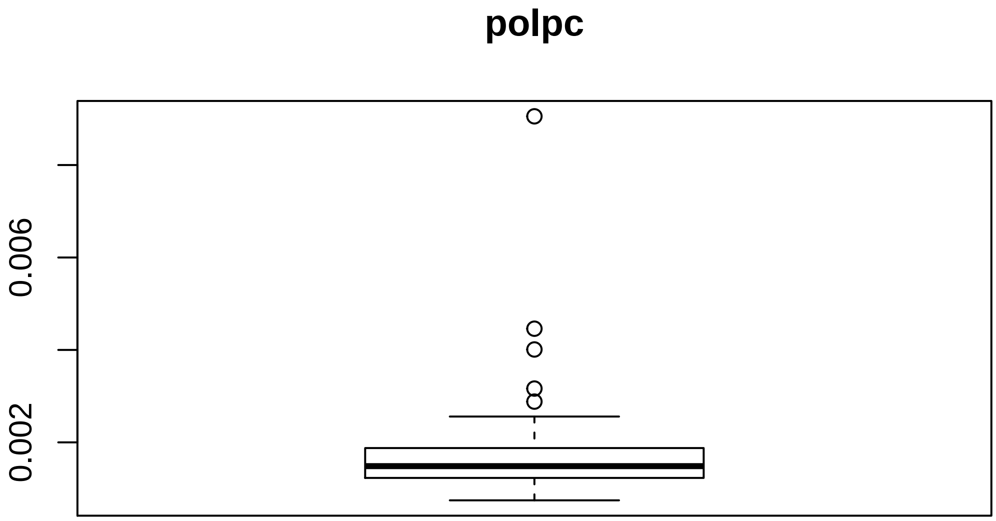
```r
# get rid of rows with missing values (this only kills the 6
# blank rows)
crime_df <- crime_df[complete.cases(crime_df), ]

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

2.3 Outlier Identification

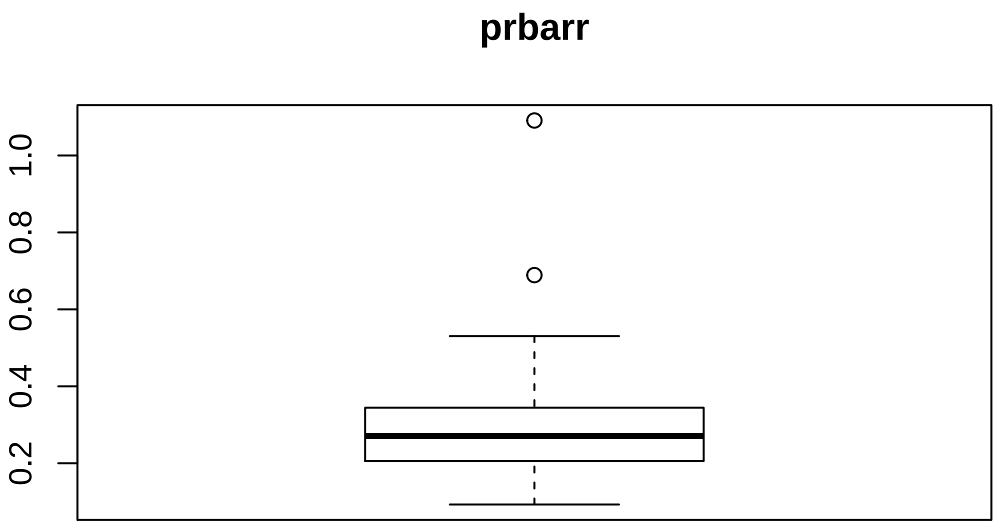
TO DO: Write function that computes outliers by column

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.

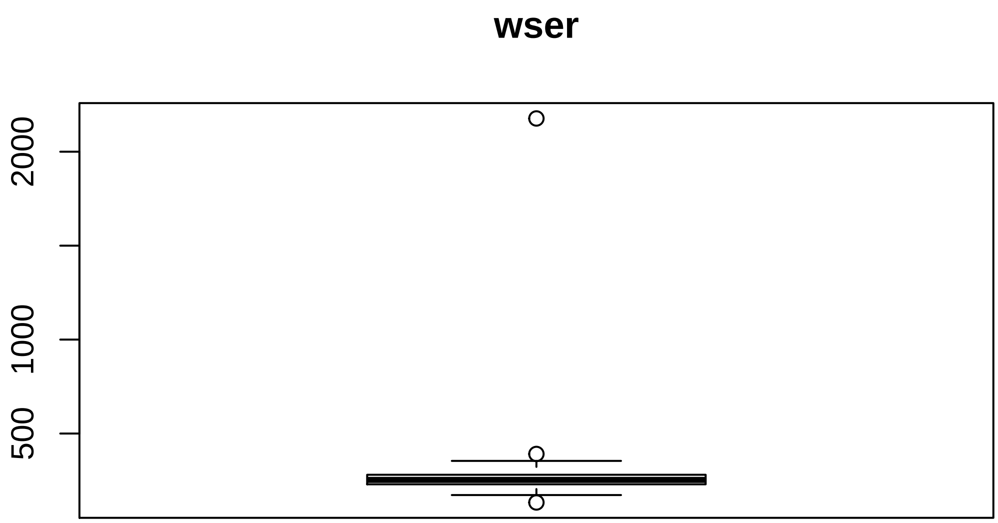
```r
boxplot(crime_df$polpc, main = "polpc")
```



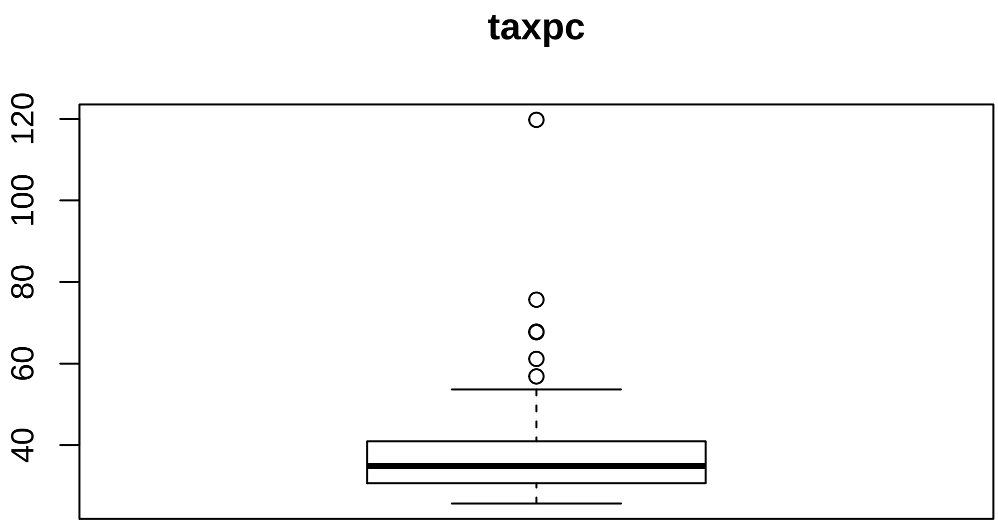

