predicting-crime-final-project

Jose Cruz

12/11/2021

Final Project:

- In this report there will be 2 parts:
- the first part will describe the process through the code and let you see everything step by step with explanation of the process and reasons.
- the second part will discuss the final model chosenand how each variable influences the response and why we choosed them and anything else we did.

The Code

```
# look at our initial dataset

head(crimeDataset)

##   X county year    crmrte   prbarr  prbconv  prbpris avgsen     polpc
density
## 1 1      1   81 0.0398849 0.289696 0.402062 0.472222   5.61 0.0017868
2.307159
## 2 2      1   82 0.0383449 0.338111 0.433005 0.506993   5.59 0.0017666
2.330254
## 3 3      1   83 0.0303048 0.330449 0.525703 0.479705   5.80 0.0018358
2.341801
## 4 4      1   84 0.0347259 0.362525 0.604706 0.520104   6.89 0.0018859
2.346420
## 5 5      1   85 0.0365730 0.325395 0.578723 0.497059   6.55 0.0019244
2.364896
## 6 6      1   86 0.0347524 0.326062 0.512324 0.439863   6.90 0.0018952
2.385681
##      taxpc  pctmin     wcon      wtuc     wtrd     wfir     wser   wmfg
wfed
## 1 25.69763 20.2187 206.4803  333.6209 182.3330 272.4492 215.7335 229.12
409.37
## 2 24.87425 20.2187 212.7542  369.2964 189.5414 300.8788 231.5767 240.33
419.70
## 3 26.45144 20.2187 219.7802 1394.8030 196.6395 309.9696 240.1568 269.70
438.85
## 4 26.84235 20.2187 223.4238  398.8604 200.5629 350.0863 252.4477 281.74
459.17
## 5 28.14034 20.2187 243.7562  358.7830 206.8827 383.0707 261.0861 298.88
490.43
## 6 29.74098 20.2187 257.9139  369.5465 218.5165 409.8842 269.6129 322.65
```

```
478.67
##      wsta   wloc       mix   pctymle
## 1 236.24 231.47 0.0999179 0.0876968
## 2 253.88 236.79 0.1030491 0.0863767
## 3 250.36 248.58 0.0806787 0.0850909
## 4 261.93 264.38 0.0785035 0.0838333
## 5 281.44 288.58 0.0932486 0.0823065
## 6 286.91 306.70 0.0973228 0.0800806

#checks for nulls

is.null(crimeDataset)

## [1] FALSE
```

- Looked at the data that we have and checked the head of our data set and checked if we had any null values and found none

```
# names of variables
names(crimeDataset)

##  [1] "X"        "county"   "year"     "crmrte"   "prbarr"   "prbconv"  "prbpris"
##  [8] "avgsen"   "polpc"    "density"  "taxpc"    "pctmin"   "wcon"     "wtuc"
## [15] "wtrd"     "wfir"     "wser"     "wmfg"     "wfed"     "wsta"     "wloc"
## [22] "mix"      "pctymle"

dim(crimeDataset)

## [1] 630  23

#drop x column from the dataset

crimeDataset=subset(crimeDataset,select = -c(X))
```

- Here we looked at the names and dimensions of our data set and found a redundant variable called x and removed it from our data set.

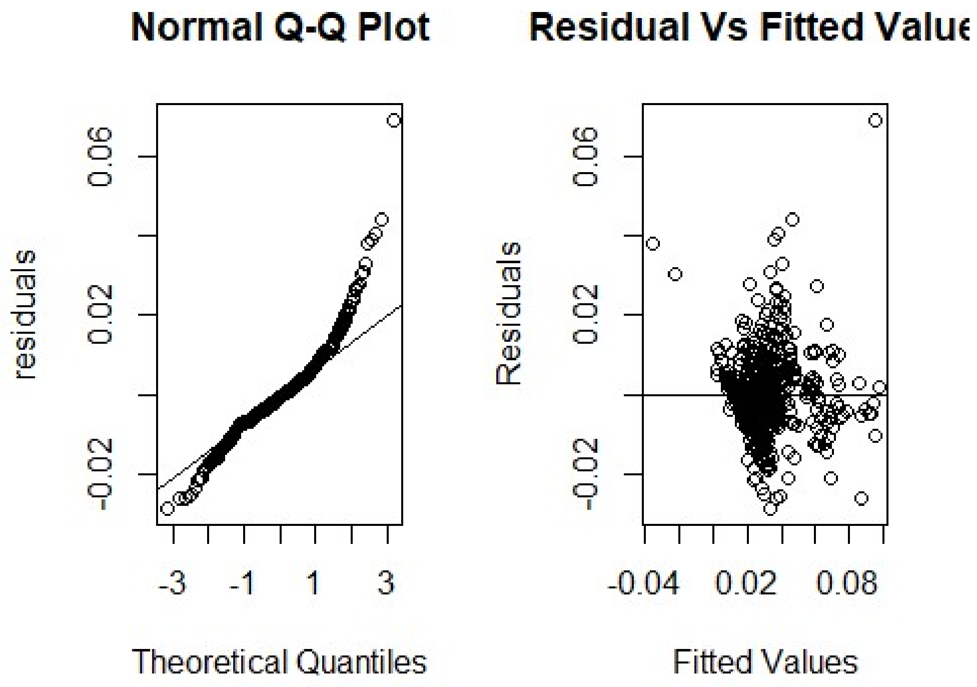
```
summary(crimeDataset)

##      county           year        crmrte             prbarr
##  Min.   :  1.0   Min.   :81   Min.   :0.001812   Min.   :0.05882
##  1st Qu.: 51.0   1st Qu.:82   1st Qu.:0.018352   1st Qu.:0.21790
##  Median :103.0   Median :84   Median :0.028441   Median :0.27824
##  Mean   :100.6   Mean   :84   Mean   :0.031588   Mean   :0.30737
##  3rd Qu.:151.0   3rd Qu.:86   3rd Qu.:0.038406   3rd Qu.:0.35252
##  Max.   :197.0   Max.   :87   Max.   :0.163835   Max.   :2.75000
##     prbconv            prbpris           avgsen           polpc
##  Min.   : 0.06838   Min.   :0.1489   Min.   : 4.220   Min.   :0.0004585
##  1st Qu.: 0.34769   1st Qu.:0.3744   1st Qu.: 7.160   1st Qu.:0.0011913
##  Median : 0.47437   Median :0.4286   Median : 8.495   Median :0.0014506
##  Mean   : 0.68862   Mean   :0.4255   Mean   : 8.955   Mean   :0.0019168
##  3rd Qu.: 0.63560   3rd Qu.:0.4832   3rd Qu.:10.197   3rd Qu.:0.0018033
##  Max.   :37.00000   Max.   :0.6786   Max.   :25.830   Max.   :0.0355781
##     density           taxpc            pctmin            wcon
##  Min.   :0.1977   Min.   : 14.30   Min.   : 1.284   Min.   :  65.62
##  1st Qu.:0.5329   1st Qu.: 23.43   1st Qu.:10.005   1st Qu.: 201.66
##  Median :0.9526   Median : 27.79   Median :24.852   Median : 236.46
##  Mean   :1.3861   Mean   : 30.24   Mean   :25.713   Mean   : 245.67
##  3rd Qu.:1.5078   3rd Qu.: 33.27   3rd Qu.:38.223   3rd Qu.: 269.69
##  Max.   :8.8277   Max.   :119.76   Max.   :64.348   Max.   :2324.60
##       wtuc             wtrd              wfir               wser
##  Min.   :  28.86   Min.   :  16.87   Min.   :  3.516   Min.   :   1.844
##  1st Qu.: 317.60   1st Qu.: 168.05   1st Qu.:235.705   1st Qu.: 191.319
##  Median : 358.20   Median : 185.48   Median :264.423   Median : 216.475
##  Mean   : 406.10   Mean   : 192.82   Mean   :272.059   Mean   : 224.671
##  3rd Qu.: 411.02   3rd Qu.: 204.82   3rd Qu.:302.440   3rd Qu.: 247.155
##  Max.   :3041.96   Max.   :2242.75   Max.   :509.466   Max.   :2177.068
##       wmfg            wfed            wsta            wloc
##  Min.   :101.8   Min.   :255.4   Min.   :173.0   Min.   :163.6
##  1st Qu.:234.0   1st Qu.:361.5   1st Qu.:258.2   1st Qu.:226.8
##  Median :271.6   Median :404.0   Median :289.4   Median :253.1
##  Mean   :285.2   Mean   :403.9   Mean   :296.9   Mean   :258.0
##  3rd Qu.:320.0   3rd Qu.:444.6   3rd Qu.:331.5   3rd Qu.:289.3
##  Max.   :646.9   Max.   :598.0   Max.   :548.0   Max.   :388.1
##       mix              pctymle
##  Min.   :0.002457   Min.   :0.06216
##  1st Qu.:0.075324   1st Qu.:0.07859
##  Median :0.102089   Median :0.08316
##  Mean   :0.139396   Mean   :0.08897
##  3rd Qu.:0.149009   3rd Qu.:0.08919
##  Max.   :4.000000   Max.   :0.27436
```

