majumder_mitchell_singh_lab3

November 27, 2018

Lab 3
W203 Statistics for Data Science
Author Names: Jake Mitchell, Rishikesh Majumder, Tej Singh
Section Number: Fall_Wed_06:30

In [2]:
```
# Load necessary libraries
library(ggplot2)
library(GGally)
library(ggcorrplot)
library(ggpubr)
library(car)
library(carData)
library(stats)
library(stargazer)
library(lmtest)
library(sandwich)

#library(dplyr)

options(repr.matrix.max.rows = 100)
options(repr.matrix.max.cols = 30)
options(repr.plot.width=10, repr.plot.height=10)

# Run Appendix - A if wishing to repeat results
```

Introduction
Team
We are a team from a leading analytical consulting firm on the East Coast. We specialize in empirical analysis of demographic data and provide a wide band of predictable outcomes which help in shaping the legislative agenda.
Agenda
As part of the next year's election campaign, we are tasked with analyzing historical crime data from various counties in North Carolina. The goal of this project is to predict the reason or a set of reasons behind the high crime rate. Once causal estimates are shown through statistical inferences, we will address the issues with possible policy changes.
Analytical Process Steps

1. Import and explore the data to get a feeling of data quality.

2. Transform data to remove or replace unexpected values.
3. Analyze relationships between different variables and choose statistically significant for the regression process.
4. Create multiple linear models and compare their robustness/effectiveness through the model summaries.
5. Detect omitted variable bias, and provide analysis on what effects the omitted variables have.
6. Propose a set of policy changes to the concerned authority that may help reduce crime rates.

0.0.1 EDA

Import the data and take a brief look at first few rows.

```
In [6]: #Import data
        data <- read.csv(file = 'crime_v2.csv')
        #Peek
        head(data)
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc
1	87	0.0356036	0.298270	0.527595997	0.436170	6.71	0.00182786	2.4226327	30.99368
3	87	0.0152532	0.132029	1.481480002	0.450000	6.35	0.00074588	1.0463320	26.89208
5	87	0.0129603	0.444444	0.267856985	0.600000	6.76	0.00123431	0.4127659	34.81605
7	87	0.0267532	0.364760	0.525424004	0.435484	7.14	0.00152994	0.4915572	42.94759
9	87	0.0106232	0.518219	0.476563007	0.442623	8.22	0.00086018	0.5469484	28.05474
11	87	0.0146067	0.524664	0.068376102	0.500000	13.00	0.00288203	0.6113361	35.22974

Get summary of data and spot anomalies

```
In [7]: summary(data)
```

```
     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
```

All the variables are numeric with different range and scale. For all the columns in the data frame, there are 6 observations without any values and the 'prbconv' contains a bad value, '''. This small number of anomalies should be transformed with appropriate adjusted values before the OLS regression.

A close look at the county and year variables show that they lack variability with observations and those may not contribute much to the analysis process.

Transformation Find out all the rows with 'NA' values.

```
In [8]: data[!complete.cases(data),]
```

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	centr
92	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA
93	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA
94	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA
95	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA
96	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA
97	NA	NA	NA	NA	'	NA	NA	NA	NA	NA	NA	NA

It seems 6 observations have NA values for all the variables. These rows could easily be removed as they are not useful in further analysis. In one of these rows there is also a tick mark " ' " that needs to be removed.

```
In [35]: #Deleting rows with NA values
         cdata <- data[complete.cases(data),]
         #Remove the non-numeric value
         cdata$prbconv <- as.numeric(gsub("[^0-9.]+", "", cdata$prbconv))
```

The probability of arrest and probability of conviction variables are actually ratios and there are a few values greater than one. We will be treating them as probabilities in this analysis, and so we will scale them as such. The ratios larger than one will be limited to 1.