```r
boxplot(crime_df$prbarr, main = "prbarr")
```




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




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




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

3.0 Model Building Process

TO DO: remove this instruction text

Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Is there a thorough univariate analysis of the outcome variable. Did the team identify at least two key explanatory variables and perform a thorough univariate analysis of each? Did the team clearly state why they chose these explanatory variables, does this explanation make sense in term of their research question? Did the team consider available variable transformations and select them with an eye towards model plausibility and interperability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?

TO DO: Mention we looked at histograms of each variable TO DO: Figure out where we want to mention statistical significance of calculated coefficients. At end of each model section? after producing all models?

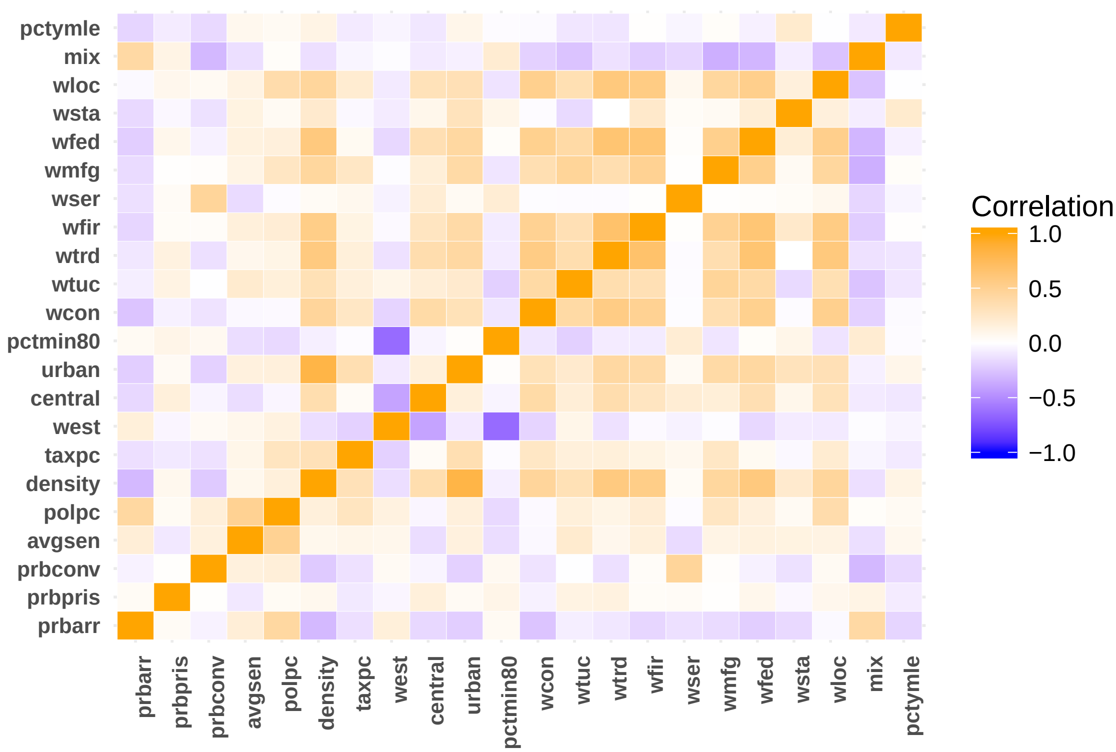
3.1 Check for multicolinearity

TO DO: clean up this sentence

To understand the correlation of each variable in the dataset to crime rate and to detect any collinear relationships between explanatory variables, a correlation matrix was constructed as shown below. This will be useful information as additional variables are added to intial models.

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.

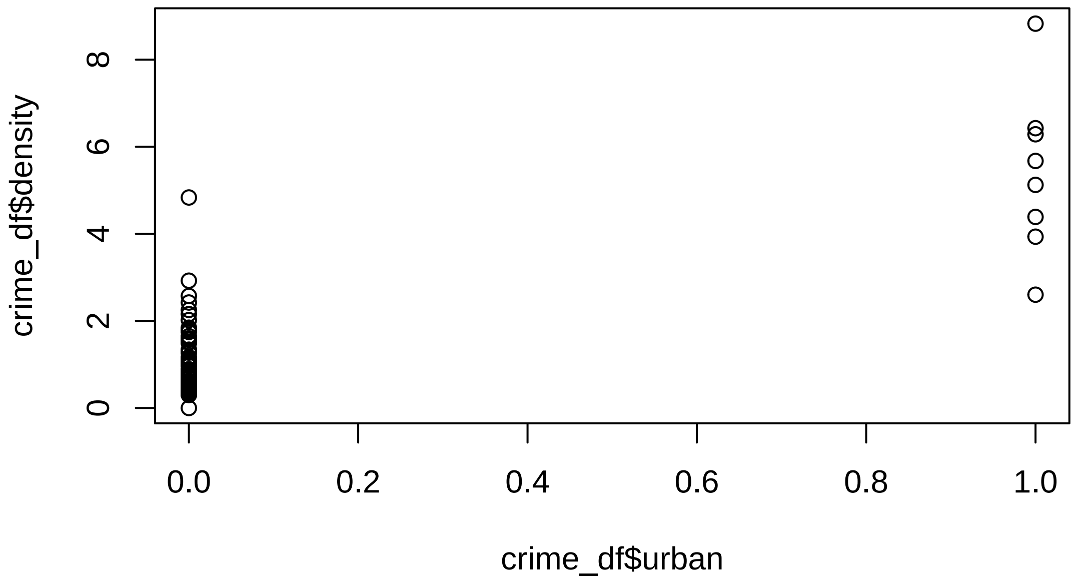
```
# TODO - fix matrix sizing