- Here we looked at our variables stats to check for any abnormal metrics. As far as I can see the data look good.

```
#create models

crimeModelOne=lm(crmrte~.,data = crimeDataset)

summary(crimeModelOne)

## 
## Call:
## lm(formula = crmrte ~ ., data = crimeDataset)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.028669 -0.005226 -0.000813  0.004056  0.069055 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8.764e-02  2.985e-02   2.936 0.003448 ** 
## county       3.562e-06  6.991e-06   0.509 0.610625    
## year        -1.154e-03  4.024e-04  -2.869 0.004261 ** 
## prbarr      -3.111e-02  2.796e-03 -11.126  < 2e-16 ***
## prbconv     -2.365e-03  3.077e-04  -7.686 6.13e-14 ***
## prbpris      1.299e-03  4.791e-03   0.271 0.786428    
## avgsen      -9.361e-05  1.512e-04  -0.619 0.536084    
## polpc        2.552e+00  1.730e-01  14.749  < 2e-16 ***
## density      6.504e-03  3.833e-04  16.970  < 2e-16 ***
## taxpc        1.650e-04  4.216e-05   3.914 0.000101 ***
## pctmin       2.494e-04  2.511e-05   9.934  < 2e-16 ***
## wcon        -4.618e-07  3.398e-06  -0.136 0.891945    
## wtuc        -3.140e-07  1.499e-06  -0.209 0.834175    
## wtrd         2.612e-06  4.791e-06   0.545 0.585761    
## wfir        -1.576e-05  1.133e-05  -1.391 0.164867    
## wser        -6.962e-06  3.947e-06  -1.764 0.078216 .  
## wmfg        -5.180e-06  6.617e-06  -0.783 0.434022    
## wfed         4.333e-05  1.051e-05   4.124 4.24e-05 ***
## wsta        -1.368e-05  1.092e-05  -1.253 0.210729    
## wloc         4.413e-05  1.980e-05   2.229 0.026185 *  
## mix         -4.166e-04  2.322e-03  -0.179 0.857657    
## pctymle      1.016e-01  1.764e-02   5.757 1.36e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.009781 on 608 degrees of freedom
## Multiple R-squared:  0.7184, Adjusted R-squared:  0.7086 
## F-statistic: 73.85 on 21 and 608 DF,  p-value: < 2.2e-16
```

- Creating our first model with all the predictors included we see a r^2 of 0.7184 , f stat of 73.85 and p value of 0.7086

```
# look at chart for normality with qq plots

crimeModelOneResiduals=crimeModelOne$residuals
crimeModelOneFitted=crimeModelOne$fitted





par(mfrow=c(1,2))
qqnorm(crimeModelOneResiduals,ylab="residuals")
qqline(crimeModelOneResiduals)




plot(crimeModelOneFitted,crimeModelOneResiduals,xlab="Fitted
Values",ylab="Residuals",main="Residual Vs Fitted Values")
abline(h=0)
```



- creating a normal qq plot and residuals vs fitted values we see that both plots are exhibiting problems with normality and variance
- this is further verify when using a shapiro test

```
# normalty test
shapiro.test(crimeModelOneResiduals)

## 
##  Shapiro-Wilk normality test
## 
## data:  crimeModelOneResiduals
## W = 0.92396, p-value < 2.2e-16
```

- Using our shapiro test we see our p value is 2.2 e -16 which indicate a very bad normality that can’t reject null hypothesis

```
#Normality test
library(lmtest)

## Loading required package: zoo

## 
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric

bptest(crimeModelOne)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  crimeModelOne
## BP = 225.41, df = 21, p-value < 2.2e-16
```

- Using pagan test our variance is also not in a good place with p value being less that 2.2 e-16 and unable to reject null hypothesis

```
# Check vif

#install.packages("car")
library(car)

## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

vif(crimeModelOne)

##   county     year   prbarr  prbconv  prbpris   avgsen    polpc  density
## 1.082374 4.264980 1.506879 1.778552 1.148894 1.062004 1.471866 2.001673
##    taxpc   pctmin     wcon     wtuc     wtrd     wfir     wser     wmfg
## 1.533076 1.184494 1.129754 1.049689 1.179221 2.626023 1.126062 1.952816
##     wfed     wsta     wloc      mix  pctymle
## 2.886326 2.237875 4.407395 1.711292 1.213429
```