```
In [37]: cdata[4:5] <- lapply(cdata[4:5], function(x) ifelse( x > 1, 1, x))
```

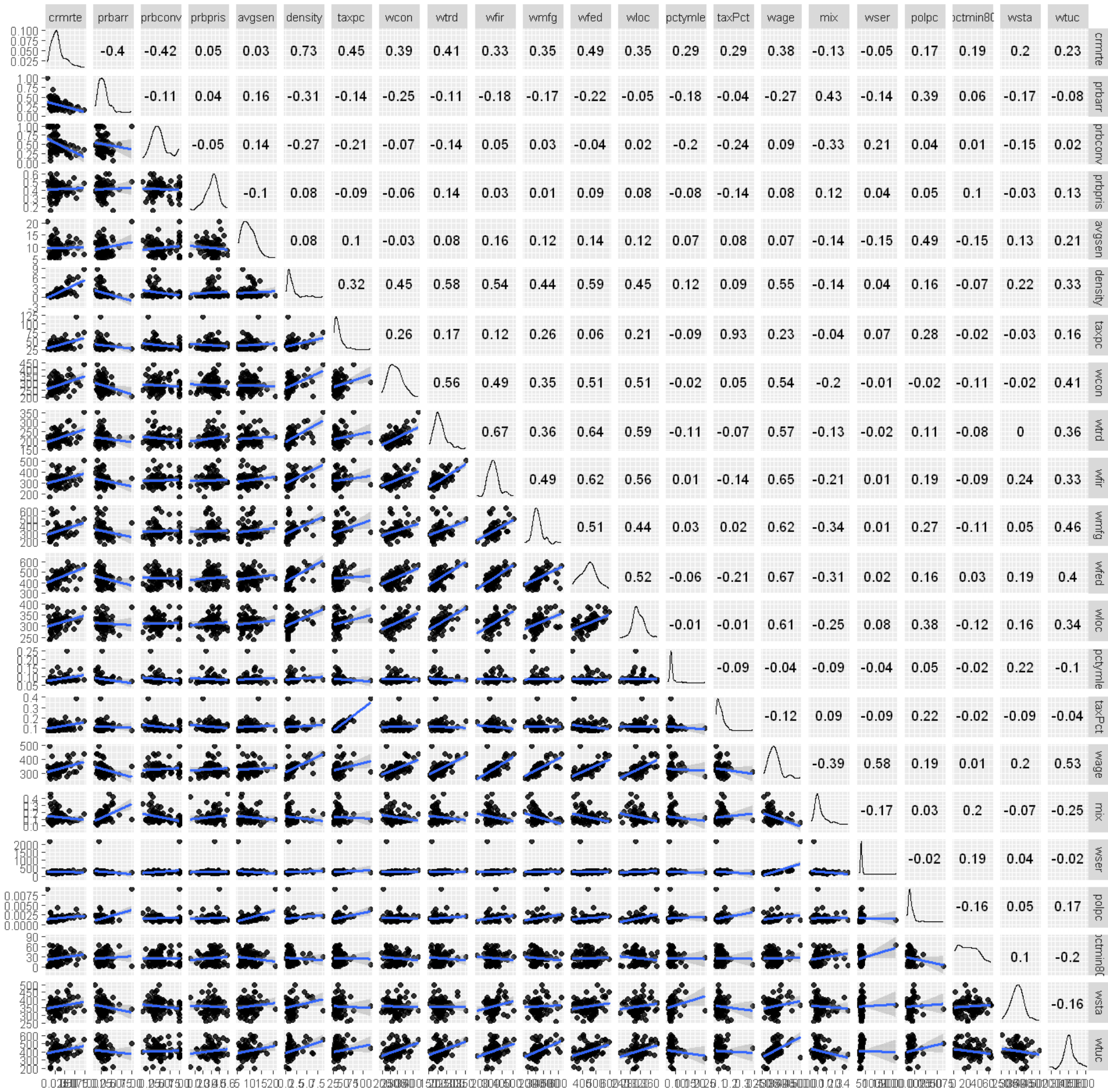
Feature Engineering All of the different wage amounts have a decent amount of colinearity because it is representative of how much people get paid. This means that using a single average wage variable will make broad analysis on the effect of wages on crime easier. In addition because the tax per capita is known, by divinding it by the average wage a tax percent can be calculated. This may be more useful to a politcal campaign because taxes are expressed in percents, not dollars per capita.

```
In [38]: #Average wage
         cdata$wage = (cdata$wcon + cdata$wtuc + cdata$wtrd + cdata$wfir + cdata$wser + cdata$
         #Overall tax percent
         cdata$taxPct = cdata$taxpc/cdata$wage
```

Distribution of Data Since the objective is the find out the of reason(s) behind higher crimes, the variable 'crmrte' should be the dependent variable. Dependent variables which are highly correlated with 'crmrte' should be useful to create a robust model.

```
In [42]: smoothing_method = "glm"
         options(repr.plot.width=12, repr.plot.height=12)
         ggscatmat(cdata[,c("crmrte", "prbarr", "prbconv", "prbpris", "avgsen", "density", "ta
                            "wtrd", "wfir", "wmfg", "wfed", "wloc", "pctymle", "taxPct", "wage
                            "wser", "polpc", "pctmin80", "wsta", "wtuc")], alpha=0.8) +
         geom_smooth(method=smoothing_method)
```