# 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())
```



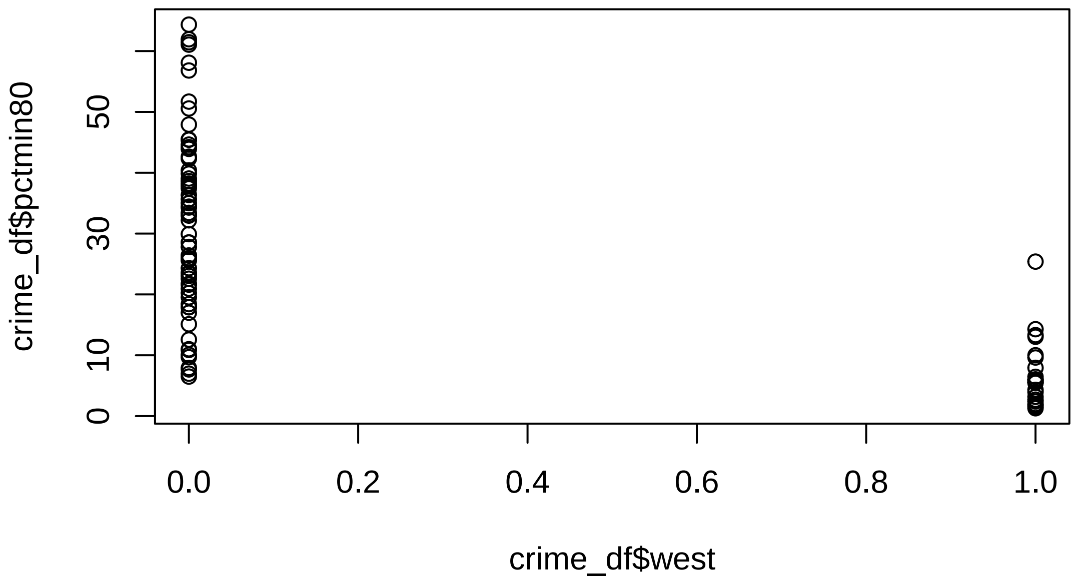
One of the assumptions for multiple OLS regression is to avoid perfect multicollinearity between independent variables. This, however, is not common in practical cases. Less than perfect multicollinearity is a more common problem that will not cause bias in the OLS, but would introduce large variances 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

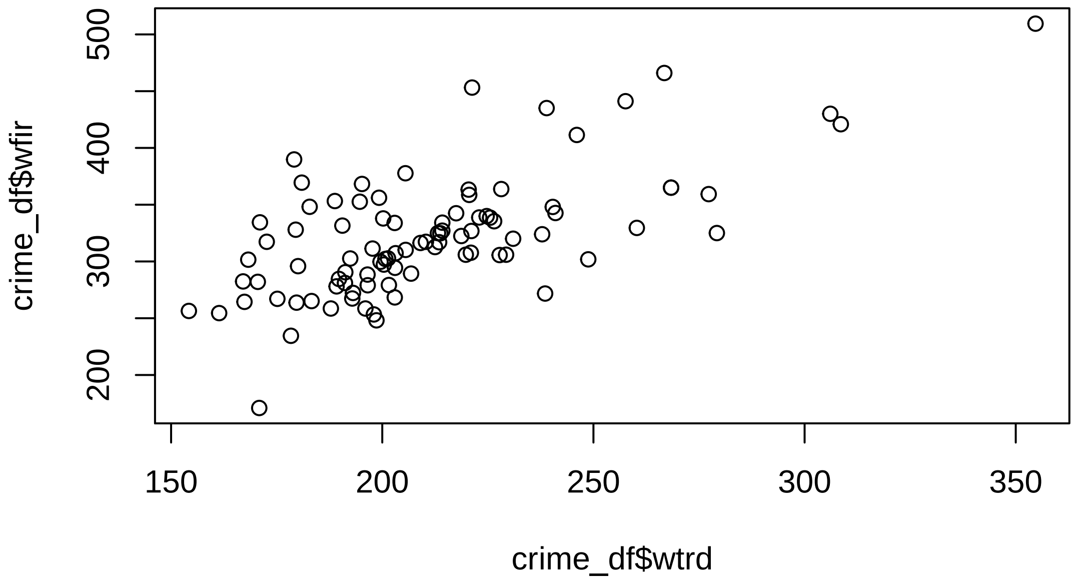
```
plot(crime_df$urban, crime_df$density)
```



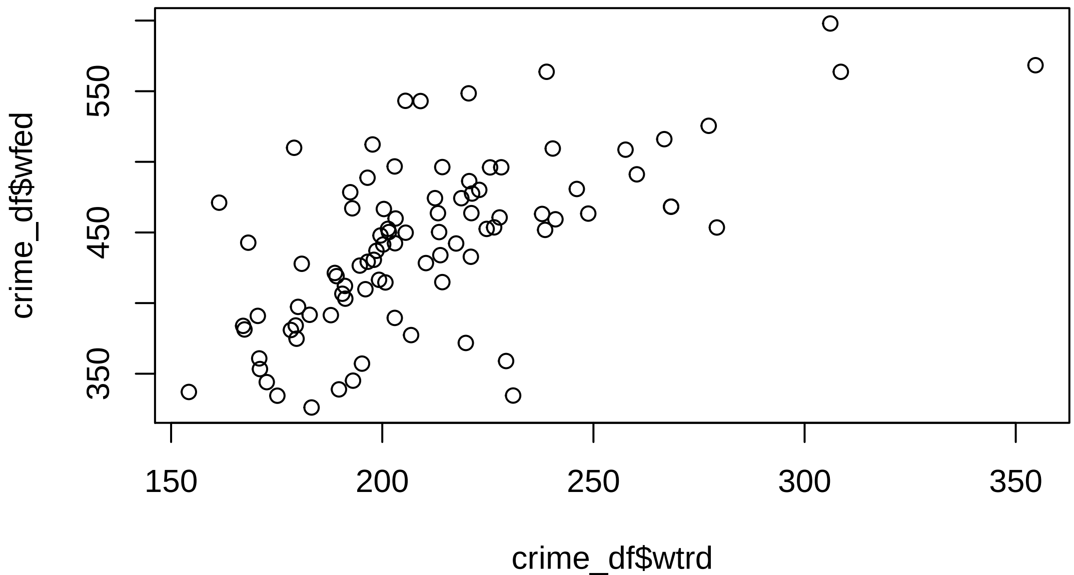

