PROJECT REPORT: CRIME RATES

CSC 423 – DATA ANALYSIS AND REGRESSION

WINTER QUARTER 2016 - 2017

MEMBERS-

TIMOTHY WARD

TENZIN BHUTIA

VISHNU VARDHAN ANNABATTIN

NON-TECHNICAL SUMMARY

Purpose

The goal of the project is to formulate successful predictive models to predict the rates of serious crimes and identify significant predictors that affect crime rates. The dataset provided has county demographic information for 440 most populous counties in the US for the years 1990-1992 and has 17 different qualitative and quantitative variables including crime rates, which is our dependent variable.

Findings

Based on our final model we narrowed down to Region, Poverty, Total Population, per Capita income, Percentage of Beds in a County and percentage of population between ages 18-34 as the most significant predictors of crime rates. It is highly probable that some regions will have higher crime rate, while others low. If the percentage of younger (ages 18-34) population is high then we may most likely see rise in crime rates. Another well-known hypothesis our model supports is that poverty and low income is not healthy and that it does contribute to growth in crimes.

Some predictors which we hypothesised would be significant, proved insignificant otherwise after our analysis and we were able exclude several of those from our model. Some of the surprising predictors we eliminated were unemployment and education. We initially assumed that higher unemployment and lower education rate would significantly contribute to higher crime rate. It is worth noticing based on our model that poverty and per capita income are good predictors but factors like unemployment and per capita income, which has effect on poverty and per capita income may not be good indicator of crime rate.

We also pondered over whether to include percentage of hospital beds in a county in our final model even though analysis showed it was significant. The reason behind this was that rise in number of hospital beds were consequence of increase in crime rate and not the other way round. We did eventually decide to include it in our final model as in the event that of any other predictors had data missing, having this in our predictive model would be helpful. For example, if we did not know what the crime rates were or what the per capita income was, but if we knew the percentage of bed increased, we could predict that crime rate has increased.

Limitations

We felt the dataset was limited. Though it had multiple variables related to education, income and hospital, it did not have any predictors related to law enforcement, which could be significant predictors for crime rates. We had findings in our model suggesting that higher percentage of population between ages 18-34 would contribute to higher crime rate. If we had additional supplementary data for this age group, for example, say if we knew whether they were victims of substance abuse or orphans, could have helped us further in our analysis. Taking data from random regions and drawing comparisons skewed our results. For example Kingston had readings way of chart compared to rest of the counties. County specific model would have been much more accurate.

TECHNICAL SUMMARY

Crime Rate data set consists of serious crime rates in 440 county observations across US with attributes County, State, Land area, Total Population, Population between ages of 18 to 34, Population over the age of 65 & above, rate of professionally active nonfederal physicians per 1000 population, Rate of beds, cribs and bassinets per 1000 population, Rate of serious crimes, Percent of adult population (25 years old or older) who completed 12 or more years of school, Percent of adult population (25 years old or older) with bachelor's degree, Percent of 1990 population with income below poverty, Percent of 1990 labor force that is unemployed, Per capita income, Total personal income of 1990 population (in millions of dollars) and Region which is geographic region classification (1=NE, 2=NC, 3=S, 4=W).

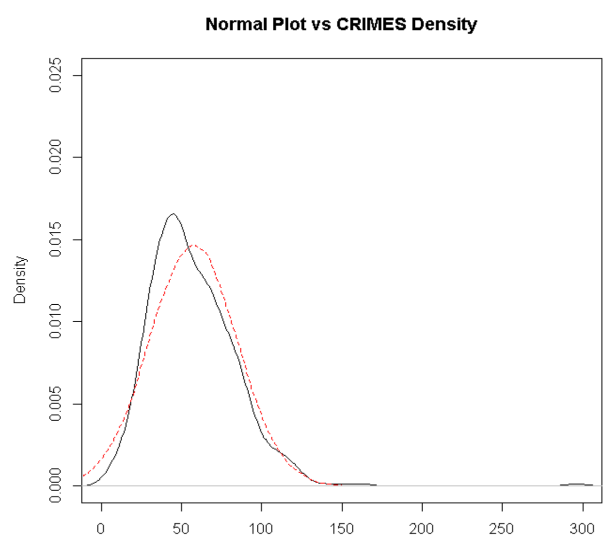
Below is the small projection of the project data set.