- A good thing for our first model we see that most of our variables do not have multicolinarity issues

```
# use forward selection to see which variable to get rid off to reduce model.

#backstep selection

backStepModel= step(crimeModelOne,direction = "backward")

## Start:  AIC=-5808.78
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wcon + wtuc + wtrd + wfir +
##     wser + wmfg + wfed + wsta + wloc + mix + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wcon     1 0.0000018 0.058171 -5810.8
## - mix      1 0.0000031 0.058172 -5810.7
## - wtuc     1 0.0000042 0.058173 -5810.7
## - prbpris  1 0.0000070 0.058176 -5810.7
## - county   1 0.0000248 0.058194 -5810.5
## - wtrd     1 0.0000284 0.058198 -5810.5
## - avgsen   1 0.0000367 0.058206 -5810.4
## - wmfg     1 0.0000586 0.058228 -5810.1
## - wsta     1 0.0001502 0.058319 -5809.2
## <none>                 0.058169 -5808.8
## - wfir     1 0.0001850 0.058354 -5808.8
## - wser     1 0.0002977 0.058467 -5807.6
## - wloc     1 0.0004753 0.058644 -5805.6
## - year     1 0.0007875 0.058957 -5802.3
## - taxpc    1 0.0014659 0.059635 -5795.1
## - wfed     1 0.0016271 0.059796 -5793.4
## - pctymle  1 0.0031711 0.061340 -5777.3
## - prbconv  1 0.0056515 0.063821 -5752.4
## - pctmin   1 0.0094407 0.067610 -5716.0
## - prbarr   1 0.0118434 0.070013 -5694.0
## - polpc    1 0.0208119 0.078981 -5618.1
## - density  1 0.0275510 0.085720 -5566.5
##
## Step:  AIC=-5810.76
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wtuc + wtrd + wfir + wser +
##     wmfg + wfed + wsta + wloc + mix + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - mix      1 0.0000031 0.058174 -5812.7
## - wtuc     1 0.0000041 0.058175 -5812.7
## - prbpris  1 0.0000073 0.058178 -5812.7
## - county   1 0.0000248 0.058196 -5812.5
## - wtrd     1 0.0000283 0.058199 -5812.5
## - avgsen   1 0.0000366 0.058207 -5812.4
## - wmfg     1 0.0000586 0.058229 -5812.1
## - wsta     1 0.0001484 0.058319 -5811.2
## <none>                 0.058171 -5810.8
## - wfir     1 0.0001864 0.058357 -5810.7
## - wser     1 0.0002985 0.058469 -5809.5
## - wloc     1 0.0004737 0.058645 -5807.6
## - year     1 0.0007955 0.058966 -5804.2
## - taxpc    1 0.0014646 0.059636 -5797.1
## - wfed     1 0.0016265 0.059797 -5795.4
## - pctymle  1 0.0031705 0.061341 -5779.3
## - prbconv  1 0.0056524 0.063823 -5754.3
## - pctmin   1 0.0095358 0.067707 -5717.1
## - prbarr   1 0.0118426 0.070013 -5696.0
## - polpc    1 0.0208808 0.079052 -5619.5
## - density  1 0.0275533 0.085724 -5568.5
##
## Step:  AIC=-5812.72
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wtuc + wtrd + wfir + wser +
##     wmfg + wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wtuc     1 0.0000039 0.058178 -5814.7
## - prbpris  1 0.0000067 0.058181 -5814.7
## - county   1 0.0000248 0.058199 -5814.5
## - wtrd     1 0.0000283 0.058202 -5814.4
## - avgsen   1 0.0000351 0.058209 -5814.3
## - wmfg     1 0.0000572 0.058231 -5814.1
## - wsta     1 0.0001479 0.058322 -5813.1
## <none>                 0.058174 -5812.7
## - wfir     1 0.0001863 0.058360 -5812.7
## - wser     1 0.0002959 0.058470 -5811.5
## - wloc     1 0.0004737 0.058648 -5809.6
## - year     1 0.0008018 0.058976 -5806.1
## - taxpc    1 0.0014623 0.059636 -5799.1
## - wfed     1 0.0016439 0.059818 -5797.2
## - pctymle  1 0.0031679 0.061342 -5781.3
## - prbconv  1 0.0076969 0.065871 -5736.4
## - pctmin   1 0.0097835 0.067957 -5716.8
## - prbarr   1 0.0139667 0.072141 -5679.2
## - polpc    1 0.0209633 0.079137 -5620.8
## - density  1 0.0279556 0.086130 -5567.5
## 
## Step:  AIC=-5814.68
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen + 
##     polpc + density + taxpc + pctmin + wtrd + wfir + wser + wmfg + 
##     wfed + wsta + wloc + pctymle
## 
##           Df Sum of Sq      RSS     AIC
## - prbpris  1 0.0000070 0.058185 -5816.6
## - county   1 0.0000280 0.058206 -5816.4
## - wtrd     1 0.0000284 0.058206 -5816.4
## - avgsen   1 0.0000361 0.058214 -5816.3
## - wmfg     1 0.0000580 0.058236 -5816.1
## - wsta     1 0.0001451 0.058323 -5815.1
## <none>                 0.058178 -5814.7
## - wfir     1 0.0001903 0.058368 -5814.6
## - wser     1 0.0002955 0.058473 -5813.5
## - wloc     1 0.0004751 0.058653 -5811.6
## - year     1 0.0008078 0.058986 -5808.0
## - taxpc    1 0.0014653 0.059643 -5801.0
## - wfed     1 0.0016499 0.059828 -5799.1
## - pctymle  1 0.0031893 0.061367 -5783.1
## - prbconv  1 0.0076934 0.065871 -5738.4
## - pctmin   1 0.0098259 0.068004 -5718.4
## - prbarr   1 0.0139715 0.072149 -5681.1
## - polpc    1 0.0209675 0.079145 -5622.8
## - density  1 0.0279553 0.086133 -5569.5
## 
## Step:  AIC=-5816.61
## crmrte ~ county + year + prbarr + prbconv + avgsen + polpc + 
##     density + taxpc + pctmin + wtrd + wfir + wser + wmfg + wfed + 
##     wsta + wloc + pctymle
## 
##           Df Sum of Sq      RSS     AIC
## - county   1 0.0000276 0.058212 -5818.3
## - wtrd     1 0.0000283 0.058213 -5818.3
## - avgsen   1 0.0000360 0.058221 -5818.2
## - wmfg     1 0.0000580 0.058243 -5818.0
## - wsta     1 0.0001477 0.058333 -5817.0
## <none>                 0.058185 -5816.6
## - wfir     1 0.0001949 0.058380 -5816.5
## - wser     1 0.0002981 0.058483 -5815.4
## - wloc     1 0.0004785 0.058663 -5813.4
## - year     1 0.0008127 0.058998 -5809.9
## - taxpc    1 0.0014658 0.059651 -5802.9
## - wfed     1 0.0016712 0.059856 -5800.8
## - pctymle  1 0.0032391 0.061424 -5784.5
## - prbconv  1 0.0077182 0.065903 -5740.1
## - pctmin   1 0.0102116 0.068396 -5716.7
## - prbarr   1 0.0140684 0.072253 -5682.2
## - polpc    1 0.0209656 0.079150 -5624.7
## - density  1 0.0286424 0.086827 -5566.4
##
## Step:  AIC=-5818.31
## crmrte ~ year + prbarr + prbconv + avgsen + polpc + density +
##     taxpc + pctmin + wtrd + wfir + wser + wmfg + wfed + wsta +
##     wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wtrd     1 0.0000268 0.058239 -5820.0
## - avgsen   1 0.0000331 0.058246 -5819.9
## - wmfg     1 0.0000560 0.058269 -5819.7
## - wsta     1 0.0001374 0.058350 -5818.8
## <none>                 0.058212 -5818.3
## - wfir     1 0.0001984 0.058411 -5818.2
## - wser     1 0.0002959 0.058508 -5817.1
## - wloc     1 0.0004938 0.058706 -5815.0
## - year     1 0.0008295 0.059042 -5811.4
## - taxpc    1 0.0014420 0.059655 -5804.9
## - wfed     1 0.0016576 0.059870 -5802.6
## - pctymle  1 0.0032941 0.061507 -5785.6
## - prbconv  1 0.0076906 0.065903 -5742.1
## - pctmin   1 0.0103162 0.068529 -5717.5
## - prbarr   1 0.0141520 0.072364 -5683.2
## - polpc    1 0.0213140 0.079527 -5623.8
## - density  1 0.0286159 0.086828 -5568.4
##
## Step:  AIC=-5820.02
## crmrte ~ year + prbarr + prbconv + avgsen + polpc + density +
##     taxpc + pctmin + wfir + wser + wmfg + wfed + wsta + wloc +
##     pctymle
##
##           Df Sum of Sq      RSS     AIC
## - avgsen   1 0.0000307 0.058270 -5821.7
## - wmfg     1 0.0000500 0.058289 -5821.5
## - wsta     1 0.0001421 0.058381 -5820.5
## <none>                 0.058239 -5820.0
## - wfir     1 0.0001923 0.058432 -5819.9
## - wser     1 0.0002934 0.058533 -5818.8
## - wloc     1 0.0005047 0.058744 -5816.6
## - year     1 0.0008233 0.059063 -5813.2
## - taxpc    1 0.0014242 0.059664 -5806.8
## - wfed     1 0.0016768 0.059916 -5804.1
## - pctymle  1 0.0032776 0.061517 -5787.5
## - prbconv  1 0.0077222 0.065962 -5743.6
## - pctmin   1 0.0104165 0.068656 -5718.4
## - prbarr   1 0.0141651 0.072404 -5684.9
## - polpc    1 0.0213686 0.079608 -5625.1
## - density  1 0.0295669 0.087806 -5563.4
##
## Step:  AIC=-5821.68
## crmrte ~ year + prbarr + prbconv + polpc + density + taxpc +
##     pctmin + wfir + wser + wmfg + wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wmfg     1 0.0000553 0.058325 -5823.1
## - wsta     1 0.0001494 0.058419 -5822.1
## <none>                 0.058270 -5821.7
## - wfir     1 0.0002039 0.058474 -5821.5
## - wser     1 0.0002811 0.058551 -5820.7
## - wloc     1 0.0004955 0.058765 -5818.3
## - year     1 0.0007967 0.059067 -5815.1
## - taxpc    1 0.0014090 0.059679 -5808.6
## - wfed     1 0.0017021 0.059972 -5805.5
## - pctymle  1 0.0032554 0.061525 -5789.4
## - prbconv  1 0.0077714 0.066041 -5744.8
## - pctmin   1 0.0104615 0.068732 -5719.7
## - prbarr   1 0.0142995 0.072569 -5685.4
## - polpc    1 0.0214092 0.079679 -5626.5
## - density  1 0.0295400 0.087810 -5565.3
##
## Step:  AIC=-5823.09
## crmrte ~ year + prbarr + prbconv + polpc + density + taxpc +
##     pctmin + wfir + wser + wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wsta     1 0.0001372 0.058462 -5823.6
## <none>                 0.058325 -5823.1
## - wfir     1 0.0002743 0.058600 -5822.1
## - wser     1 0.0002897 0.058615 -5822.0
## - wloc     1 0.0004832 0.058808 -5819.9
## - year     1 0.0007970 0.059122 -5816.5
## - taxpc    1 0.0013644 0.059690 -5810.5
## - wfed     1 0.0016601 0.059985 -5807.4
## - pctymle  1 0.0032063 0.061532 -5791.4
## - prbconv  1 0.0077406 0.066066 -5746.6
## - pctmin   1 0.0109685 0.069294 -5716.5
## - prbarr   1 0.0144545 0.072780 -5685.6
## - polpc    1 0.0213570 0.079682 -5628.5
## - density  1 0.0294849 0.087810 -5567.3
##
## Step:  AIC=-5823.61
## crmrte ~ year + prbarr + prbconv + polpc + density + taxpc +
##     pctmin + wfir + wser + wfed + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## <none>                  0.058462 -5823.6
## - wser     1 0.0002893 0.058752 -5822.5
## - wfir     1 0.0002982 0.058761 -5822.4
## - wloc     1 0.0004454 0.058908 -5820.8
## - year     1 0.0012923 0.059755 -5811.8
## - taxpc    1 0.0013618 0.059824 -5811.1
## - wfed     1 0.0016269 0.060089 -5808.3
## - pctymle  1 0.0030722 0.061535 -5793.3
## - prbconv  1 0.0077120 0.066175 -5747.5
## - pctmin   1 0.0108627 0.069325 -5718.2
## - prbarr   1 0.0143330 0.072796 -5687.5
## - polpc    1 0.0212204 0.079683 -5630.5
## - density  1 0.0293532 0.087816 -5569.3
```

```
summary(backStepModel)
```

```
##
## Call:
## lm(formula = crmrte ~ year + prbarr + prbconv + polpc + density +
##     taxpc + pctmin + wfir + wser + wfed + wloc + pctymle, data =
crimeDataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028839 -0.005139 -0.000517  0.003981  0.069151
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.009e-01  2.673e-02   3.774 0.000176 ***
## year        -1.337e-03  3.619e-04  -3.693 0.000241 ***
## prbarr      -3.140e-02  2.553e-03 -12.299  < 2e-16 ***
## prbconv     -2.386e-03  2.644e-04  -9.022  < 2e-16 ***
## polpc        2.546e+00  1.701e-01  14.965  < 2e-16 ***
## density      6.478e-03  3.681e-04  17.601  < 2e-16 ***
## taxpc        1.501e-04  3.959e-05   3.791 0.000165 ***
## pctmin       2.535e-04  2.368e-05  10.707  < 2e-16 ***
## wfir        -1.924e-05  1.085e-05  -1.774 0.076564 .
## wser        -6.824e-06  3.905e-06  -1.747 0.081050 .
## wfed         4.136e-05  9.981e-06   4.144 3.90e-05 ***
## wloc         4.224e-05  1.948e-05   2.168 0.030541 *
## pctymle      9.587e-02  1.684e-02   5.694 1.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009734 on 617 degrees of freedom
## Multiple R-squared:  0.7169, Adjusted R-squared:  0.7114
## F-statistic: 130.2 on 12 and 617 DF,  p-value: < 2.2e-16
```

```
# using both steps to see if we get what same model

bothStep=step(crimeModelOne,direction = "both")
```

```
## Start:  AIC=-5808.78
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wcon + wtuc + wtrd + wfir +
##     wser + wmfg + wfed + wsta + wloc + mix + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wcon     1 0.0000018 0.058171 -5810.8
## - mix      1 0.0000031 0.058172 -5810.7
## - wtuc     1 0.0000042 0.058173 -5810.7
## - prbpris  1 0.0000070 0.058176 -5810.7
## - county   1 0.0000248 0.058194 -5810.5
## - wtrd     1 0.0000284 0.058198 -5810.5
## - avgsen   1 0.0000367 0.058206 -5810.4
## - wmfg     1 0.0000586 0.058228 -5810.1
## - wsta     1 0.0001502 0.058319 -5809.2
## <none>                 0.058169 -5808.8
## - wfir     1 0.0001850 0.058354 -5808.8
## - wser     1 0.0002977 0.058467 -5807.6
## - wloc     1 0.0004753 0.058644 -5805.6
## - year     1 0.0007875 0.058957 -5802.3
## - taxpc    1 0.0014659 0.059635 -5795.1
## - wfed     1 0.0016271 0.059796 -5793.4
## - pctymle  1 0.0031711 0.061340 -5777.3
## - prbconv  1 0.0056515 0.063821 -5752.4
## - pctmin   1 0.0094407 0.067610 -5716.0
## - prbarr   1 0.0118434 0.070013 -5694.0
## - polpc    1 0.0208119 0.078981 -5618.1
## - density  1 0.0275510 0.085720 -5566.5
##
## Step:  AIC=-5810.76
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wtuc + wtrd + wfir + wser +
##     wmfg + wfed + wsta + wloc + mix + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - mix      1 0.0000031 0.058174 -5812.7
## - wtuc     1 0.0000041 0.058175 -5812.7
## - prbpris  1 0.0000073 0.058178 -5812.7
## - county   1 0.0000248 0.058196 -5812.5
## - wtrd     1 0.0000283 0.058199 -5812.5
## - avgsen   1 0.0000366 0.058207 -5812.4
## - wmfg     1 0.0000586 0.058229 -5812.1
## - wsta     1 0.0001484 0.058319 -5811.2
## <none>                 0.058171 -5810.8
## - wfir     1 0.0001864 0.058357 -5810.7
## - wser     1 0.0002985 0.058469 -5809.5
## + wcon     1 0.0000018 0.058169 -5808.8
## - wloc     1 0.0004737 0.058645 -5807.6
## - year     1 0.0007955 0.058966 -5804.2
## - taxpc    1 0.0014646 0.059636 -5797.1
## - wfed     1 0.0016265 0.059797 -5795.4
## - pctymle  1 0.0031705 0.061341 -5779.3
## - prbconv  1 0.0056524 0.063823 -5754.3
## - pctmin   1 0.0095358 0.067707 -5717.1
## - prbarr   1 0.0118426 0.070013 -5696.0
## - polpc    1 0.0208808 0.079052 -5619.5
## - density  1 0.0275533 0.085724 -5568.5
##
## Step:  AIC=-5812.72
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wtuc + wtrd + wfir + wser +
##     wmfg + wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wtuc     1 0.0000039 0.058178 -5814.7
## - prbpris  1 0.0000067 0.058181 -5814.7
## - county   1 0.0000248 0.058199 -5814.5
## - wtrd     1 0.0000283 0.058202 -5814.4
## - avgsen   1 0.0000351 0.058209 -5814.3
## - wmfg     1 0.0000572 0.058231 -5814.1
## - wsta     1 0.0001479 0.058322 -5813.1
## <none>                 0.058174 -5812.7
## - wfir     1 0.0001863 0.058360 -5812.7
## - wser     1 0.0002959 0.058470 -5811.5
## + mix      1 0.0000031 0.058171 -5810.8
## + wcon     1 0.0000018 0.058172 -5810.7
## - wloc     1 0.0004737 0.058648 -5809.6
## - year     1 0.0008018 0.058976 -5806.1
## - taxpc    1 0.0014623 0.059636 -5799.1
## - wfed     1 0.0016439 0.059818 -5797.2
## - pctymle  1 0.0031679 0.061342 -5781.3
## - prbconv  1 0.0076969 0.065871 -5736.4
## - pctmin   1 0.0097835 0.067957 -5716.8
## - prbarr   1 0.0139667 0.072141 -5679.2
## - polpc    1 0.0209633 0.079137 -5620.8
## - density  1 0.0279556 0.086130 -5567.5
```