```
Warning message:
"Removed 23023 rows containing non-finite values (stat_smooth)."
```



Variable Selection The variable selection process is based on two consecutive criteria - practical and statistical significance. After taking notes on local law enforcement, IRS and other legislative authorities, we can narrow the list to fewer key variables: "prbarr", "prbconv", "density", "taxpc", "wage", "pctymle", "polpc", "pctmin80", "mix".

The chosen variables are generally part of three different domains: Certainty of punishment - probability of arrest and probability of conviciton; Demographic variations - density, percent young male, percent minorty; and Financial standing - taxes and wages

Other variables may not play a useful role in the analysis for the following reasons

- The 'county' and 'year' variables can be ignored as those do not vary with the crime rate.
- The average sentence days ('avgsen') are determined by a long process of court trials and federally standardized protocols. Criminal activities appear to be more heavily influenced by the probability of receiving a sentence rather than by the sentencing terms.

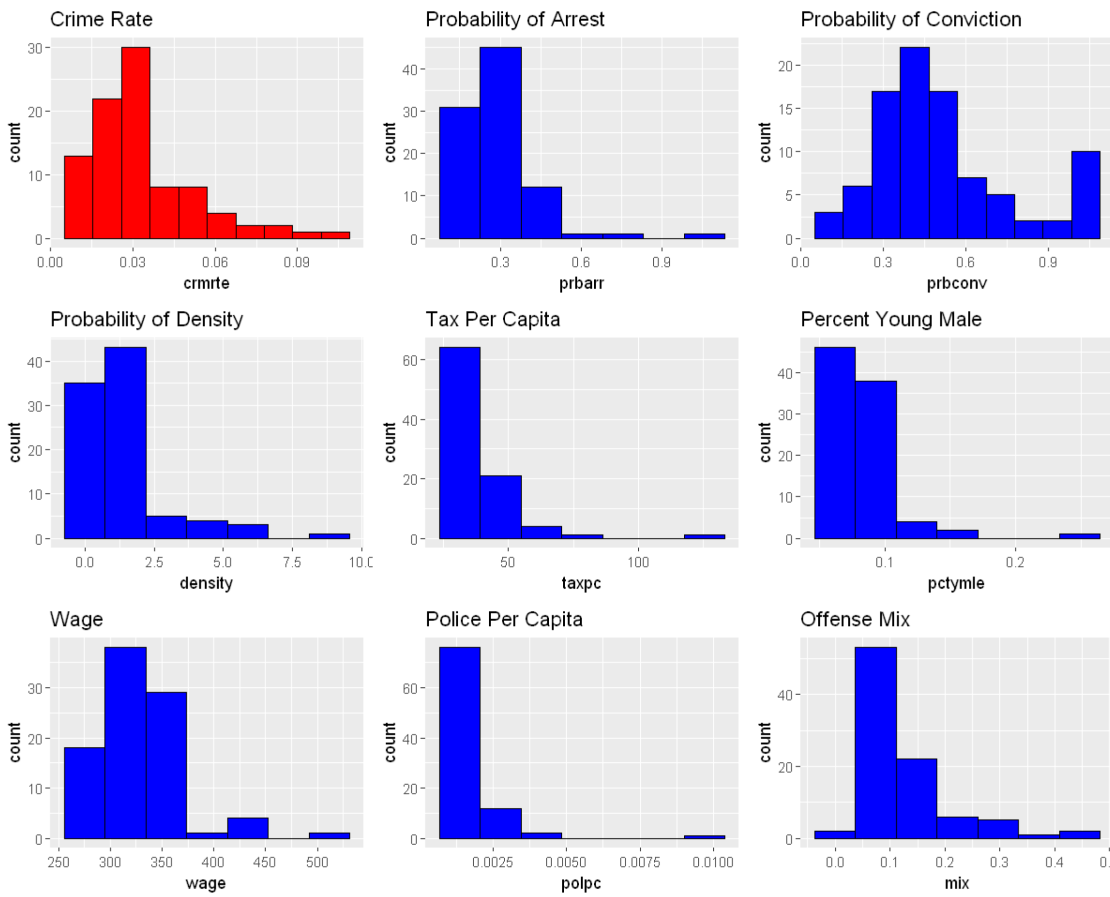
- County location alone may not be a good factor of crime rate data. Other variables like, police per capita ('polpc') and density should be closely related with urban and rural locations. So, any distinct information about density or police per capita will be diluted as density has statistical significance but likely contains the effects of both variables.
- As previously discussed, all the wage variables are consolidated into one variable, "wage", since individual use of those in a model may not contribute much. It is worth noting however that the only wages that political party has direct control over are the federal, state and minimum wages.

Statistical significance of these variables is analyzed as part of the model creation and EDA. Initially, independent variables will be chosen for the OLS regression, if they are significantly correlated with the dependent variable, crime rate.

Distribution

In [65]:
```
options(repr.plot.width=10, repr.plot.height=8)
c <- ggplot(data=cdata, aes(crmrte)) +
    geom_histogram(bins=10, fill="red", color="black") +
    ggtitle(paste("Crime Rate"))

ar <- ggplot(data=cdata, aes(prbarr)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Probability of Arrest"))

cn <- ggplot(data=cdata, aes(prbconv)) +
    geom_histogram(bins=10, fill="blue", color="black") +
    ggtitle(paste("Probability of Conviction"))

dn <- ggplot(data=cdata, aes(density)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Probability of Density"))

tp <- ggplot(data=cdata, aes(taxpc)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Tax Per Capita"))

pct <- ggplot(data=cdata, aes(pctymle)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Percent Young Male"))

wc <- ggplot(data=cdata, aes(wage)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Wage"))

plp <- ggplot(data=cdata, aes(polpc)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Police Per Capita"))
```

```
mx <- ggplot(data=cdata, aes(mix)) +
    geom_histogram(bins=7, fill="blue", color="black") +
    ggtitle(paste("Offense Mix"))