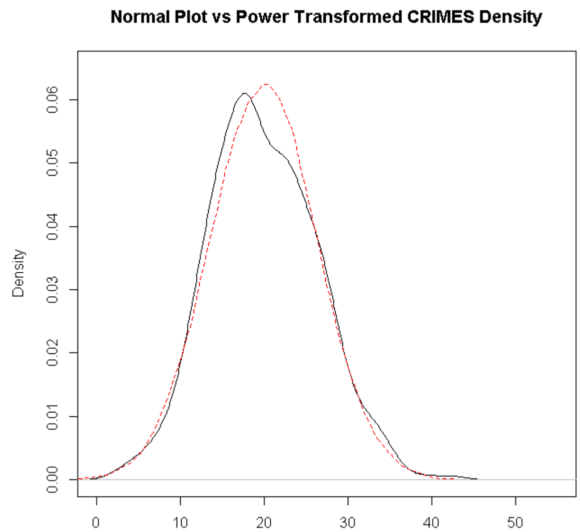
```
> head(data)
  ID      county State Land total_Pop Pop18_34 Pop65plus     DOCS     BEDS    CRIMES Hsgrads Bgrads poverty unemp Pcimcome Pers_income region
1  1 Los_Angeles    CA 4060   8863164     32.1       9.7 2.671394 3.125295  77.73026    70.0   22.3    11.6   8.0    20786      184230      4
2  2        Cook    IL  946   5105067     29.2      12.4 2.968227 4.221296  85.58869    73.4   22.8    11.1   7.2    21729      110928      2
3  3      Harris    TX 1729   2818199     31.3       7.1 2.680080 4.417360  89.96029    74.9   25.4    12.5   5.7    19517       55003      3
4  4   San_Diego    CA 4205   2498016     33.5      10.9 2.363876 2.473563  69.58362    81.9   25.3     8.1   6.1    19588       48931      4
5  5      Orange    CA  790   2410556     32.6       9.2 2.514773 2.642129  59.95463    81.2   27.8     5.2   4.8    24400       58818      4
6  6       Kings    NY   71   2300664     28.3      12.4 2.112868 3.886704 295.98672    63.7   16.6    19.5   9.5    16803       38658      1
>
```

Objective of the project activity is to predict the crimes rates in various county locations using provided data set and to analyze which predictors are more significant to explain changes in crime rates.

Exploratory data analysis:

Below graph explains the distribution of crimes rates in comparison to standard normal curve. Plot is somewhat close to normal curve yet transformation can be applied to get it close to normal curve. based on the mean, median, skewness, and kurtosis CRIMES may benefit from a transformation to be closer to normal. Also, it CRIMES has one extremely big outlier that probably should be removed.





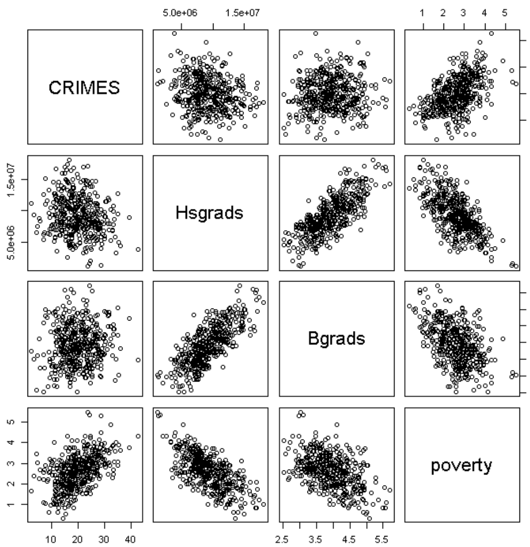
The distribution of box cox transformation of CRIMES seems to provide a better normal fit than CRIMES alone. This can be confirmed by the mean and median being almost exactly same, skewness being very close to zero and the Kurtosis being very close to zero (Appendix has code outputs of this transformation). Even if we examine the density plot alone, we can see that the plot overlay very closely on top of each other with exception of a small shift towards the peak of the plot. Last, this transformation has helped greatly with the extreme outlier we had.

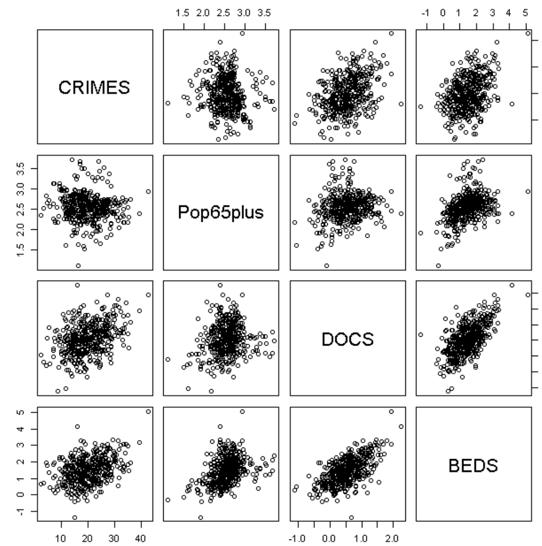
When reviewing the graphs and correlation matrix, the power transformation improved the linear relationship of many variables visually and numerically. However, for some variables such as unemployment, Bgrads and Pop65 plus the transformation has made it clearer that the relationship with CRIMES is almost non-linear which was a bit ambiguous. Overall, the transformation has made the linear relationship a lot more apparent.