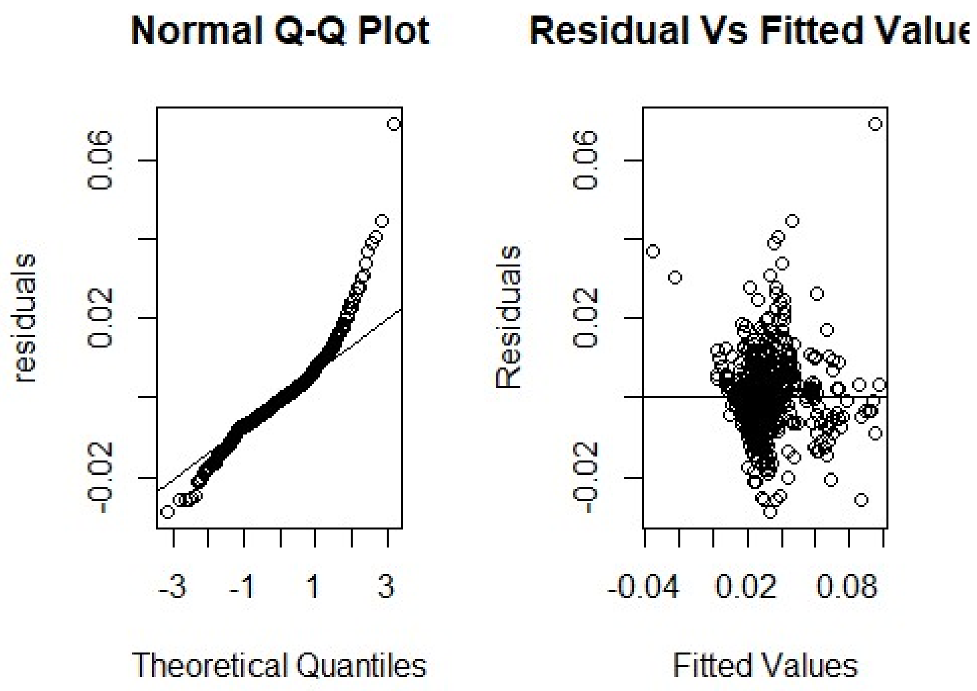
```
##
## Step:  AIC=-5814.68
## crmrte ~ county + year + prbarr + prbconv + prbpris + avgsen +
##     polpc + density + taxpc + pctmin + wtrd + wfir + wser + wmfg +
##     wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - prbpris  1 0.0000070 0.058185 -5816.6
## - county   1 0.0000280 0.058206 -5816.4
## - wtrd     1 0.0000284 0.058206 -5816.4
## - avgsen   1 0.0000361 0.058214 -5816.3
## - wmfg     1 0.0000580 0.058236 -5816.1
## - wsta     1 0.0001451 0.058323 -5815.1
## <none>                 0.058178 -5814.7
## - wfir     1 0.0001903 0.058368 -5814.6
## - wser     1 0.0002955 0.058473 -5813.5
## + wtuc     1 0.0000039 0.058174 -5812.7
## + mix      1 0.0000029 0.058175 -5812.7
## + wcon     1 0.0000017 0.058176 -5812.7
## - wloc     1 0.0004751 0.058653 -5811.6
## - year     1 0.0008078 0.058986 -5808.0
## - taxpc    1 0.0014653 0.059643 -5801.0
## - wfed     1 0.0016499 0.059828 -5799.1
## - pctymle  1 0.0031893 0.061367 -5783.1
## - prbconv  1 0.0076934 0.065871 -5738.4
## - pctmin   1 0.0098259 0.068004 -5718.4
## - prbarr   1 0.0139715 0.072149 -5681.1
## - polpc    1 0.0209675 0.079145 -5622.8
## - density  1 0.0279553 0.086133 -5569.5
##
## Step:  AIC=-5816.61
## crmrte ~ county + year + prbarr + prbconv + avgsen + polpc +
##     density + taxpc + pctmin + wtrd + wfir + wser + wmfg + wfed +
##     wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - county   1 0.0000276 0.058212 -5818.3
## - wtrd     1 0.0000283 0.058213 -5818.3
## - avgsen   1 0.0000360 0.058221 -5818.2
## - wmfg     1 0.0000580 0.058243 -5818.0
## - wsta     1 0.0001477 0.058333 -5817.0
## <none>                 0.058185 -5816.6
## - wfir     1 0.0001949 0.058380 -5816.5
## - wser     1 0.0002981 0.058483 -5815.4
## + prbpris  1 0.0000070 0.058178 -5814.7
## + wtuc     1 0.0000042 0.058181 -5814.7
## + mix      1 0.0000024 0.058183 -5814.6
## + wcon     1 0.0000019 0.058183 -5814.6
## - wloc     1 0.0004785 0.058663 -5813.4
## - year     1 0.0008127 0.058998 -5809.9
## - taxpc    1 0.0014658 0.059651 -5802.9
## - wfed     1 0.0016712 0.059856 -5800.8
## - pctymle  1 0.0032391 0.061424 -5784.5
## - prbconv  1 0.0077182 0.065903 -5740.1
## - pctmin   1 0.0102116 0.068396 -5716.7
## - prbarr   1 0.0140684 0.072253 -5682.2
## - polpc    1 0.0209656 0.079150 -5624.7
## - density  1 0.0286424 0.086827 -5566.4
##
## Step:  AIC=-5818.31
## crmrte ~ year + prbarr + prbconv + avgsen + polpc + density +
##     taxpc + pctmin + wtrd + wfir + wser + wmfg + wfed + wsta +
##     wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wtrd     1 0.0000268 0.058239 -5820.0
## - avgsen   1 0.0000331 0.058246 -5819.9
## - wmfg     1 0.0000560 0.058269 -5819.7
## - wsta     1 0.0001374 0.058350 -5818.8
## <none>                 0.058212 -5818.3
## - wfir     1 0.0001984 0.058411 -5818.2
## - wser     1 0.0002959 0.058508 -5817.1
## + county   1 0.0000276 0.058185 -5816.6
## + wtuc     1 0.0000075 0.058205 -5816.4
## + prbpris  1 0.0000066 0.058206 -5816.4
## + mix      1 0.0000023 0.058210 -5816.3
## + wcon     1 0.0000018 0.058211 -5816.3
## - wloc     1 0.0004938 0.058706 -5815.0
## - year     1 0.0008295 0.059042 -5811.4
## - taxpc    1 0.0014420 0.059655 -5804.9
## - wfed     1 0.0016576 0.059870 -5802.6
## - pctymle  1 0.0032941 0.061507 -5785.6
## - prbconv  1 0.0076906 0.065903 -5742.1
## - pctmin   1 0.0103162 0.068529 -5717.5
## - prbarr   1 0.0141520 0.072364 -5683.2
## - polpc    1 0.0213140 0.079527 -5623.8
## - density  1 0.0286159 0.086828 -5568.4
##
## Step:  AIC=-5820.02
## crmrte ~ year + prbarr + prbconv + avgsen + polpc + density +
##     taxpc + pctmin + wfir + wser + wmfg + wfed + wsta + wloc +
##     pctymle
##
##           Df Sum of Sq      RSS     AIC
## - avgsen   1 0.0000307 0.058270 -5821.7
## - wmfg     1 0.0000500 0.058289 -5821.5
## - wsta     1 0.0001421 0.058381 -5820.5
## <none>                 0.058239 -5820.0
## - wfir     1 0.0001923 0.058432 -5819.9
## - wser     1 0.0002934 0.058533 -5818.8
## + wtrd     1 0.0000268 0.058212 -5818.3
## + county   1 0.0000262 0.058213 -5818.3
## + wtuc     1 0.0000075 0.058232 -5818.1
## + prbpris  1 0.0000065 0.058233 -5818.1
## + mix      1 0.0000023 0.058237 -5818.0
## + wcon     1 0.0000017 0.058238 -5818.0
## - wloc     1 0.0005047 0.058744 -5816.6
## - year     1 0.0008233 0.059063 -5813.2
## - taxpc    1 0.0014242 0.059664 -5806.8
## - wfed     1 0.0016768 0.059916 -5804.1
## - pctymle  1 0.0032776 0.061517 -5787.5
## - prbconv  1 0.0077222 0.065962 -5743.6
## - pctmin   1 0.0104165 0.068656 -5718.4
## - prbarr   1 0.0141651 0.072404 -5684.9
## - polpc    1 0.0213686 0.079608 -5625.1
## - density  1 0.0295669 0.087806 -5563.4
##
## Step:  AIC=-5821.68
## crmrte ~ year + prbarr + prbconv + polpc + density + taxpc +
##     pctmin + wfir + wser + wmfg + wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wmfg     1 0.0000553 0.058325 -5823.1
## - wsta     1 0.0001494 0.058419 -5822.1
## <none>                 0.058270 -5821.7
## - wfir     1 0.0002039 0.058474 -5821.5
## - wser     1 0.0002811 0.058551 -5820.7
## + avgsen   1 0.0000307 0.058239 -5820.0
## + wtrd     1 0.0000244 0.058246 -5819.9
## + county   1 0.0000235 0.058246 -5819.9
## + wtuc     1 0.0000086 0.058261 -5819.8
## + prbpris  1 0.0000064 0.058264 -5819.8
## + wcon     1 0.0000017 0.058268 -5819.7
## + mix      1 0.0000011 0.058269 -5819.7
## - wloc     1 0.0004955 0.058765 -5818.3
## - year     1 0.0007967 0.059067 -5815.1
## - taxpc    1 0.0014090 0.059679 -5808.6
## - wfed     1 0.0017021 0.059972 -5805.5
## - pctymle  1 0.0032554 0.061525 -5789.4
## - prbconv  1 0.0077714 0.066041 -5744.8
## - pctmin   1 0.0104615 0.068732 -5719.7
## - prbarr   1 0.0142995 0.072569 -5685.4
## - polpc    1 0.0214092 0.079679 -5626.5
## - density  1 0.0295400 0.087810 -5565.3
##
## Step:  AIC=-5823.09
## crmrte ~ year + prbarr + prbconv + polpc + density + taxpc +
##     pctmin + wfir + wser + wfed + wsta + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## - wsta     1 0.0001372 0.058462 -5823.6
## <none>                 0.058325 -5823.1
## - wfir     1 0.0002743 0.058600 -5822.1
## - wser     1 0.0002897 0.058615 -5822.0
## + wmfg     1 0.0000553 0.058270 -5821.7
## + avgsen   1 0.0000359 0.058289 -5821.5
## + county   1 0.0000217 0.058304 -5821.3
## + wtrd     1 0.0000183 0.058307 -5821.3
## + wtuc     1 0.0000097 0.058316 -5821.2
## + prbpris  1 0.0000064 0.058319 -5821.2
## + wcon     1 0.0000016 0.058324 -5821.1
## + mix      1 0.0000003 0.058325 -5821.1
## - wloc     1 0.0004832 0.058808 -5819.9
## - year     1 0.0007970 0.059122 -5816.5
## - taxpc    1 0.0013644 0.059690 -5810.5
## - wfed     1 0.0016601 0.059985 -5807.4
## - pctymle  1 0.0032063 0.061532 -5791.4
## - prbconv  1 0.0077406 0.066066 -5746.6
## - pctmin   1 0.0109685 0.069294 -5716.5
## - prbarr   1 0.0144545 0.072780 -5685.6
## - polpc    1 0.0213570 0.079682 -5628.5
## - density  1 0.0294849 0.087810 -5567.3
##
## Step:  AIC=-5823.61
## crmrte ~ year + prbarr + prbconv + polpc + density + taxpc +
##     pctmin + wfir + wser + wfed + wloc + pctymle
##
##           Df Sum of Sq      RSS     AIC
## <none>                 0.058462 -5823.6
## + wsta     1 0.0001372 0.058325 -5823.1
## - wser     1 0.0002893 0.058752 -5822.5
## - wfir     1 0.0002982 0.058761 -5822.4
## + wmfg     1 0.0000431 0.058419 -5822.1
## + avgsen   1 0.0000427 0.058420 -5822.1
## + wtrd     1 0.0000226 0.058440 -5821.9
## + county   1 0.0000125 0.058450 -5821.7
## + prbpris  1 0.0000089 0.058454 -5821.7
## + wtuc     1 0.0000045 0.058458 -5821.7
## + mix      1 0.0000001 0.058462 -5821.6
## + wcon     1 0.0000000 0.058462 -5821.6
## - wloc     1 0.0004454 0.058908 -5820.8
## - year     1 0.0012923 0.059755 -5811.8
## - taxpc    1 0.0013618 0.059824 -5811.1
## - wfed     1 0.0016269 0.060089 -5808.3
## - pctymle  1 0.0030722 0.061535 -5793.3
## - prbconv  1 0.0077120 0.066175 -5747.5
## - pctmin   1 0.0108627 0.069325 -5718.2
## - prbarr   1 0.0143330 0.072796 -5687.5
## - polpc    1 0.0212204 0.079683 -5630.5
## - density  1 0.0293532 0.087816 -5569.3
```

```
summary(bothStep)

##
## Call:
## lm(formula = crmrte ~ year + prbarr + prbconv + polpc + density +
##     taxpc + pctmin + wfir + wser + wfed + wloc + pctymle, data =
crimeDataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.028839 -0.005139 -0.000517  0.003981  0.069151
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.009e-01  2.673e-02   3.774 0.000176 ***
## year        -1.337e-03  3.619e-04  -3.693 0.000241 ***
## prbarr      -3.140e-02  2.553e-03 -12.299  < 2e-16 ***
## prbconv     -2.386e-03  2.644e-04  -9.022  < 2e-16 ***
## polpc        2.546e+00  1.701e-01  14.965  < 2e-16 ***
## density      6.478e-03  3.681e-04  17.601  < 2e-16 ***
## taxpc        1.501e-04  3.959e-05   3.791 0.000165 ***
## pctmin       2.535e-04  2.368e-05  10.707  < 2e-16 ***
## wfir        -1.924e-05  1.085e-05  -1.774 0.076564 .
## wser        -6.824e-06  3.905e-06  -1.747 0.081050 .
## wfed         4.136e-05  9.981e-06   4.144 3.90e-05 ***
## wloc         4.224e-05  1.948e-05   2.168 0.030541 *
## pctymle      9.587e-02  1.684e-02   5.694 1.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009734 on 617 degrees of freedom
## Multiple R-squared:  0.7169, Adjusted R-squared:  0.7114
## F-statistic: 130.2 on 12 and 617 DF,  p-value: < 2.2e-16
```