ggarrange(c, ar, cn, dn, tp, pct, wc, plp, mx,
          ncol = 3, nrow = 3)
```



Although few of the distributions are skewed, taking log values did not improve any of the distributions.

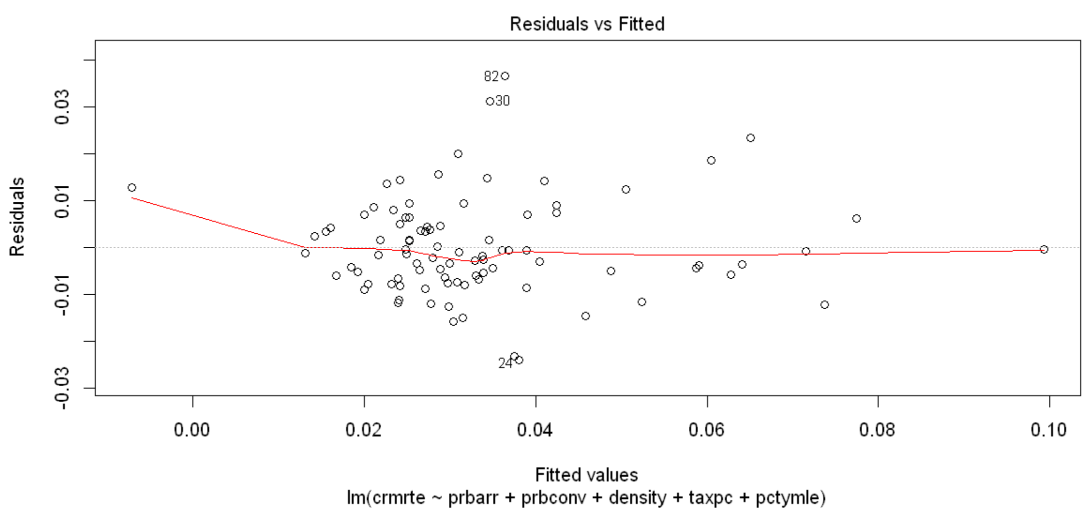
From Appendix - B, it is clear that applying different forms of data transformations (natural log, square, square root) did not make correlational improvements with the crime rate variable.

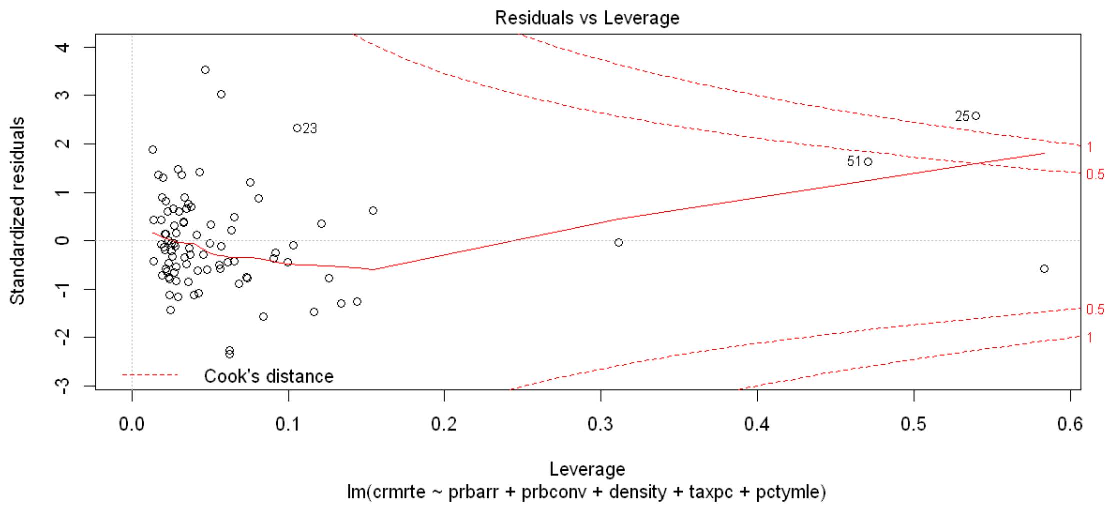
0.0.2 Model Creation

Model 1 - Key Interest Variables First we will make a model based solely on the key variables.

```
In [55]: options(repr.plot.width=10, repr.plot.height=5)
         model1 <- lm(crmrte ~  prbarr + prbconv + density + taxpc  + pctymle, data = cdata)
         summary(model1)
         plot(model1, which = 1)
         plot(model1, which = 5)
```

```
Call:
lm(formula = crmrte ~ prbarr + prbconv + density + taxpc + pctymle,
    data = cdata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.023978 -0.006244 -0.001428  0.005540  0.036508

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.769e-02  8.600e-03   2.057 0.042766 *
prbarr      -2.893e-02  9.386e-03  -3.083 0.002767 **
prbconv     -1.758e-02  5.187e-03  -3.389 0.001065 **
density      6.388e-03  8.437e-04   7.571 4.09e-11 ***
taxpc        3.271e-04  9.271e-05   3.528 0.000678 ***
pctymle      1.380e-01  5.076e-02   2.718 0.007948 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01059 on 85 degrees of freedom
Multiple R-squared:  0.7005,Adjusted R-squared:  0.6829
F-statistic: 39.76 on 5 and 85 DF,  p-value: < 2.2e-16
```






```
In [56]: paste("AIC Score: ", AIC(model1))
         paste("Covariation of coefficients - ")
         diag(vcov(model1))
         paste("Mean residuals: ", mean(model1$residuals))
```

'AIC Score: -561.616468314666'

'Covariation of coefficients - '

(Intercept) 7.39546217619811e-05 **prbarr** 8.80933663240813e-05 **prbconv** 2.6905246669517e-05 **density** 7.11869781993171e-07 **taxpc** 8.59520626986017e-09 **pctymle** 0.00257643082667928

'Mean residuals: -7.14905115536995e-20'

Model 1 - Interpretation Statistical Figures

1. Low Residuals Median: -0.002907 Mean: 2.85166207566554e-19
2. Low Coefficients and low variation of coefficients
3. Low RSE: 0.01065
4. Significantly high R-squared/Adjusted R-squared values - 0.7007, 0.6793
5. Highest AIC score: -559.671140806715
6. One value more than 1 cook's distance. A few close to $\frac{1}{2}$ cook's distance.

Quality and Measurement of OLS Assumptions

- From the Fitted and Residual Plot, the spline curve shows a good alignment with the fitted line. It proves a linear relationship between dependent and independent variables. In addition, the plot does show biases as the data points are not random.
- The model efficiency is high since the coefficient variations or robust standard error values are low. It also means the estimators are consistent around the regression line.
- All variables are highly statistically significant (< 0.01).

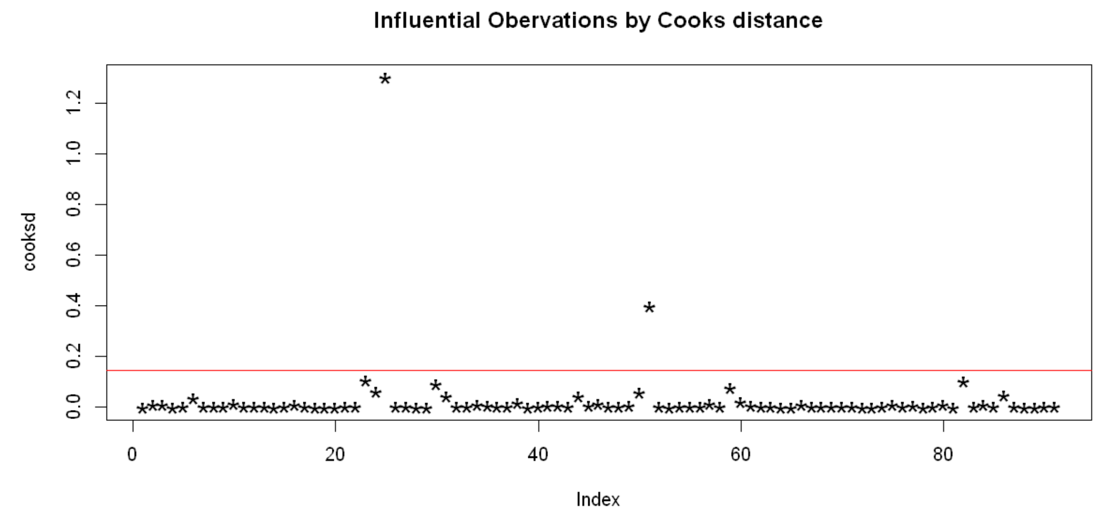
- High Adjusted R-squared value implies goodness of fit, although it may be inflated due to the number of inputs.
- Two observations can be considered as outliers and those may influence the estimation (25, 51).

Model 2 - Key Variables + Covariates Adjustments after Model1 results - 1. Removing two outliers 2. Removing statistically insignificant variable, wage. 3. Introducing new variables which have practical significance: -- Wage, Police Per Capita, and Mix

Note - These variables have a weak correlation with other independent variables.

Removing outliers