```
plot(crime_df$west, crime_df$pctmin80)
```



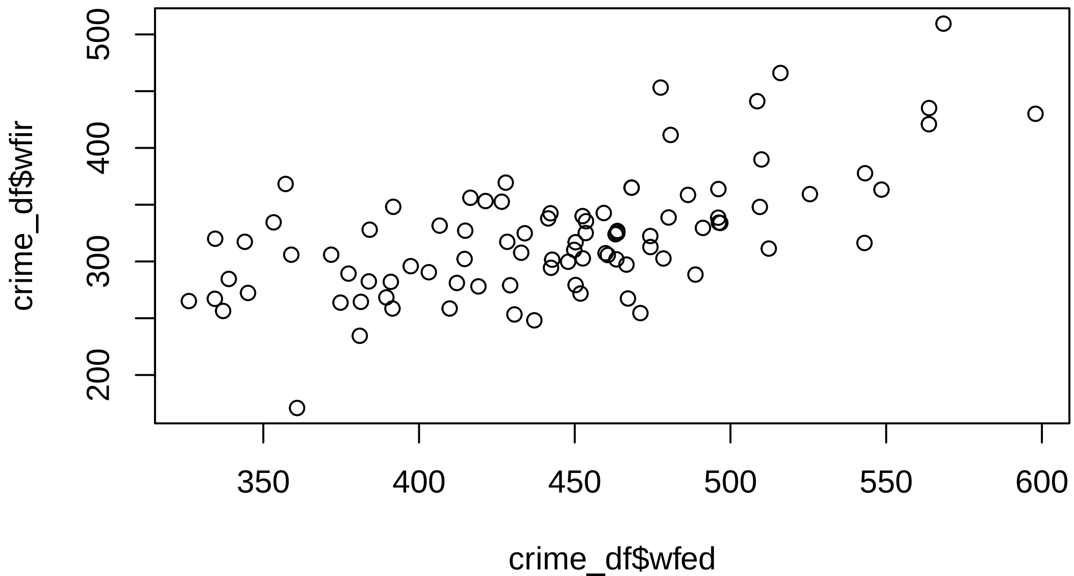

```
plot(crime_df$wtrd, crime_df$wfir)
```



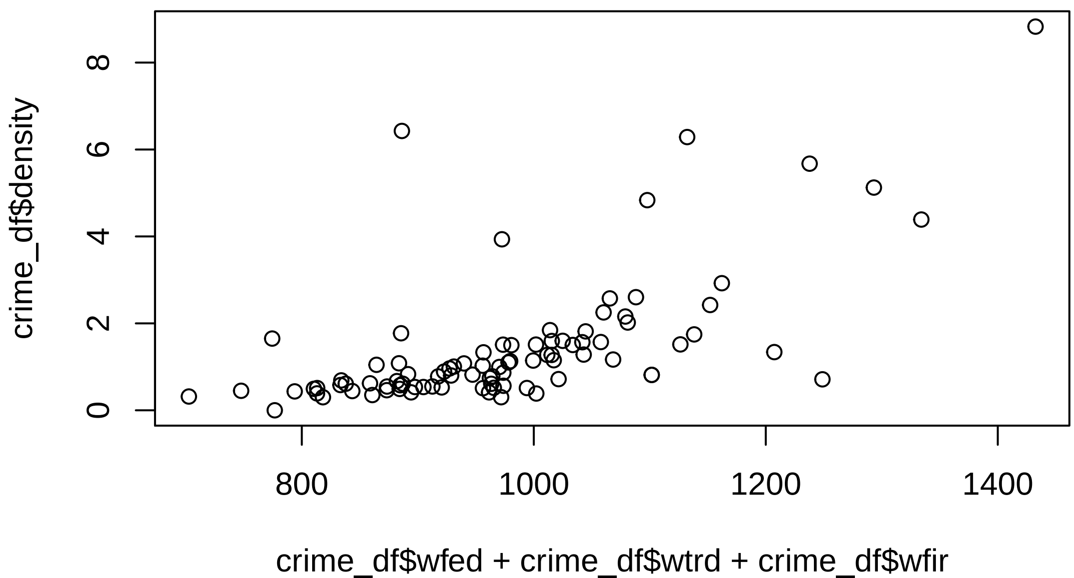

```r
plot(crime_df$wtrd, crime_df$wfed)
```




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



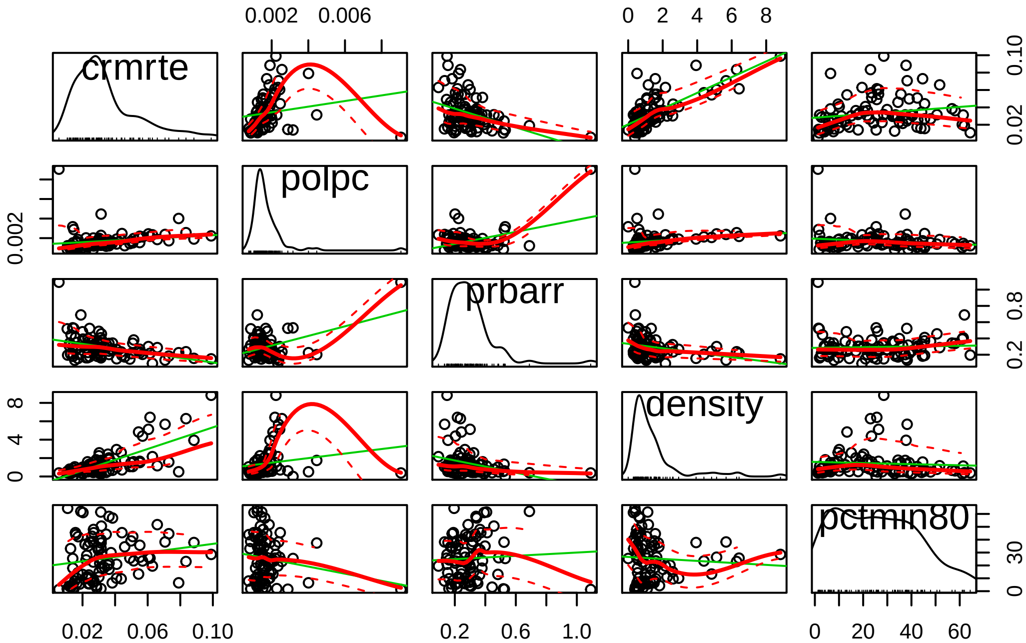