- Since we have many variables to start with we are going to start eliminated unnecessary ones using feature selection.
- With feature selection we reduced our model from 21 to 11 variables

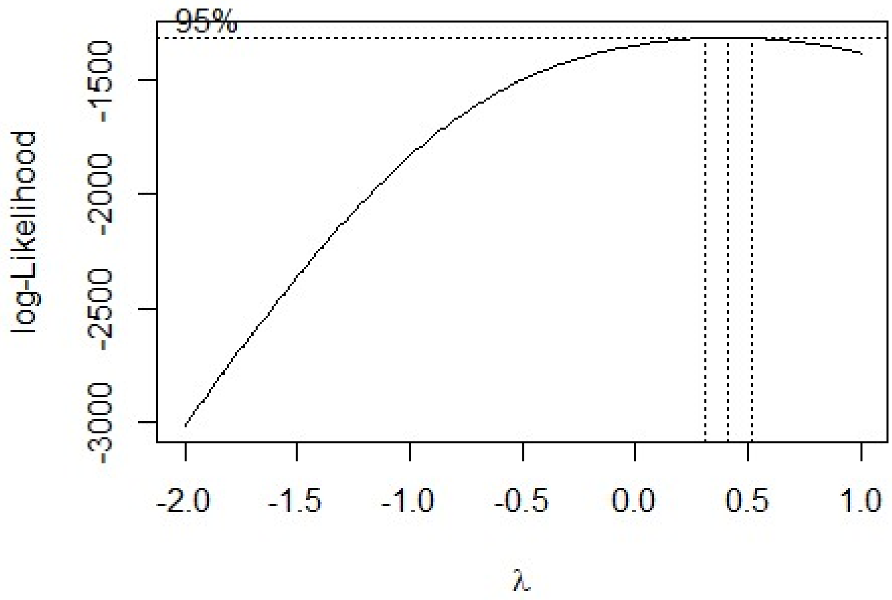
```
bothStepResidual=bothStep$residuals
bothStepFitted=crimeModelOne$fitted





par(mfrow=c(1,2))
qqnorm(bothStepResidual,ylab="residuals")
qqline(bothStepResidual)
```

```
plot(bothStepFitted,bothStepResidual,xlab="Fitted
Values",ylab="Residuals",main="Residual Vs Fitted Values")
abline(h=0)
```



- Even with reducing the count of predictors our results are still the same so now we will find a transformation to help increase our variance and linearity.

```
# Preform box cox analysis
library(MASS)


bothStepY= crimeDataset$crmrte # this is the y

bothStepX =
cbind(1,crimeDataset$year,crimeDataset$prbarr,crimeDataset$prbconv,crimeDatas
et$polpc,crimeDataset$density,crimeDataset$taxpc,crimeDataset$pctmin,crimeDat
aset$wfir,crimeDataset$wser,crimeDataset$wfed,crimeDataset$wloc,crimeDataset$
pctymle)


boxCoxResult=boxcox(bothStepY~bothStepX, lambda= seq(from=-2, to=1, by=0.01))
```



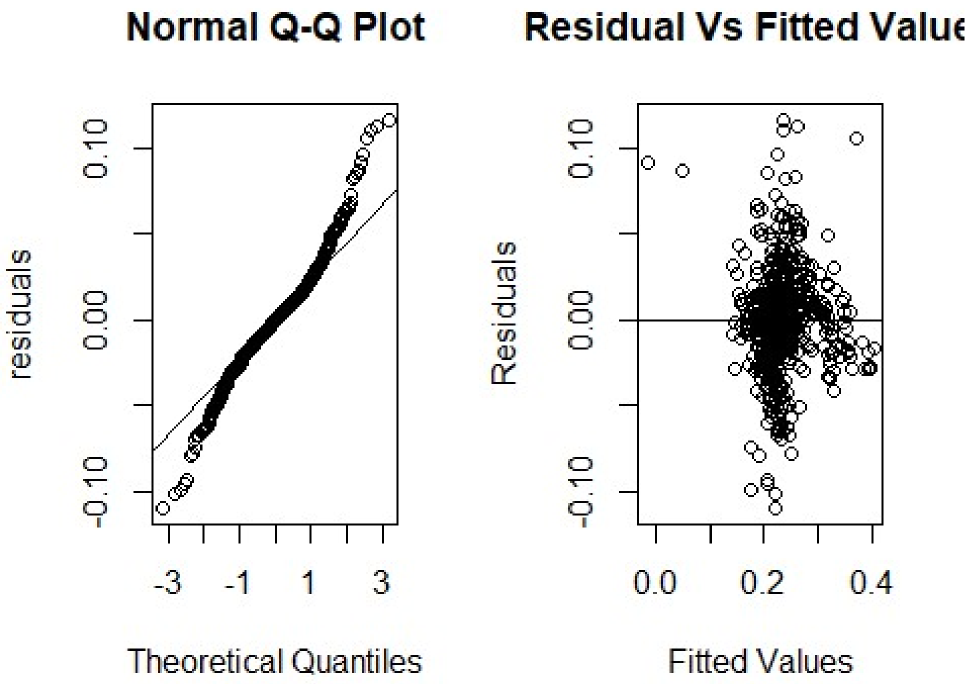
```
maxVariable=boxCoxResult$x[boxCoxResult$y==max(boxCoxResult$y)]
```