```
In [70]: #Cooks distance measurement
         options(repr.plot.width=10, repr.plot.height=5)
         cooksd <- cooks.distance(model1)
         plot(cooksd, pch="*", cex=2, main="Influential Obervations by Cooks distance")
         abline(h = 5*mean(cooksd, na.rm=T), col="red")
```




```
In [71]: #Influential Outliers
         influential <- as.numeric(names(cooksd)[(cooksd > 5*mean(cooksd, na.rm=T))])
         cooksd[influential]
```

25 1.30107603445457 **51** 0.399487570962988

```
In [72]: #Remove outliers
         ctdata <- cdata[(cdata$county != cdata[influential[1],]$county) & (cdata$county != cda
```

```
In [74]: #Create model2
         options(repr.plot.width=10, repr.plot.height=5)
```

```r
model2 <- lm(crmrte ~ prbarr + prbconv + density + taxpc + wage + pctymle + polpc + pc
summary(model2)
plot(model2, which = 1)
plot(model2, which = 5)
```

```
Call:
lm(formula = crmrte ~ prbarr + prbconv + density + taxpc + wage +
    pctymle + polpc + pctmin80 + mix, data = ctdata)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0173362 -0.0043928 -0.0000093  0.0050776  0.0217707

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.799e-02  1.221e-02   3.110   0.0026 **
prbarr      -4.586e-02  1.096e-02  -4.183 7.40e-05 ***
prbconv     -2.319e-02  4.709e-03  -4.926 4.53e-06 ***
density      7.203e-03  8.592e-04   8.384 1.56e-12 ***
taxpc       -4.211e-05  1.101e-04  -0.382   0.7031
wage        -2.678e-05  2.988e-05  -0.896   0.3730
pctymle      8.137e-02  4.177e-02   1.948   0.0550 .
polpc        3.751e+00  1.941e+00   1.932   0.0569 .
pctmin80     3.700e-04  5.484e-05   6.748 2.26e-09 ***
mix         -1.863e-02  1.458e-02  -1.277   0.2052
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008245 on 79 degrees of freedom
Multiple R-squared:  0.8147,Adjusted R-squared:  0.7936
F-statistic:  38.6 on 9 and 79 DF,  p-value: < 2.2e-16
```