```r
plot(crime_df$wfed + crime_df$wtrd + crime_df$wfir, crime_df$density)
```



3.2 Scatterplot Matrix

To visualize the relationship between crime rate and our explanatory variables of interest, a scatterplot matrix was generated.

```r
spm(~crmrte + polpc + prbarr + density + pctmin80, data = crime_df)
```



The plots reveal that each of the selected explanatory variables shows a relationship with crime rate. There is some degree of nonlinear relationship between `polpc` and `crmrte` and between `pctmin80` and `crmrte`. However, transforming these variables would distort the practical interpretability of any model slope coefficients. Therefore, the variables will not be transformed.

4.0 Regression Models: Base Model

The initial model created contains only those variables directly related to the candidate's positions on being pro-police, for strict enforcement, and concern with inner city and minority communities. Therefore, the variables we have chosen to represent these positions are: probability of arrest (prbarr), density, police per capita (polpc), and the percentage of minorities (pctmin80).

```
model1 <- lm(crmrte ~ prbarr + density + polpc + pctmin80, data = crime_df)
```

After creating the model, we will start by evaluating it against the six Classical Linear Model assumptions.

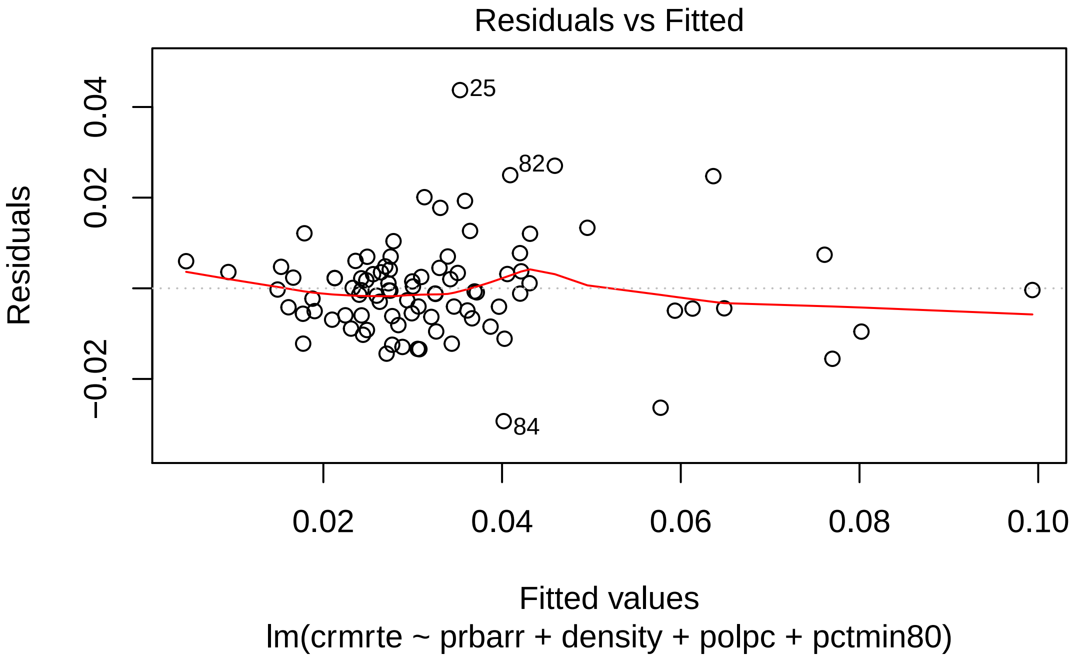
CLM 1. Linear population model: We do not have to worry about this assumption at the moment because we haven't constrained the error term.

CLM 2. Random Sampling: To check random sampling, we need domain knowledge and an understanding of how the data were collected. There are 100 counties in North Carolina, and there are data for 91 of them. Without knowledge of the 9 excluded counties, no statement regarding the validity of random sampling can be made.

CLM 3. No perfect multicollinearity: There is no need to explicitly check for perfect collinearity, because R would've reported a warning if this occurred. Furthermore, the correlation matrix shown in section "**TO DO___**" also shows that there is no perfect collinearity.

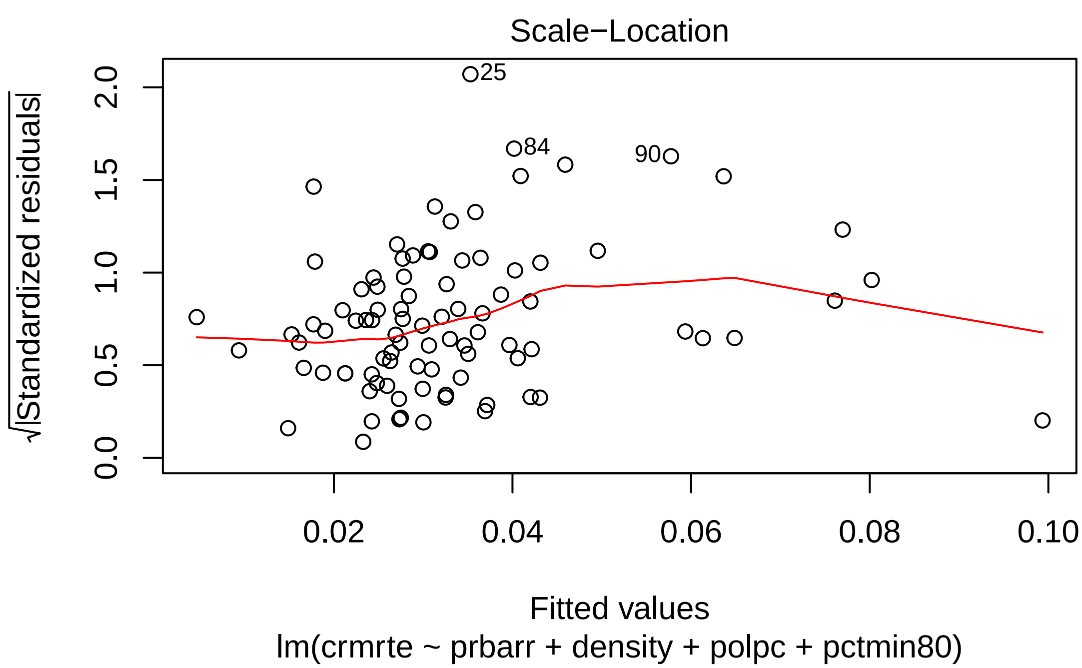
CLM 4. Zero Conditional Mean: $E(u|x) = 0$. For this model, the residuals vs. fitted values plot shown below reveals a relatively flat spline centered around zero. Therefore, there does not seem to be a clear deviation from the zero conditional mean and the assumption holds.