- To find our best transformation I decided to do a box cox with 95 percent certainty.
- Preforming our box cox we found a max of 0.41.

```
# Try a log transformation

#logCrimeModel=lm((crmrte^(0.41))~.,data = crimeDataset)


logCrimeModel2=lm((crmrte^(0.41))~year+prbarr+prbconv+polpc+density+taxpc+pct
min+wfir+wser+wfed+wloc+pctymle,data = crimeDataset)


summary(logCrimeModel2)

##
## Call:
## lm(formula = (crmrte^(0.41)) ~ year + prbarr + prbconv + polpc +
##     density + taxpc + pctmin + wfir + wser + wfed + wloc + pctymle,
##     data = crimeDataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109737 -0.015631  0.000215  0.014645  0.116772
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.747e-01  8.279e-02   5.734 1.54e-08 ***
## year        -4.527e-03  1.121e-03  -4.039 6.05e-05 ***
## prbarr      -1.056e-01  7.907e-03 -13.359  < 2e-16 ***
## prbconv     -7.189e-03  8.190e-04  -8.778  < 2e-16 ***
## polpc        5.591e+00  5.270e-01  10.609  < 2e-16 ***
## density      1.634e-02  1.140e-03  14.332  < 2e-16 ***
## taxpc        3.474e-04  1.226e-04   2.833  0.00476 ** 
## pctmin       8.049e-04  7.333e-05  10.976  < 2e-16 ***
## wfir        -7.144e-05  3.359e-05  -2.127  0.03383 *  
## wser        -2.375e-05  1.210e-05  -1.963  0.05007 .  
## wfed         1.904e-04  3.091e-05   6.158 1.33e-09 ***
## wloc         1.217e-04  6.035e-05   2.017  0.04414 *  
## pctymle      3.236e-01  5.215e-02   6.206 9.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.03015 on 617 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6835 
## F-statistic: 114.2 on 12 and 617 DF,  p-value: < 2.2e-16
```

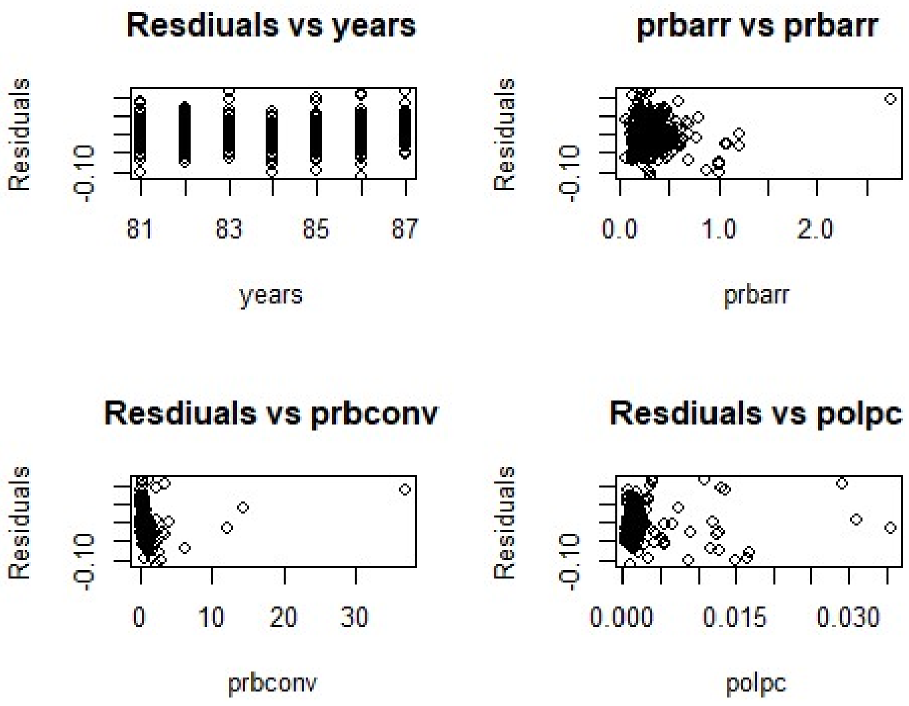
```
#normality and linearity
logCrimeRes=logCrimeModel2$residuals
logCrimeFitted=logCrimeModel2$fitted



par(mfrow=c(1,2))
qqnorm(logCrimeRes,ylab="residuals")
qqline(logCrimeRes)




plot(logCrimeFitted,logCrimeRes,xlab="Fitted
Values",ylab="Residuals",main="Residual Vs Fitted Values")
abline(h=0)
```



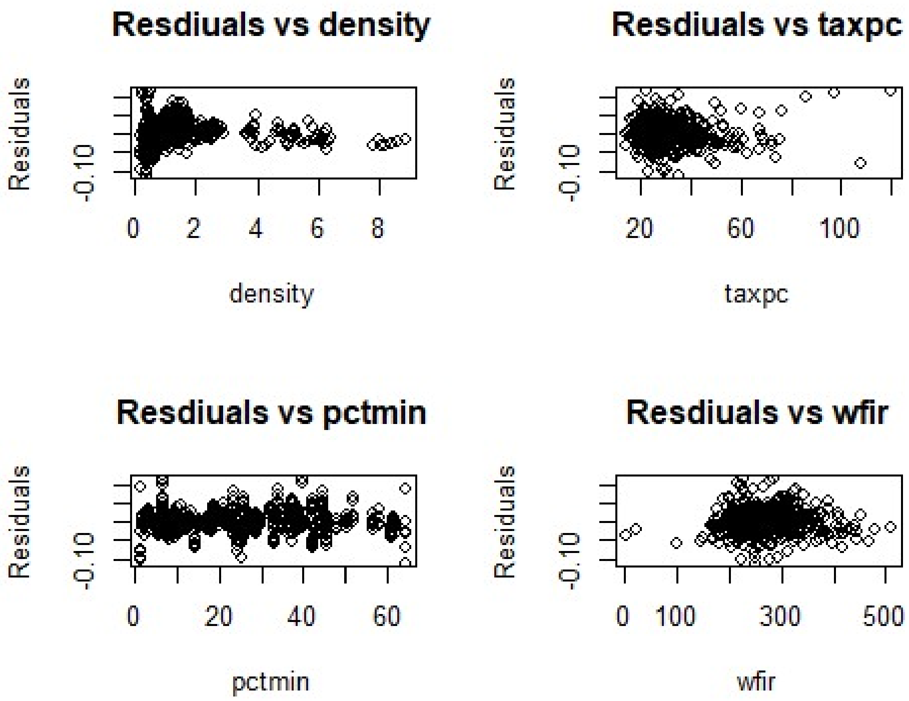
```
# wilk and bausan tests:


shapiro.test(logCrimeRes)

##
##  Shapiro-Wilk normality test
##
## data:  logCrimeRes
## W = 0.97137, p-value = 9.409e-10

bptest(logCrimeModel2)

##
##  studentized Breusch-Pagan test
##
## data:  logCrimeModel2
## BP = 168.1, df = 12, p-value < 2.2e-16

vif(logCrimeModel2)

##     year   prbarr  prbconv    polpc  density    taxpc   pctmin     wfir
## 3.483647 1.268020 1.326204 1.437181 1.863887 1.365250 1.063257 2.428251
##     wser     wfed     wloc  pctymle
## 1.113312 2.630283 4.310658 1.115669
```

- Applying the transformation of 0.41 to the y in our model increase linearity but variance remains unaffected
- All of our variable have no signs of multicollinerity according to our VIF

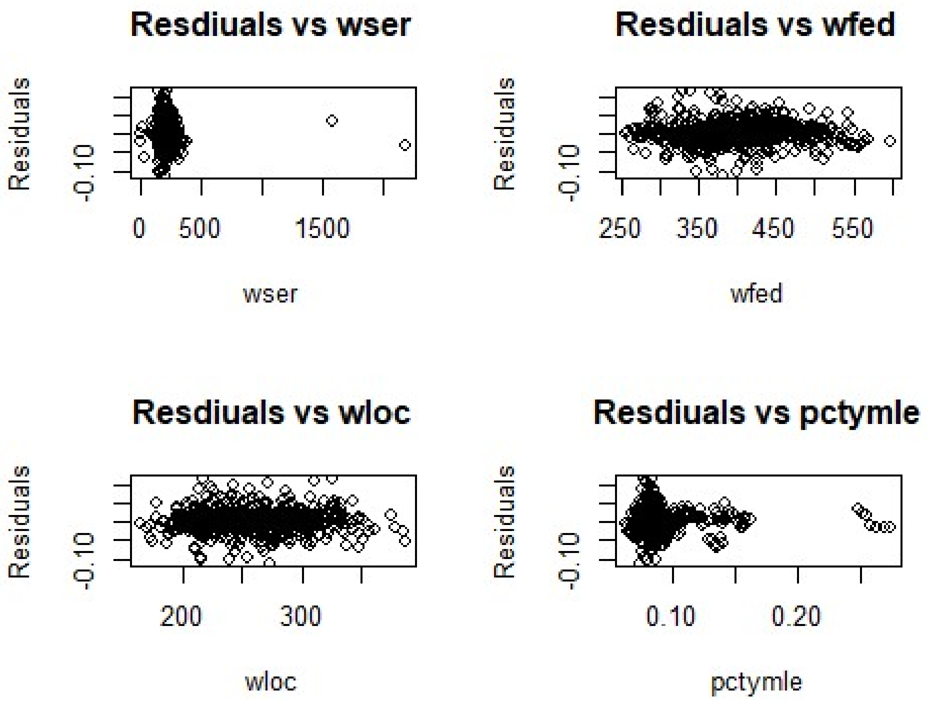
```
# looks at residuals vs resdiual graphs


par(mfrow=c(2,2))

plot(crimeDataset$year,logCrimeRes,main = "Resdiuals vs
years",xlab="years",ylab = "Residuals")
plot(crimeDataset$prbarr,logCrimeRes,main = "prbarr vs
prbarr",xlab="prbarr",ylab = "Residuals")
plot(crimeDataset$prbconv,logCrimeRes,main = "Resdiuals vs
prbconv",xlab="prbconv",ylab = "Residuals")
plot(crimeDataset$polpc,logCrimeRes,main = "Resdiuals vs
polpc",xlab="polpc",ylab = "Residuals")
```



```
plot(crimeDataset$density,logCrimeRes,main = "Resdiuals vs
density",xlab="density",ylab = "Residuals")
plot(crimeDataset$taxpc,logCrimeRes,main = "Resdiuals vs
taxpc",xlab="taxpc",ylab = "Residuals")
plot(crimeDataset$pctmin,logCrimeRes,main = "Resdiuals vs
pctmin",xlab="pctmin",ylab = "Residuals")
plot(crimeDataset$wfir,logCrimeRes,main = "Resdiuals vs
wfir",xlab="wfir",ylab = "Residuals")
```