```
In [75]: paste("AIC Score: ", AIC(model2))
         paste("Covariation of coefficients - ")
         diag(vcov(model2))
         paste("Mean residuals: ", mean(model2$residuals))
```

'AIC Score: -590.104195478824'

'Covariation of coefficients - '

(Intercept) 0.000149165817174937 **prbarr** 0.000120209075253182 **prbconv** 2.21700965083303e-05 **density** 7.38213807907202e-07 **taxpc** 1.21225355679025e-08 **wage**

8.92971991302478e-10 **pctymle** 0.00174455627190874 **polpc** 3.76812986297666 **pctmin80** 3.00750515242381e-09 **mix** 0.00021263931365299

'Mean residuals: -3.08679268011339e-19'

Model 2 - Interpretation Statistical Figures

1. Low Residuals Median: -0.0002695 Mean: -1.68519108301068e-19
2. Mostly low coefficients and low variation of coeffcients (except 'polpc')
3. Low RSE: 0.008235
4. Significantly high R-squared/Adjusted R-squared values - 0.8128, 0.7941.

5. Lower AIC score: -591.204239814494
6. No outliers per Cook's distance

Quality and Measurement of OLS Assumptions

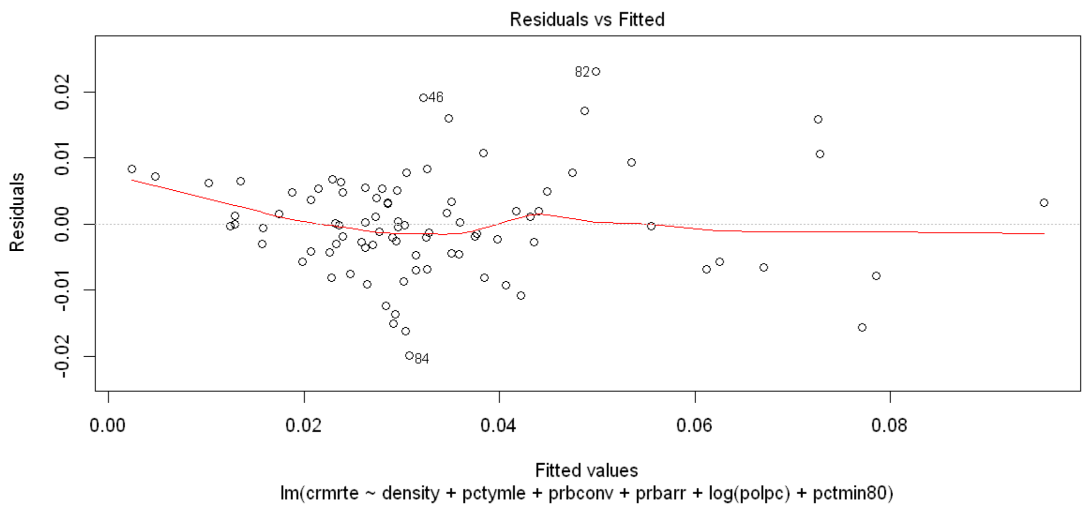
- From the Fitted and Residual Plot, the spline curve shows similar behavior as the Model1 and it still shows a sign of biases. However, the data points have become more randomly distributed.
- The model efficiency is relatively high since the coefficient variations or robust standard error values are low. It also means the estimators are consistent around the regression line.
- 'taxpc', 'polpc' and 'mix' failed to project strong statistical significance.
- High Adjusted R-squared value implies goodness of fit, however it is likely due to the additional explanatory variables.
- No influencial outliers.

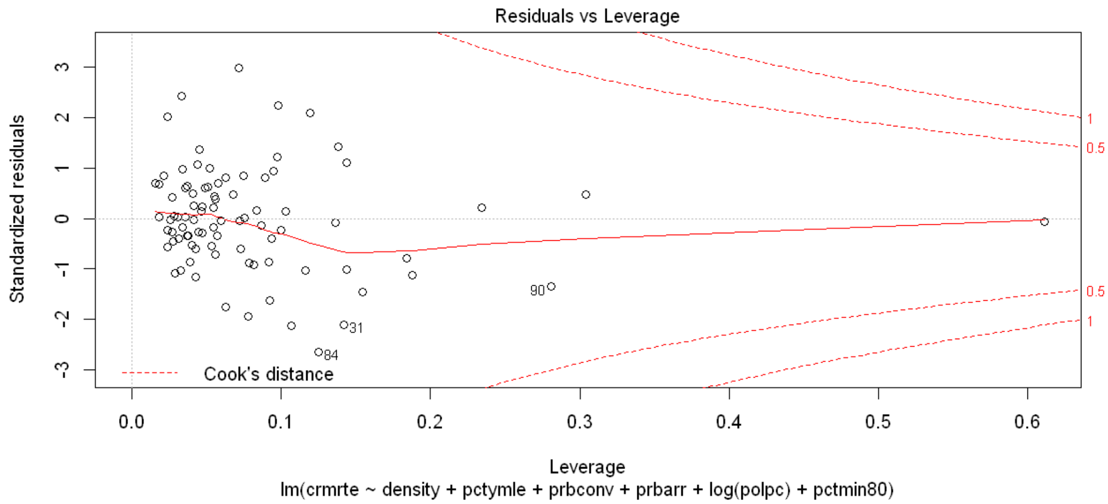
Model 3 Adjustments after Model2 results - 1. Removing statistically insignificant variables, 'mix', 'taxpc'. 2. Since 'polpc' has potential practical significance and has a skewed distribution, taking log value and keeping as an independent variable.

```
In [76]: options(repr.plot.width=10, repr.plot.height=5)
         model3 <- lm(crmrte ~ density + pctymle + prbconv + prbarr + log(polpc) + pctmin80, da
         summary(model3)
         plot(model3, which = 1)
         plot(model3, which = 5)
```

```
Call:
lm(formula = crmrte ~ density + pctymle + prbconv + prbarr +
    log(polpc) + pctmin80, data = ctdata)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0198864 -0.0044373 -0.0002281  0.0047602  0.0230956

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.115e-02  2.260e-02   3.591 0.000561 ***
```

```
density      6.626e-03  7.129e-04   9.296 1.86e-14 ***
pctymle      8.906e-02  3.925e-02   2.269 0.025881 *
prbconv     -2.143e-02  4.470e-03  -4.794 7.20e-06 ***
prbarr      -4.986e-02  9.790e-03  -5.093 2.21e-06 ***
log(polpc)   7.554e-03  3.208e-03   2.354 0.020945 *
pctmin80     3.514e-04  5.192e-05   6.768 1.80e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008052 on 82 degrees of freedom
Multiple R-squared:  0.8166,Adjusted R-squared:  0.8032
F-statistic: 60.85 on 6 and 82 DF,  p-value: < 2.2e-16
```