Below is the correlation matrix of the data set. CRIMES have **reasonable positive correlation** with Poverty and Region “South”.

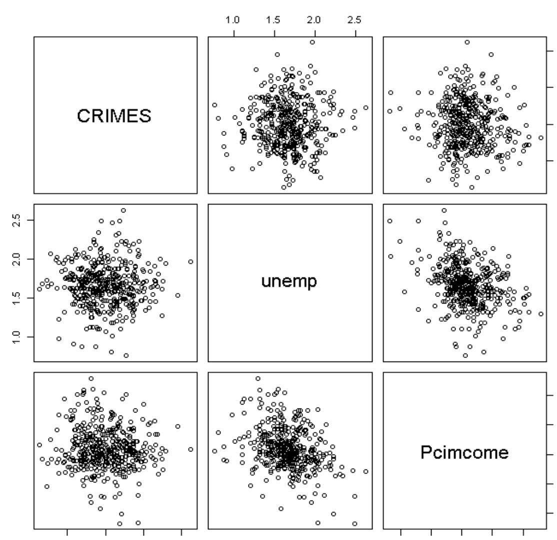
```
> cor(data_new5)
                 Land    total_Pop     Pop18_34    Pop65plus        DOCS         BEDS       CRIMES     Hsgrads       Bgrads     poverty
Land       1.00000000 0.0088576001 -0.1471164507 0.0193463697 -0.23433856 -0.2068018886 -0.035335462 -0.07494338 -0.197997643 0.18435375
total_Pop  0.00885760 1.0000000000  0.1236109055 0.0007902513  0.44841020  0.0939435064  0.381916147  0.13144584  0.353953768 -0.05980487
Pop18_34  -0.14711645 0.1236109055  1.0000000000 -0.6676283458  0.22537335 -0.0032069860  0.225558793  0.25133594  0.463664270 0.04258416
Pop65plus  0.01934637 0.0007902513 -0.6676283458 1.0000000000  0.16563436  0.3854073475 -0.068925570 -0.38205610 -0.367118874 0.11818664
DOCS      -0.23433856 0.4484102021  0.2253733542 0.1656343560  1.00000000  0.6175591743  0.379559266  0.17604582  0.530927304 0.04778874
BEDS      -0.20680189 0.0939435064 -0.0032069860 0.3854073475  0.61755917  1.0000000000  0.370098727 -0.29818706 -0.066711774 0.44761389
CRIMES    -0.03533546 0.3819161468  0.2255587931 -0.0689255705  0.37955927  0.3700987268  1.000000000 -0.19352176  0.080282524 0.48942254
Hsgrads   -0.07494338 0.1314458387  0.2513359386 -0.3820560973  0.17604582 -0.2981870643 -0.193521760  1.00000000  0.758209672 -0.66357989
Bgrads    -0.19799764 0.3539537678  0.4636642704 -0.3671188735  0.53092730 -0.0667117742  0.080282524  0.75820967  1.000000000 -0.47066163
poverty    0.18435375 -0.0598048731 0.0425841584 0.1181866435  0.04778874  0.4476138850  0.489422539 -0.66357989 -0.470661627 1.00000000
unemp      0.18382188 -0.0434728597 -0.3152802645 0.3274871393 -0.29943750  0.0020317722  0.005169391 -0.58885039 -0.608163303 0.39328253
Pcimcome  -0.31396814 0.4399672230  0.0010531296 0.0331958104  0.45188149 -0.0491099170 -0.066796345  0.53761010  0.673912513 -0.69856580
NE        -0.14316756 0.0750893412 -0.0619673613 0.2563781640  0.09627401  0.0009487865 -0.413570062 -0.01512609  0.060821877 -0.30238630
NC        -0.15079258 -0.1029085748 -0.0008666719 -0.0383244088 -0.10777632  0.1231018796 -0.134636823  0.13163863 -0.108064017 -0.07418040
S         -0.12923961 -0.0734146797  0.0784530600 -0.1283008479  0.01709171  0.0977378920  0.410470952 -0.24030882 -0.008835663 0.27670701
                unemp    Pcimcome           NE           NC            S
Land       0.183821881 -0.31396814 -0.1431675629 -0.1507925807 -0.129239606
total_Pop -0.043472860  0.43996722  0.0750893412 -0.1029085748 -0.073414680
Pop18_34  -0.315280264  0.00105313 -0.0619673613 -0.0008666719  0.078453060
Pop65plus  0.327487139  0.03319581  0.2563781640 -0.0383244088 -0.128300848
DOCS      -0.299437501  0.45188149  0.0962740063 -0.1077763235  0.017091713
BEDS       0.002031772 -0.04910992  0.0009487865  0.1231018796  0.097737892
CRIMES     0.005169391 -0.06679634 -0.4135700624 -0.1346368231  0.410470952
Hsgrads   -0.588850390  0.53761010 -0.0151260920  0.1316386260 -0.240308822
Bgrads    -0.608163303  0.67391251  0.0608218773 -0.1080640167 -0.008835663
poverty    0.393282532 -0.69856580 -0.3023862962 -0.0741803980  0.276707008
unemp      1.000000000 -0.33487286  0.1821632326 -0.0838539737 -0.102916492
Pcimcome  -0.334872857  1.00000000  0.2767055174 -0.0036404304 -0.214587411
NE         0.182163233  0.27670552  1.0000000000 -0.3142555933 -0.400374105
NC        -0.083853974 -0.00364043 -0.3142555933  1.0000000000 -0.415698758
S         -0.102916492 -0.21458741 -0.4003741052 -0.4156987575  1.000000000
>
```