```
# Residuals vs. Fitted Plot
plot(model1, which = 1)
```



CLM 5. Homoscedasticity: In the residuals vs. fitted values plot shown in **CLM 4**, the data points seem to form a cone shape which suggests some heteroscedasticity. In the scale-location plot below, there seems to be a slight positive slope across the range of fitted values between 0.02 and 0.04. Furthermore, the Breusch-Pagan test shown below has a p-value of 6.278e-05 which indicates that the null hypothesis of homoscedasticity can be rejected. When evaluating the statistical significance of calculated model coefficients, heteroscedastic-robust standard errors will be used.

```
# Scale-Location Plot
plot(model1, which = 3)
```

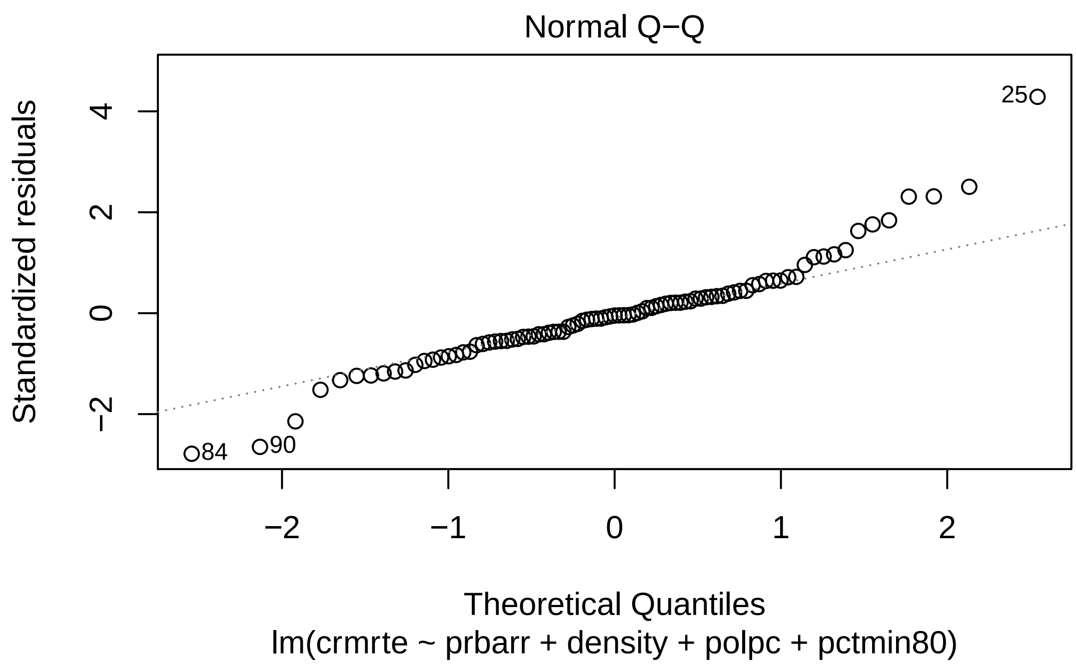



```
# Breusch-Pagan
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 24.521, df = 4, p-value = 6.278e-05
```

CLM 6. Normality of errors: In the Q-Q plot shown below, the bulk of the error terms seem to follow the straight line which suggests a fairly normal distribution. However, the standardized residuals show some deviation from the straight line at the extreme ends of the distribution. This suggests some skew at the extreme ends of our residuals. Furthermore, the Shapiro test shown below has a p value of 0.0002 which means we can reject the null hypothesis of the residuals having a normal distribution.

```
# Q-Q plot of Standardized Residuals
plot(model1, which = 2)
```

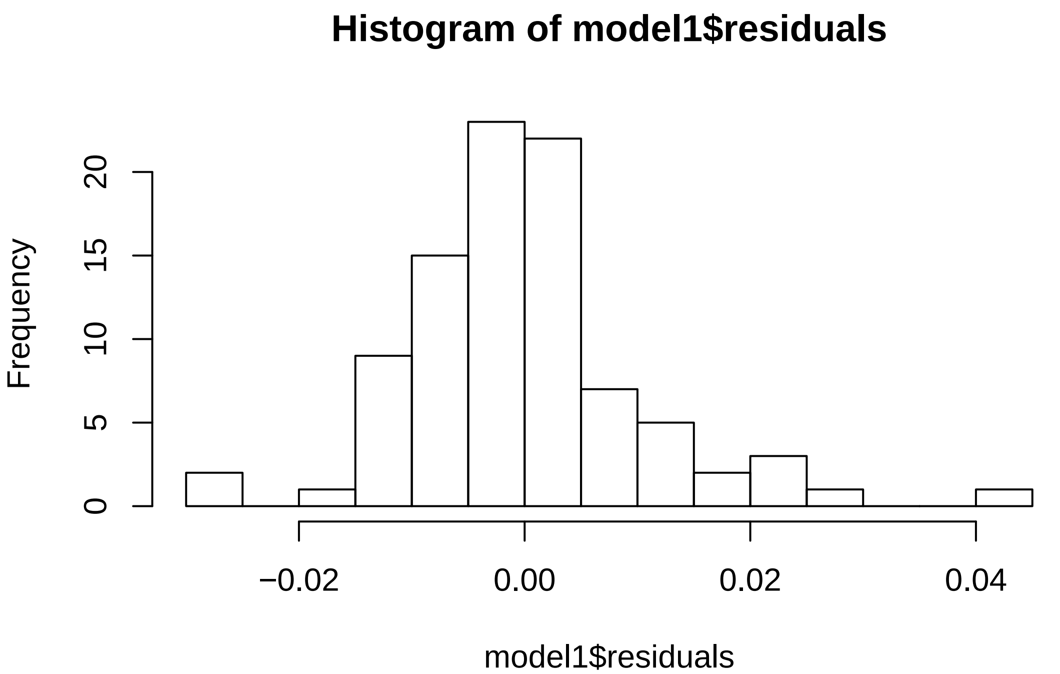