```
In [39]: paste("AIC Score: ", AIC(model3))
         paste("Covariation of coefficients - ")
         diag(vcov(model3))
         paste("Mean residuals: ", mean(model3$residuals))
```

'AIC Score: -597.014127577175'

'Covariation of coefficients - '

(Intercept) 0.000510863610100357 **density** 5.08156688819428e-07 **pctymle** 0.00154020668920739 **prbconv** 1.99783790553895e-05 **prbarr** 9.58367417214052e-05 **log(polpc)** 1.02929769182186e-05 **pctmin80** 2.69614492426135e-09

'Mean residuals: -2.40142168117705e-19'

Model 3 - Interpretation Statistical Figures

1. Very Low Residuals Median: -0.0002281 Mean: -2.40142168117705e-19
2. Low coefficients and low variation of coefficients
3. Low RSE: 0.008052
4. Significantly high R-squared/Adjusted R-squared values - 0.8166, 0.8032

5. Low AIC score: -597.014127577175
6. No outliers

Quality and Measurement of OLS Assumptions

- From the Fitted and Residual Plot, the spline curve shows similar behavior as the other models. The initial curve shape should be because of less number of data points with low crime rate values. After the initial few data points, the spline curve ,in fact, becomes almost aligned with the fitted line with **random data points**. In this case, the OLS assumption **zero conditional mean** and **exogeneity** have been met.

- **Proven Multicollinearity Assumption** - OLS estimators can produce a strong linear regression model if the estimators are unbiased and not correlated to each other. The dependent variables influence each other and that positive or negative bias is captured by the error term(s). Depending on the direction of influence, if the degree of correlation is high the estimation will be overrated or underrated. The correlation numbers with the variables and residuals (APPENDIX B) proves that Multicollinearity assumption is still protected.
- The model efficiency is relatively high since the coefficient variations or robust standard error values are still very low. It also means the estimators are consistent around the regression line.
- All the independent variables have strong statistical significance.
- High Adjusted R-squared value implies goodness of fit.
- No influncial outliers.

Causal Estimation The model3 shows strong alignment with required OLS assumptions. With Zero conditional mean, exogeneity and very low residuals, the model3 seems to showing a good indication for causal estimation. However, the apparent causation could well be a lot weaker for the following reasons: - With more data, the model robustness could easily go down. Currently, a relatively low number of observations (~90) are used to create the models. - Biased sampling methods and poor data collection strategies could have a negative impact on the estimation. - In a practical setting, it is very hard to make sure models contain all of the data points relevant to the thing that we are attempting to model. This leads to most models having some level of omitted variable bias.

Finally, even if the causal estimation is hard to come by for this analysis, we can confidently conclude this is an associative model.

Comparing Models

```
In [77]: se.model1 = sqrt(diag(vcovHC(model1)))
         se.model2 = sqrt(diag(vcovHC(model2)))
         se.model3 = sqrt(diag(vcovHC(model3)))

         stargazer(model1, model2, model3, type = "text", omit.stat = "f",
                   se = list(se.model1, se.model2),
                   star.cutoffs = c(0.05, 0.01, 0.001))
```

```
============================================================
                            Dependent variable:
                  ------------------------------------------
                                 crmrte
                      (1)           (2)           (3)
------------------------------------------------------------
prbarr              -0.029      -0.046***     -0.050***
                    (0.016)      (0.011)       (0.010)

log(polpc)                                     0.008*
                                               (0.003)
```

prbconv	-0.018*	-0.023***	-0.021***
	(0.007)	(0.007)	(0.004)
density	0.006***	0.007***	0.007***
	(0.001)	(0.002)	(0.001)
taxpc	0.0003	-0.00004	
	(0.0003)	(0.0002)	
wage		-0.00003	
		(0.0001)	
pctymle	0.138*	0.081	0.089*
	(0.061)	(0.042)	(0.039)
polpc		3.751	
		(2.959)	
pctmin80		0.0004***	0.0004***
		(0.0001)	(0.0001)
mix		-0.019	
		(0.017)	
Constant	0.018	0.038	0.081***
	(0.015)	(0.024)	(0.023)
Observations	91	89	89
R2	0.701	0.815	0.817
Adjusted R2	0.683	0.794	0.803
Residual Std. Error	0.011 (df = 85)	0.008 (df = 79)	0.008 (df = 82)
Note:			*p<0.05; **p<0.01; ***p<0.001

From the model interpretation and this comparison table, it is obvious that the model3 has a lower standard error and higher Adjusted R-squared values. In addition, from the residual and fitted plot, the model3 shows good efficiency and consistency.

0.0.3 Proposed Policy Changes

Since the probability of arrest, probability of conviction, minority percentage, and density of a county have a major contribution for the model3, there are a few suggestions that may help lower the crime rate.

- A legislative push for a lower tolerance level for criminal activities should have a major impact on a number of criminal cases.

- More automated surveillance activities in densely populated areas may be necessary to reduce crime rates.
- A series of public circulations on stricter policies which is targetted towards specific demographics should have a greater impact.
  - The strong effect of the minority percentage in a county may be a cumulation of several factors that are disproportionately affecting minorities. This is something that warrants further research.

0.0.4 Appendix - A

Helper functions

```
In [3]: getLogTran <- function(df, cols, tranType, tranTypeDep){
            corList <- list()

            for(coln in cols){

                if(tranType == ''){
                  corList <- append(corList, cor(df[[coln]], tranVar(df$crmrte, tranTypeDep)))

                }else{

                  newColN <- paste(coln, tranType, sep="_")
                  df[[newColN]] <- tranVar(df[[coln]], tranType)
                  corList <- append(corList, cor(df[[newColN]], tranVar(df$crmrte, tranTypeDep
              }

            }

            return (corList)
        }

        tranVar <- function(var1, tranType){
            if(tranType == 'lg'){
                return (log(var1))
            } else if(tranType == 'sq'){
                return (var1^2)
            } else if(tranType == 'sqt'){
                return (sqrt(var1))
            } else {
                return (var1)
            }
        }
```

0.0.5 Appendix - B

Compare Corr for Different Transformation with Crime Rate

```
In [47]: #tdata <- data.frame(matrix(ncol = 12, nrow = 0))
         cols <- c("crmrte", "prbarr", "prbconv", "density", "taxpc", "wage", "pctymle")
         #colnames(tdata) <- cols