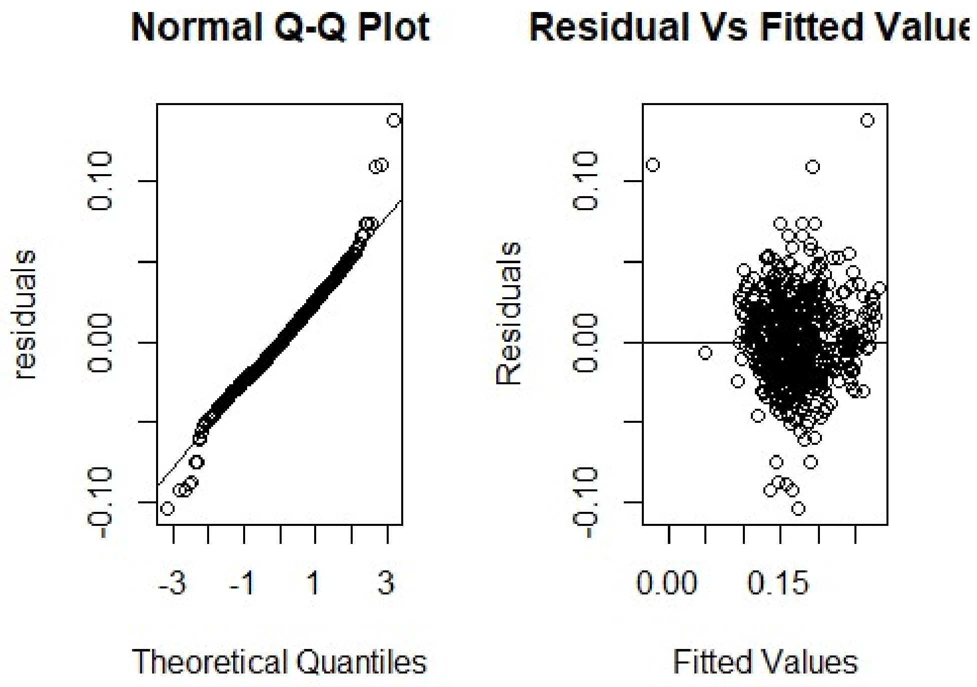
```
plot(crimeDataset$wser,logCrimeRes,main = "Resdiuals vs
wser",xlab="wser",ylab = "Residuals")
plot(crimeDataset$wfed,logCrimeRes,main = "Resdiuals vs
wfed",xlab="wfed",ylab = "Residuals")
plot(crimeDataset$wloc,logCrimeRes,main = "Resdiuals vs
wloc",xlab="wloc",ylab = "Residuals")
plot(crimeDataset$pctymle,logCrimeRes,main = "Resdiuals vs
pctymle",xlab="pctymle",ylab = "Residuals")
```



- To get more insights into our plot I decided to do a residuals vs predictor plot to see if any has signs of variance and linearity
- Looking at the year charts it can be left out but will keep due to how high our t value is in our current model with the transformation.
- prbconv,prbarr,polpc amd pctymle are at near zero, could apply a log transformation.
- The rest a scattered with either spread across the plot or together in a single spot.

```
# trying more log transformation on your near zeros

logCrimeModel2=lm((crmrte^(0.5))~year+sqrt(prbarr)+sqrt(prbconv)+sqrt(polpc)+
log(density)+sqrt(taxpc)+(1/pctmin)+wfir+wser+wfed+wloc+pctymle,data =
crimeDataset)


summary(logCrimeModel2)

##
## Call:
## lm(formula = (crmrte^(0.5)) ~ year + sqrt(prbarr) + sqrt(prbconv) +
##     sqrt(polpc) + log(density) + sqrt(taxpc) + (1/pctmin) + wfir +
##     wser + wfed + wloc + pctymle, data = crimeDataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.103534 -0.018181 -0.001046  0.017090  0.137887
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.784e-01  7.517e-02   6.364 3.84e-10 ***
## year          -4.167e-03  1.042e-03  -4.000 7.10e-05 ***
## sqrt(prbarr)  -1.317e-01  1.049e-02 -12.560  < 2e-16 ***
## sqrt(prbconv) -4.939e-02  3.999e-03 -12.351  < 2e-16 ***
## sqrt(polpc)    1.204e+00  8.728e-02  13.795  < 2e-16 ***
## log(density)   2.244e-02  2.265e-03   9.907  < 2e-16 ***
## sqrt(taxpc)    7.150e-03  1.416e-03   5.050 5.83e-07 ***
## wfir          -4.823e-05  3.032e-05  -1.591 0.112233
## wser          -7.969e-06  1.092e-05  -0.730 0.465677
## wfed           1.132e-04  2.954e-05   3.830 0.000141 ***
## wloc           7.558e-05  5.546e-05   1.363 0.173465
## pctymle        1.504e-01  4.917e-02   3.060 0.002312 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02724 on 618 degrees of freedom
## Multiple R-squared:  0.6815, Adjusted R-squared:  0.6758
## F-statistic: 120.2 on 11 and 618 DF,  p-value: < 2.2e-16

#normality and linearity
logCrimeRes=logCrimeModel2$residuals
logCrimeFitted=logCrimeModel2$fitted


par(mfrow=c(1,2))
qqnorm(logCrimeRes,ylab="residuals")
qqline(logCrimeRes)



plot(logCrimeFitted,logCrimeRes,xlab="Fitted
Values",ylab="Residuals",main="Residual Vs Fitted Values")
abline(h=0)
```