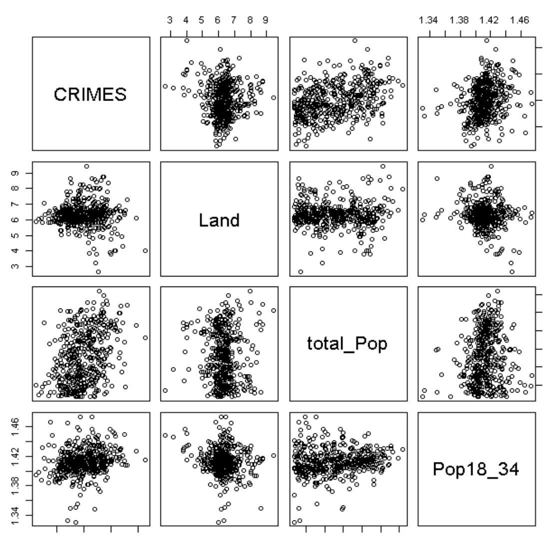
Below are Pair wise association between dependent and independent variables. These plots are plotted between dependent and independent variable after box cox transformation. CRIMES have some linear association with Poverty, DOCS & BEDS but not much association with Hsgrads, Bgrads & Population with age of 65 and above.





Below plots doesn’t provide any insight into linear association of CRIMES with Unemployment, Per capita income, Land, total population and with population with ages between 18 & 34.





Modeling:

Evaluated the whole model for any existence of collinearity among independent variables and below is the output of VIF function,

```
> vif(fit)
      Land  total_Pop   Pop18_34  Pop65plus       DOCS       BEDS    Hsgrads     Bgrads    poverty      unemp   Pcimcome Pers_income
  1.526062  79.685999   2.608395   2.063859   3.348289   3.516911   4.900643   7.156282   4.380737   2.143236   6.292047   85.135133
      d_NC       d_NE        d_S
  2.710021   2.810854   3.110685
>
```

Variables **total_Pop** and **Pers_income** have high collinearity with **VIF > 10**. Examining other variables there wasn't a correlation stronger than about 70%. So it is necessary to only remove one of the two variables to avoid multi collinearity. Hence removing **Pers_income** variable from the model and keeping **total_Pop.**

Regression analysis has been performed for the initial model after removing the multi collinearity variable. With inputs from initial model, variable selection methods such as Step wise regression and Backward selection have been used using Adj-R2. The Adj-R2 test selected a model with **12 variables** and provided an Adj-R2 value of **0.621**. The Step-Wise Regression and Reverse Regression selected a model with **10 independent variables** and has an Adj-R2 value of 0.6212. This time it's clear the model predicting power are almost exactly same. However, given the issue of dimensionality, we favored for the model with only 10 independent variables. With feedback from these method executions, removed independent variables **Pop65plus**, **DOCS**, **HSgrads** & **unemp**. Final regression model has been executed after removing these variables and standardizing the variables. (Appendix has the code output of Initial Model, Step Wise, Backward Selection and Standardized variables).

Last, we've notice that the model with 10 independent variables has 3 variables that are not significant at the 5 % level. The variables are **Bgrads**, **land** and **South**. Since we favored of keeping **South** and removed the other two variables. This change led to an Adj-R2 value of 0.6177, a difference of 0.0035 compared to inclusion the two variables.