```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.93713, p-value = 0.0002688
```

To further verify this observation, a histogram of this model's residuals is shown below. The histogram shows approximate normality near the center of the distribution, but also some evidence of skewness; especially on the positive end. However, the Central Limit Theorem (CLT) claims that if the sample size is large enough we can assume that the residuals have a normal sampling distribution. For distributions with a very strong skew, a much larger sample size may be required, but for minor skews as in this case, the rule of thumb is that the CLT can be applied when the sample size is greater than 30. The sample size used for this model is 91 which should be enough for the CLT to hold.

```r
hist(model1$residuals, breaks = 20)
```



Based on our review of the six CLM assumptions, this is a valid linear model. We replaced the regular standard errors with the heteroskedasticity-robust standard errors. The resulting coefficients and parameters of the model are shown below:

```r
# TO DO: Use this at the end (section 4.3)

# Replace regular Standard Errors with the
# heteroskedasticity-robust Standard Errors se.model1 <-
# sqrt(diag(vcovHC(model1)))

# stargazer(model1, title = 'Base Model',type = 'text',
# report = 'vcstp', omit.stat = 'f', se = list(se.model1,
# NULL), star.cutoffs = c(0.05,0.01,0.001))

paste("adj.r.square:", summary(model1)$adj.r.squared)
```

```
## [1] "adj.r.square: 0.656841444317101"
```

```r
coeftest(model1, vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.9722e-02  7.9057e-03  2.4946 0.0145210 *
## prbarr      -4.6441e-02  1.9565e-02 -2.3737 0.0198401 *
## density      7.5082e-03  1.1606e-03  6.4690  5.78e-09 ***
## polpc        5.0116e+00  4.2773e+00  1.1717 0.2445552
## pctmin80     3.1834e-04  8.4981e-05  3.7460 0.0003243 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

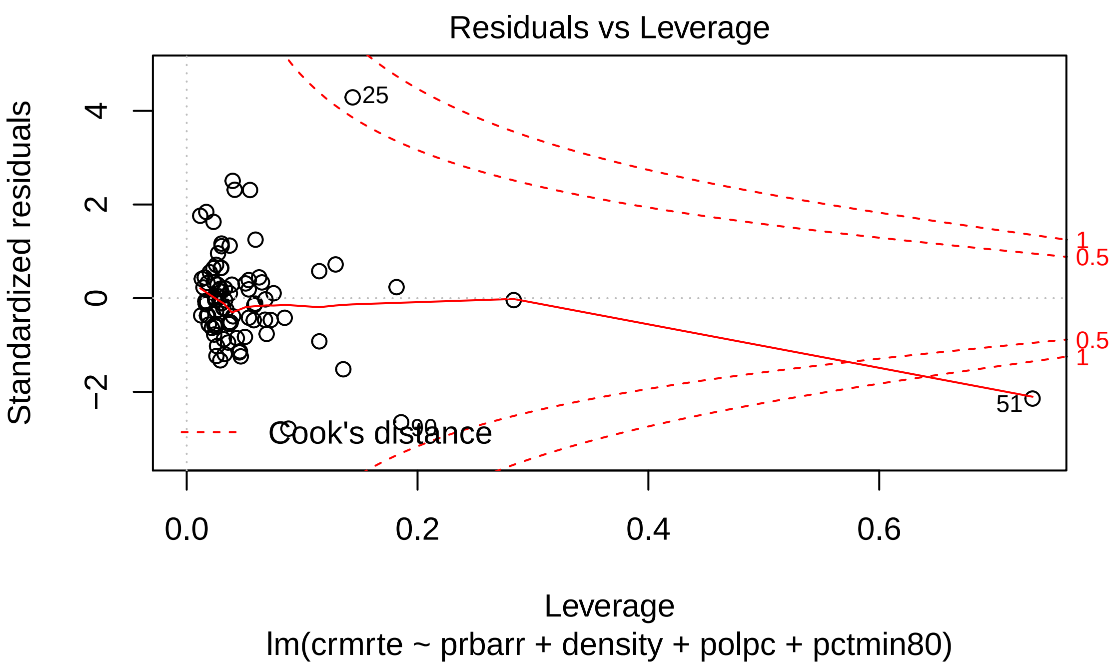
The adjusted r-squared of the model is relatively high at 0.66. This means that 66% of the variation in crime rate is explained by our input variables. Furthermore, the results of our initial model shows that the probability of arrest is statistically significant as a modulator of crime, while the density and minority percentage of each county are strongly statistically significant. The police per capita, on the other hand, is not. The slope coefficients tell us that for every 1 unit increase in `prbarr`, there is a corresponding 0.046 decrease in the crime rate. The model also suggests that by increasing the density of a county by 1 person per square mile, crime commited per person may rise by 0.008. Finally, for every percentage point increase of minorities in a county, crime commited per person may rise by 0.0003. The model also suggests that by increasing the police per capita by 1 will result in 5 additional crimes commited per person. However, this slope coefficeint is shown to be statistically insignificant.

To further assess the strength of our model, we can take a look at the residuals vs. leverage plot shown below. Here we can see that data point 51, has a Cook's distance greater than 1, meaning it has high influence over the model. As shown in section **2.3 TO DO** this data point has `polpc` and `prbarr` values multiple times higher than the next highest values for these variables. If this data point is not representative of the general population in North Carolina, then it may hurt the accuracy of our model. However, we investigated the other values of this county and could not justify removing this data point without further information.