         #for(ops in c('', 'lg', 'sq', 'sqt')){
         print('------------------Corr with Crime Rate. No trans of crmrte--------------------

         list_data <- list(cols, getLogTran(cdata, cols, 'lg', ''),
                           getLogTran(cdata, cols, 'sq', ''), getLogTran(cdata, cols, 'sqt', '

         tdata  <-  as.data.frame(matrix(unlist(list_data), nrow=length(unlist(list_data[1])))
         colnames(tdata) <- c("var", "lg", "sq", "sqt", "none")
         tdata

         print('------------------Corr with Crime Rate. Log trans of crmrte-------------------

         list_data <- list(cols, getLogTran(cdata, cols, 'lg', 'lg'),
                           getLogTran(cdata, cols, 'sq', 'lg'),
                           getLogTran(cdata, cols, 'sqt', 'lg'),
                           getLogTran(cdata, cols, '', 'lg'))

         tdata  <-  as.data.frame(matrix(unlist(list_data), nrow=length(unlist(list_data[1])))
         colnames(tdata) <- c("var", "lg", "sq", "sqt", "none")
         tdata

         print('------------------Corr with Crime Rate. Square Trans of crmrte----------------


         list_data <- list(cols, getLogTran(cdata, cols, 'lg', 'sq'),
                           getLogTran(cdata, cols, 'sq', 'sq'),
                           getLogTran(cdata, cols, 'sqt', 'sq'),
                           getLogTran(cdata, cols, '', 'sq'))

         tdata  <-  as.data.frame(matrix(unlist(list_data), nrow=length(unlist(list_data[1])))
         colnames(tdata) <- c("var", "lg", "sq", "sqt", "none")
         tdata

         print('------------------Corr with Crime Rate. SQRT Trans of crmrte------------------


         list_data <- list(cols, getLogTran(cdata, cols, 'lg', 'sqt'),
                           getLogTran(cdata, cols, 'sq', 'sqt'),
                           getLogTran(cdata, cols, 'sqt', 'sqt'),
                           getLogTran(cdata, cols, '', 'sqt'))

         tdata  <-  as.data.frame(matrix(unlist(list_data), nrow=length(unlist(list_data[1])))
         colnames(tdata) <- c("var", "lg", "sq", "sqt", "none")
         tdata
```

[1] "------------------Corr with Crime Rate. No trans of crmrte---------------------"

var	lg	sq	sqt	none
crmrte	0.94155941652815	0.963314447688739	0.986585485007721	1
prbarr	-0.419670932952259	-0.33750871021838	-0.414702800613676	-0.398879118076819
prbconv	-0.35418173866631	-0.417079154924082	-0.398569393501294	-0.417302576248551
density	0.477607761812432	0.660066274500242	0.733462985063669	0.728963158061984
taxpc	0.415564519823015	0.44182777372935	0.437572604049532	0.4509797818509
wage	0.386374319390269	0.359104947366975	0.382066158175715	0.376101960332975
pctymle	0.324972570053244	0.245450825730505	0.310144690951903	0.291248491056166

[1] "------------------Corr with Crime Rate. Log trans of crmrte---------------------"

var	lg	sq	sqt	none
crmrte	1	0.82581808440035	0.983402742416995	0.94155941652815
prbarr	-0.431410316712002	-0.459855111060795	-0.454101251802571	-0.468403863932953
prbconv	-0.341139648543617	-0.472928942629082	-0.406452145832728	-0.444316373057585
density	0.493596527034902	0.520974596124295	0.68110686628769	0.633650441803765
taxpc	0.341856418221085	0.344572481500381	0.354361129677032	0.360050785795018
wage	0.332769071484339	0.283661814996677	0.322794556355104	0.311251670277
pctymle	0.31242104733822	0.235794389929685	0.297335385260795	0.27888339420391

[1] "------------------Corr with Crime Rate. Square Trans of crmrte---------------------"

var	lg	sq	sqt	none
crmrte	0.82581808440035	1	0.908891866329352	0.963314447688739
prbarr	-0.393571355383632	-0.263180713034383	-0.374534525713388	-0.344704537711208
prbconv	-0.334557367523137	-0.347166049638351	-0.361399267209415	-0.366043379028041
density	0.425754280086151	0.739296908890496	0.717924635162036	0.753822678425409
taxpc	0.453149701355062	0.494619036717848	0.48110566503104	0.499453249054686
wage	0.388999478908761	0.381724124434679	0.389602617479143	0.388648930815095
pctymle	0.294385229646093	0.217595054143318	0.280503978888829	0.262378460036068

[1] "------------------Corr with Crime Rate. SQRT Trans of crmrte---------------------"

var	lg	sq	sqt	none
crmrte	0.983402742416995	0.908891866329352	1	0.986585485007721
prbarr	-0.427248888625485	-0.391120619995217	-0.434215188985925	-0.431107672375324
prbconv	-0.352044843027956	-0.44898794245525	-0.407197939732133	-0.43541429509193
density	0.493488003542832	0.597989175801448	0.718591754803244	0.691747794431379
taxpc	0.382756178938766	0.398040449985126	0.400385015589278	0.410210182611733
wage	0.368151444712786	0.329289255681376	0.36089293087517	0.351987512269128
pctymle	0.325431491130904	0.246550698458356	0.310362612665254	0.291511539640238

0.0.6 Appindex - B

In [41]:
```
cov(model3$residuals,ctdata$density)
cov(model3$residuals,ctdata$pctymle)
cov(model3$residuals,ctdata$prbconv)
cov(model3$residuals,ctdata$prbarr)
cov(model3$residuals,log(ctdata$polpc))
cov(model3$residuals,ctdata$pctmin80)
```

-7.14148637794782e-20
-1.75203433584513e-21
-6.8433404058887e-20
6.13546931696082e-20
3.07620949984721e-20
8.04640873319103e-18