Below are the regression equation and code output.

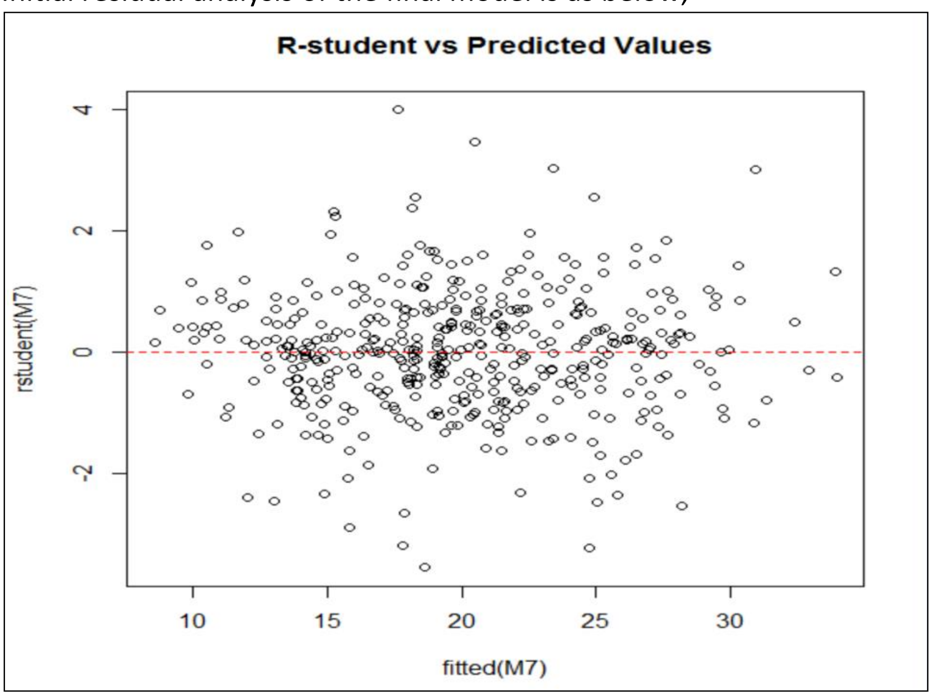
Equation:

CRIMES^(0.6629) = -2727.125 + 1258.461*Total_Pop^(-.4975) + 49.628*Pop18_34^(-0.6179) + 1.248*BEDS^(0.3309) + 3.560*Poverty^(0.2223) + 23.775 * Pcimcome^(-0.1092) – 5.315 NE – 2.143 NC + 1.779 * S

```
> M7<-lm(CRIMES~total_Pop+Pop18_34+BEDS+poverty+Pcimcome+NE+NC+S,data=data_new5)
> summary(M7)

Call:
lm(formula = CRIMES ~ total_Pop + Pop18_34 + BEDS + poverty +
    Pcimcome + NE + NC + S, data = data_new5)

Residuals:
    Min      1Q  Median      3Q     Max
-13.736  -2.282   0.113   2.461  15.337

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2727.1251   308.4033  -8.843  < 2e-16 ***
total_Pop    1258.4619   162.1222   7.762 6.14e-14 ***
Pop18_34       49.6281    10.7235   4.628 4.90e-06 ***
BEDS            1.2484     0.3203   3.898 0.000113 ***
poverty         3.5607     0.4523   7.872 2.86e-14 ***
Pcimcome       23.7755     4.9665   4.787 2.33e-06 ***
NE             -5.3151     0.6544  -8.123 4.88e-15 ***
NC             -2.1439     0.6488  -3.304 0.001032 **
S               1.7797     0.5811   3.063 0.002332 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.944 on 430 degrees of freedom
Multiple R-squared:  0.6247,    Adjusted R-squared:  0.6177
F-statistic: 89.48 on 8 and 430 DF,  p-value: < 2.2e-16

> |
```

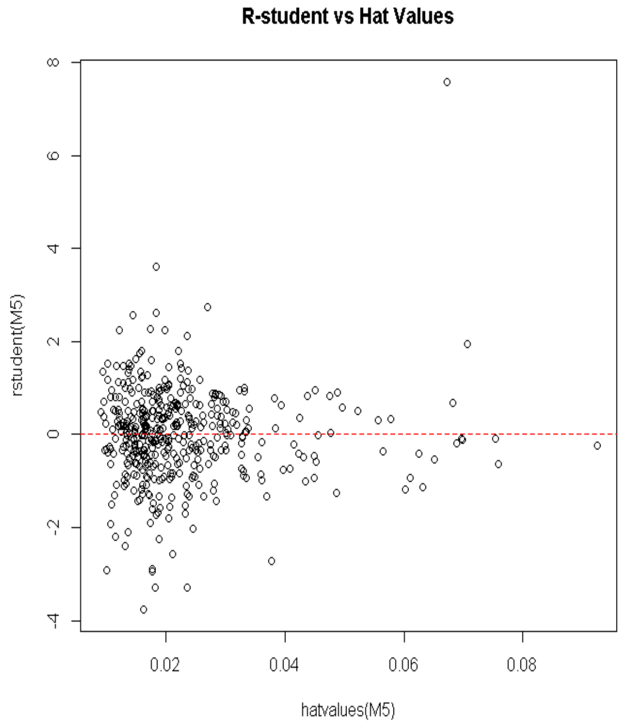
Diagnostics/residual analysis:

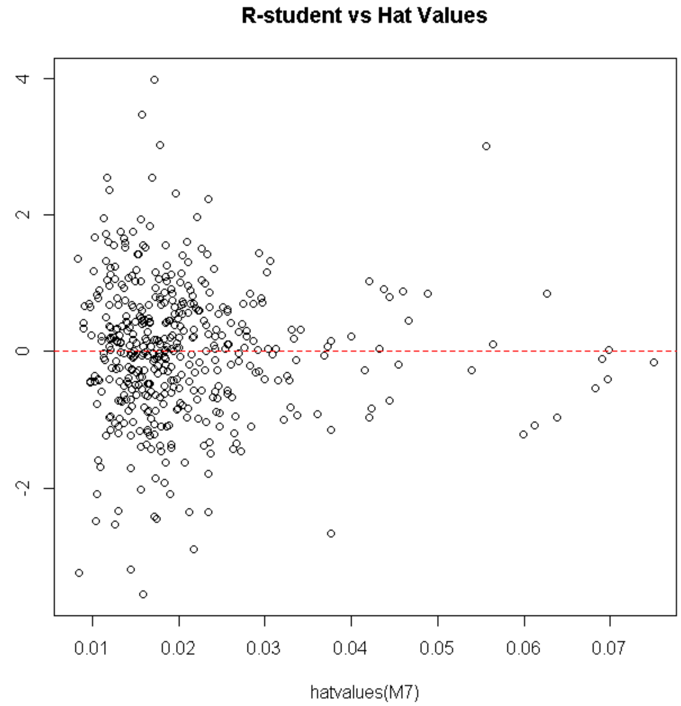
Initial residual analysis of the final model is as below,



Based on our original analysis, there was one extreme outlier that was very different from all of the rest of the points in the dataset. After deciding to remove this very influential point we can see the "R-Student vs Hat Values" residual plots improved dramatically. Some other highly possible outliers now do not deviate as far from the rest of the residuals. This overall, decrease the number of possible outliers and dramatically increased the Adj-R2 value in our analysis.

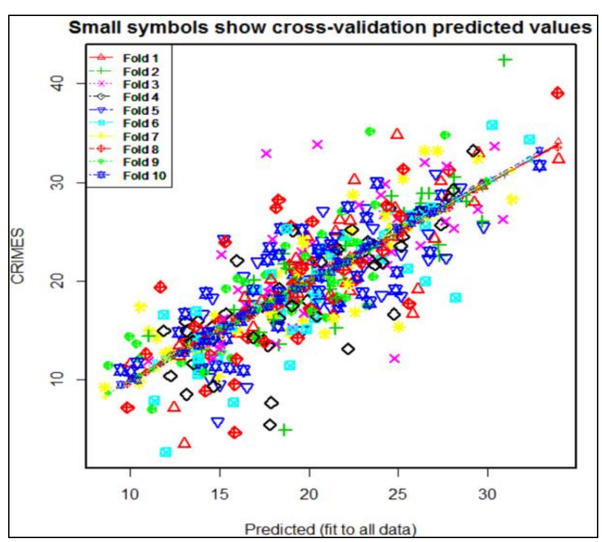
BEFORE **AFTER**





Validation:

According to the calculated values for the 10-fold cross validation, The overall MSE for all test-sets was pretty low with the average value of 16.3. This would indicate that our model does a decent job at predicting new observations without overfitting with the original data-set. Last, we've calculated a MAPE value of 19.6% indicating that on average our forecast is off by close to 20%. This isn't greatest result, but when considering the overall Adj-R2 of about 60% this level of error is to be expected.