Furthermore, a general rule is that if 1 % (or more) data points have standardized residuals > 2.5, the model contains too much error. If 5% (or more) of data points have residuals > 2, the model has too much error and represents our data poorly. In the residual vs. leverage plot below, we see that 7.7% of our data points have standardized residuals over 2. Therefore, our model has too much error and may represent our data poorly.

Because of this, we will now incorporate a few covariates that might increase the accuracy of our results.

```
plot(model1, which = 5)
```

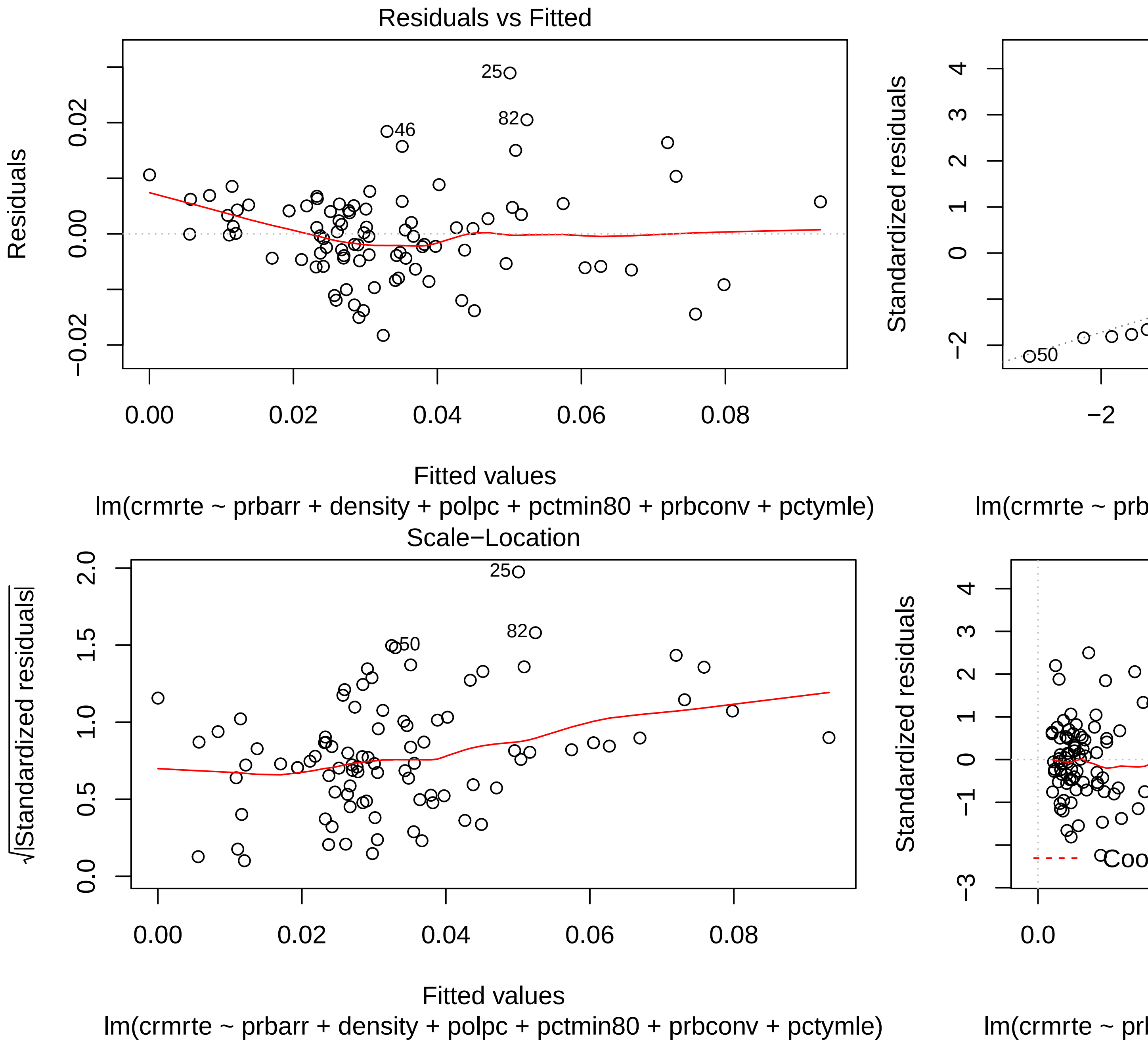


4.1 Regression Model: Second Model

Does this model include covariates meant to increase the accuracy of the regression? Has the team justified inclusion of each of these additional variables? Does the team identify what they want to measure with each coefficient? Does the team interpret the result of the regression in a thorough and convincing manner. Does the team evaluate all 6 CLM assumptions? Are the conclusions they draw based on this evaluation appropriate? Did the team interpret the results in terms of their research question?

One model that includes key explanatory variables and only covariates that you believe increase the accuracy of your results without introducing substantial bias (for example, you should not include outcome variables that will absorb some of the causal effect you are interested in). This model should strike a balance between accuracy and parsimony and reflect your best understanding of the determinants of crime.

```
# new: prbconv pctymle
model2 <- lm(crmrte ~ prbarr + density + polpc + pctmin80 + prbconv +
    pctymle, data = crime_df)
plot(model2)
```




```
# Breusch-Pagan 0.0004972
```

TO DO: Fill in discussion of CLM assumptions

4.2 Regression Third model

TO DO: add a lot of variables. Consider colinearity that we discovered in correlation matrix in modeling section.

4.3 Regression Table

TO DO: Be sure to convert SE's to robust before displaying.

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.

```
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
```

```
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.

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

summary(model_key)$r.squared
```

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

```r
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

```r
library(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
R^2 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.

TO BE SORTED LATER

TO BE SORTED LATER

TO BE SORTED LATER

TO BE SORTED LATER

Standardized Regression Model

TO DO: Eliminate. Save the comments on the diagnostic plots for use in the non-standardized model analysis.

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) 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)
```

TO DO: eliminate.

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')

#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)
```

```
subset(mar_r2_df, marginalr2 > .04)
```