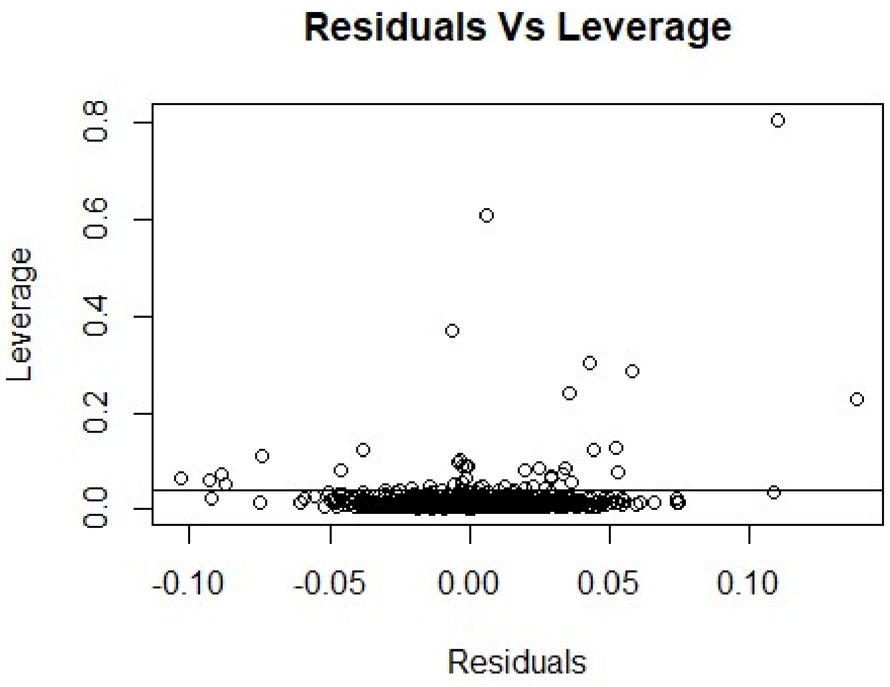


Normal Q-Q Plot

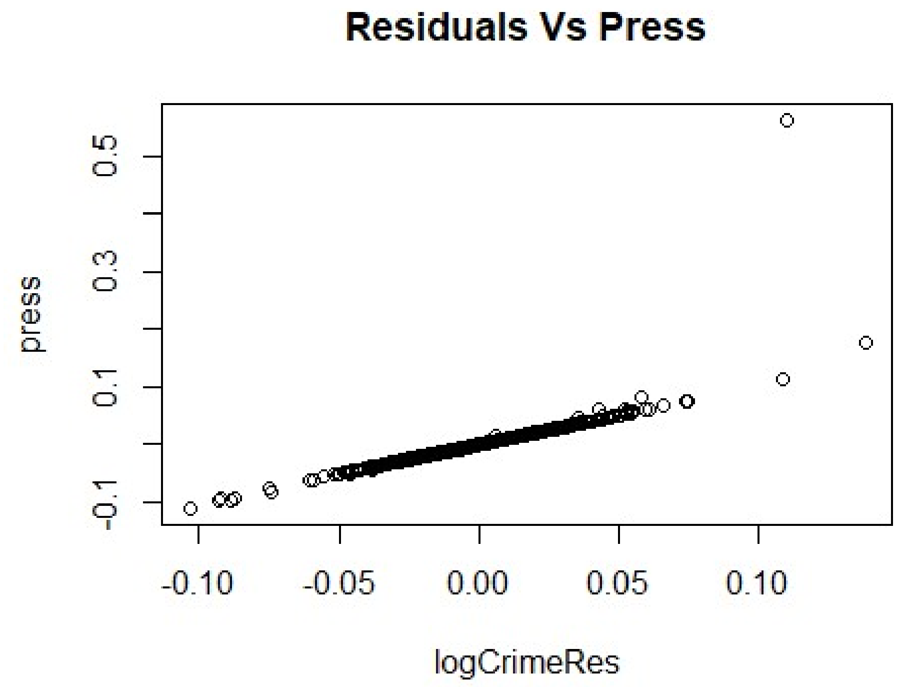
Residual Vs Fitted Valu

- Here we applied some transformations based on our graphs and the results of trial and error.
- Looking at our plots we see that some of our results seem to have improved linearity and some variance but not enough yet.

```
# check for residuals vs leverage  and  residuals vs press:

sig=summary(logCrimeModel2)$sigma

X=cbind(1,crimeDataset$year,crimeDataset$prbarr,crimeDataset$prbconv,crimeDat
aset$polpc,crimeDataset$density,crimeDataset$taxpc,crimeDataset$pctmin,crimeD
ataset$wfir,crimeDataset$wser,crimeDataset$wfed,crimeDataset$wloc,crimeDatase
t$pctymle)

hat=X%*%solve(t(X)%*%X)%*%t(X)


p=dim(X)[2]
n=length(crimeDataset$year)


plot(logCrimeRes, diag(hat), xlab='Residuals', ylab='Leverage',
main='Residuals Vs Leverage')
abline(h=2*p/n)
```



```
sum(diag(hat)>2*p/n)

## [1] 40

#press residuals

press=logCrimeRes/(1-diag(hat))
plot(logCrimeRes, press, main='Residuals Vs Press')
```



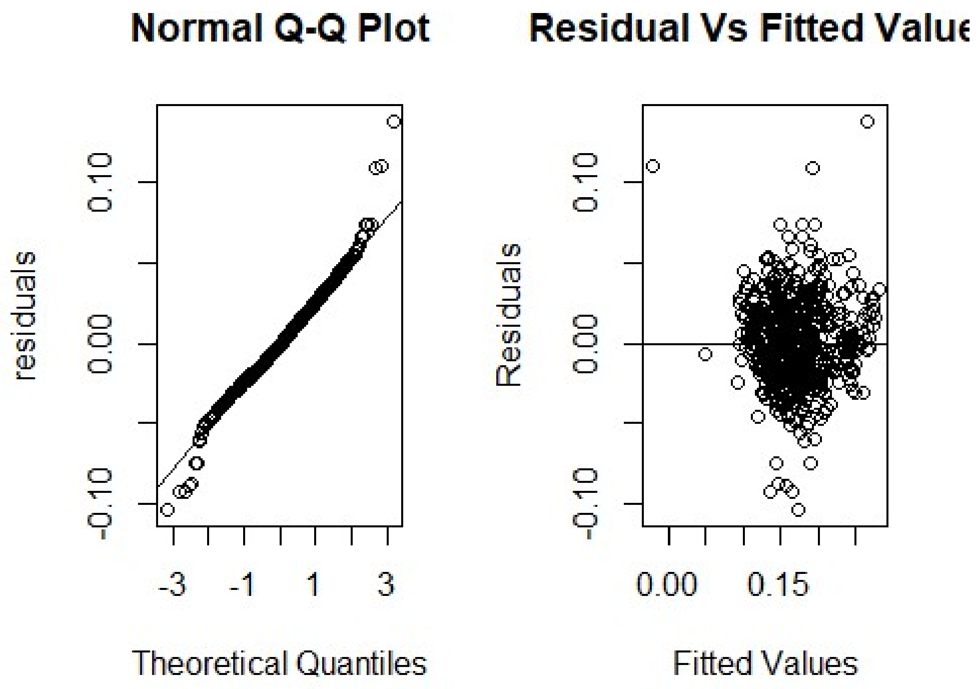
- Doing our leverage and press graph we see that there are some influential points that are affecting our results.
- Getting rid of these could yield better results for our model.

```
# remove outliers using  cooks distance and r students

#w <- abs(rstudent(bothStep)) < 3 & abs(cooks.distance(bothStep)) <
4/nrow(bothStep$model)

# noInfluenceModel <-update(bothStep, weights=as.numeric(w))



HighLeverage <- cooks.distance(bothStep) > (4/nrow(crimeDataset))
LargeResiduals <- rstudent(bothStep) > 3
hsb2 <- crimeDataset[!HighLeverage & !LargeResiduals,]
noInfluenceModel <- update(logCrimeModel2,data=hsb2)
```

- Here we used cooks distance and r student to detect influential points in our data set and updated our model to use the new data set.

```
# qqline and residuals plots

noInfluenceModelResiduals=noInfluenceModel$residuals

noInfluenceModelFitted=noInfluenceModel$fitted
```

```
par(mfrow=c(1,2))
qqnorm(logCrimeRes,ylab="residuals")
qqline(logCrimeRes)



plot(logCrimeFitted,logCrimeRes,xlab="Fitted
Values",ylab="Residuals",main="Residual Vs Fitted Values")
abline(h=0)
```



```
shapiro.test(noInfluenceModelResiduals)

##
##  Shapiro-Wilk normality test
##
## data:  noInfluenceModelResiduals
## W = 0.9971, p-value = 0.3828

bptest(noInfluenceModel)

##
##  studentized Breusch-Pagan test
##
## data:  noInfluenceModel
## BP = 33.567, df = 11, p-value = 0.0004256
```

```
vif(noInfluenceModel)

##          year  sqrt(prbarr) sqrt(prbconv)   sqrt(polpc)  log(density)
##      3.956121      1.524360      1.489649      1.216640      3.248816
##   sqrt(taxpc)          wfir          wser          wfed          wloc
##      1.526903      2.902335      2.119042      3.093525      4.576477
##       pctymle
##      1.252220

summary(noInfluenceModel)

##
## Call:
## lm(formula = (crmrte^(0.5)) ~ year + sqrt(prbarr) + sqrt(prbconv) +
##     sqrt(polpc) + log(density) + sqrt(taxpc) + (1/pctmin) + wfir +
##     wser + wfed + wloc + pctymle, data = hsb2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064494 -0.016073 -0.001727  0.014975  0.062828
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.418e-01  6.533e-02   8.294 7.74e-16 ***
## year          -5.088e-03  9.071e-04  -5.608 3.17e-08 ***
## sqrt(prbarr)  -1.007e-01  1.085e-02  -9.278  < 2e-16 ***
## sqrt(prbconv) -6.706e-02  5.771e-03 -11.621  < 2e-16 ***
## sqrt(polpc)    8.540e-01  1.243e-01   6.869 1.67e-11 ***
## log(density)   2.231e-02  2.157e-03  10.344  < 2e-16 ***
## sqrt(taxpc)    9.191e-03  1.391e-03   6.606 8.94e-11 ***
## wfir          -6.574e-06  2.897e-05  -0.227 0.820566
## wser          -9.953e-05  2.961e-05  -3.361 0.000827 ***
## wfed           1.469e-04  2.544e-05   5.775 1.26e-08 ***
## wloc           9.267e-05  4.748e-05   1.952 0.051442 .
## pctymle        1.756e-01  4.243e-02   4.139 4.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02215 on 578 degrees of freedom
## Multiple R-squared:  0.7476, Adjusted R-squared:  0.7428
## F-statistic: 155.7 on 11 and 578 DF,  p-value: < 2.2e-16
```

- Applying our updated data set with no influential points and our model with custom transformation we have achieve a much better normality and variance.
- Our variance is still below expectation but much better and where we started.
- our normality p value is 0.3828.
- our pagan test p value 0.0004256.
- There is no signs of issues of multicolinearity using our VIF.

Final Model

- For our final model I decided to go with the function : Y^0.5=year+sqrt(prbarr)+sqrt(prbconv)+sqrt(polpc)+log(density)+sqrt(taxpc)+(1/ pctmin)+wfir+wser+wfed+wloc+pctymle along side our dataset with no influential points.
- The model included 12 variables each selected through variable selection and feature transformations. Below I will expain the reason for each of their inclement and how they relate to our crime rate:
- year is included because of how negatively corelated it is to our crime rate, so as crime rate decreases so do the year it was commited.
- the square root of prbarr probability of arrest is included because of high negative t value meaning with increasing crime rate so does our probality of arrest
- square root of prbconv (probablity of conviction) is included for its high t value and how it affects crime rate. When crime rate increases so does the chances of convictions
- square root of polpc number of police per capital is included due to significant t value and how crime rate increases with the influx of police per capital.
- log of density is include due to it high its t value is and how the increase in population natuarlly increase the amount crime rate increases.
- square root of taxpc was included with how much it affects crime rate. As taxes per captial increases do does our crime rate.
- wfir is insignificant to our model but was included due to how it decrease our variance and linearity if removed.
- wser, wfed, wloc are all wage variables included to make the model better and are signifcant but play a small role in affecting our model is more used to balance out the variance and linearity of our results.
- pctymle percent young male is insignificant but was choosen mainly due to how big a part it plays in normalizing our model. removing it significantly decreases our variance and linearity.
- Other parts attempted:
  - Other things that were attempted was trying another back step selection but results proved to be negiable and our r^2 decreased as a results.
  - Removing other insignificant variables proved to decreased our variance and linearity along side our r^2
  - Applying other transformation to our varibles proved to impove our model only minisule and affected the results of our r^2 and variance to be lower than before being apply.

Final words

- The current final model selected presents a model that has good normality with low variance, but this is the best that could be achieved with my current knowledge of the tools and system. It has a r^2 of 0.7476, with F-statistic: 155.7 and p value < 2.2e-16. A possible transformation could be using ridge regression to increase the variance in exchange for more bias in the model. This concludes the result of my project in trying to solve a model that calculates the crime rate using the given data, thank you.