Predictions:

Calculated predictions and confidence interval for every 1 unit of selected independent variables and below is the output from Predict functions.

```
> predict(M7,P1,interval="prediction",level=0.95)
   fit  lwr  upr
1 28.7 20.9 36.6
> predict(M7,P1,interval="confidence",level=0.95)
   fit  lwr  upr
1 28.7 27.6 29.9
>
```

Post applying inverse transformation,

Predicted value and Band

Fit	**Lwr**	**Upr**
158.5736	**98.09324**	**228.1409**

Confidence,

Fit	**Lwr**	**Upr**
158.5736	**149.0241**	**168.3211**

Visual plot of prediction vs Average value in the confidence Interval is added to Appendix.

Conclusions:

The best model as per our analysis to predict the rate of crimes found Region, Poverty, Total Population, per Capita income, Percentage of Beds in a County and percentage of population between ages 18-34 as the most significant predictors of crime rate. We did ponder over whether to include 'percentage of beds in a county' in our model as increase in hospital beds was effect of increased crime rates and not something that contributed to actual crime rates. But because results are proportional we decided to include it. Almost all the predictors were statistically highly significant at P-value<0.001 except for some regions. This brings us our attention to region- even though region has huge influence but some counties had extreme values compared to other counties having huge effect on the final outcome. Model would have been much more accurate if we could have treated each region separately.

The overall MSE value for test-sets was pretty low averaging around 16.3. The model forecast based on MAPE score was off by 19.6%. We had multi-collinearity issue with **total_Pop** and **Pers_income,** so we eliminated **total_Pop** and retained **Pers_income** to fix the issue. The distribution of crime rate and other predictor variables were not symmetric, so to improve linearity we performed Box Cox power transformation on crime rates and the independent variables. The transformation also minimized the effect of the extreme outlier we had. Further after eliminating the influential point "R-Student vs Hat Values" residual plots improved significantly, which decreased the number of possible outliers and improved our Adj-R2 value in our final model to 61.77 %.

Appendix:

Source Code:

```
################### Create new model without Influential
point.######################
setwd("C:/Users/magic_000/Desktop/CSC 423 Data Analysis & Regression/Final
Project/CrimeratesProject")
dat<-read.table("Crimerates_data.txt",header=T)
attach(dat)


NE=(region==1)*1
NC=(region==2)*1
S=(region==3)*1
dat_new<-cbind(dat[-length(dat)],NE,NC,S)


dat_final=subset(dat_new,dat_new$CRIMES<250)
#attach(dat_final)
NE=dat_final$NE
S=dat_final$S
NC=dat_final$NC
dat_new3bc<-dat_final[c(-1,-2,-3,-16,-17,-18,-19)]

#Transformation
library(car)
bc<-powerTransform(dat_new3bc)
summary(bc)
Tform=bcPower(dat_new3bc, bc$lambda)
#New Dataset and Model
dat_new5=cbind(Tform,NE,NC,S)
names(dat_new5)[1:12]<-c("Land","total_Pop","Pop18_34","Pop65plus","DOCS","BEDS","CRIME
S","Hsgrads","Bgrads","poverty","unemp","Pcimcome")
M6<-lm(CRIMES~Land+total_Pop+Pop18_34+Pop65plus+DOCS+BEDS+Hsgrads+Bgrads+pov
erty+unemp+Pcimcome+NE+NC+S,data=dat_new5)

#############HISTOGRAM & DENSITY #####################
plot(density(dat_new5$CRIMES),main="Normal Plot vs Transformed CRIMES
Density",xlab="",ylab="",ylim=c(0,0.07),xlim=c(0,55))
par(new=T)    #Didnt execute
plot(density(rnorm(100000,mean(dat_new5$CRIMES),sd(dat_new5$CRIMES))),lty=2,col="red",
main="",xlab="",ylim=c(0,0.07),xlim=c(0,55))
sum_stat(dat_new5$CRIMES) #Didnt execute

###########SCATTER PLOTS#################
pairs(dat_new5[c(7,1,2,3)])
pairs(dat_new5[c(7,4,5,6)])
pairs(dat_new5[c(7,8,9,10)])
```

```
cor(dat_new5)


####TESTING BEST FITTING MODEL####
#ADJ-R2
library (leaps)
leapmodels=leaps(x=dat_new5[names(dat_new5[-7])], y=dat_new5$CRIMES,
names=names(dat_new5[-7]), method="adjr2")
mat=cbind(leapmodels$size,leapmodels$which, leapmodels$adjr2)
mat[order(mat[,dim(mat)[2]], decreasing=T),]
Result= subset(data.frame(mat),data.frame(mat)$V16==max(data.frame(mat)$V16))

#Stepwise regression
Base = lm(CRIMES~1,data=dat_new5)
step(Base, scope = list(upper=M6, lower=~1 ), direction = "both", trace=FALSE)
summary(step(Base, scope = list(upper=M6, lower=~1 ), direction = "both", trace=FALSE))

#Backward selection
step(M6, direction = "backward")

#################Final Model######################

M7<-lm(CRIMES~total_Pop+Pop18_34+BEDS+poverty+Pcimcome+NE+NC+S,data=dat_new5
)
summary(M7)

#Standardize Model
M7_Sd<-lm(scale(CRIMES)~scale(total_Pop)+scale(Pop18_34)+scale(BEDS)+scale(poverty)+scal
e(Pcimcome)+NE+NC+S,data=dat_new5)
summary(M7_Sd)

#################Tesing Residuals#######################
plot(rstudent(M7)~fitted(M7),main="R-student vs Predicted Values")
abline(0,0,col="red",lty=2)
qqnorm(rstudent(M7))
qqline(rstudent(M7),col="red",lty=2)
plot(density(rnorm(100000)),main="",xlim=c(-5,5),ylim=c(0,0.55),col="red",lty=2,xlab="")
par(new=T)
plot(density(rstandard(M7)),xlim=c(-5,5),ylim=c(0,0.55),main="Residual Density vs Normal
Distribution",xlab="")


##################Influential Points Outliers####################
plot(rstudent(M7)~hatvalues(M7),main="R-student vs Hat Values")
abline(0,0,col="red",lty=2)
TST<-data.frame(cbind(influence.measures(M7)$infmat,rstudent(M7)))
names(TST)[15]<-"rstudent" #ERROR
```

```
pairs((TST[-c(1:12)]))
summary(influence.measures(M7))

##############################PREDICTION
INTERVAL######################################

#Prediction Average
set.seed(25)
P1<-
data.frame(total_Pop=sample(dat_new5$total_Pop,1),Pop18_34=sample(dat_new5$Pop18_34,1)
,BEDS=sample(dat_new5$BEDS,1),poverty=sample(dat_new5$poverty,1),Pcimcome=sample(d
at_new5$Pcimcome,1),NE=0,NC=0,S=1)
predict(M7,P1,interval="prediction",level=0.95)
#Average Expectancy
predict(M7,P1,interval="confidence",level=0.95)

plot(0,28.74306,col="red",pch=16,ylim=c(15,41),main="Confidence Interval Predict Value vs
Average Value")
abline(20.90535,0,col="red",lty=2)
abline(36.58077,0,col="red",lty=2)
abline(27.58365,0,col="blue",lty=3)
abline(29.90248,0,col="blue",lty=3)

##################CROSS VALIDATION ###################
library(DAAG)
set.seed(25)
cv.lm(data=dat_new5,M7,m=10,printit=T)
MAPE=(sum(abs((cv.lm(data=dat_new5,M7,m=2)$CRIMES-cv.lm(data=dat_new5,M7,m=2)$cv
pred)/cv.lm(data=dat_new5,M7,m=2)$CRIMES))*100)/length(dat_new5[,1])
MAPE
###########################################

sum_stat<- function(data) {
	len=length(data)
	m=mean(data)
	sdev=sd(data)
	quant=quantile(data)
	med=median(data)
	vary=var(data)
	maxi=max(data)
	mini=min(data)
	num=sum((data-mean(data))^3)/length(data)
	denom=(sum((data-mean(data))^2)/length(data))^(3/2)
	num_kurt=sum((data-mean(data))^4)/length(data)
	denom_kurt=(sum((data-mean(data))^2)/length(data))^2
	cat("Obs.",len, "\n")
	cat("Average",m, "\n")
```

```
        cat("Median",med, "\n")
        cat("Std. Deviation",sdev, "\n")
        cat("Variance",vary, "\n")
        cat("Skewness",num/denom,"\n")
        cat("Kurtosis",(num_kurt/denom_kurt)-3,"\n")
        cat("Range",maxi-mini, "\n")
        cat("Min",mini, "\n")
        cat("Max",maxi, "\n")
        OLL=as.vector((quant[2]-(1.5*(quant[4]-quant[2]))))
        OUL=as.vector(((1.5*(quant[4]-quant[2]))+quant[4]))
        cat("Outlier Lower Limit",OLL,"\n")
        cat("Outlier Upper Limit",OUL,"\n\n")
        dat=0
        for (i in 1:length(data)) {
        if (data[i]>OUL | data[i]<OLL){
        dat=append(dat,data[i])
        }}
        if (length(dat)==1){
                cat("No Outliers","\n\n")
                }
        else {
                dat=dat[-1]
                cat("There are",length(dat),"Outliers.",dat,"\n\n")
          }
        return(quant)

}

#If skewness is less than -1 or greater than 1, the distribution is highly skewed.
#If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
#If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

#The "minus 3" at the end of this formula is often explained as a correction to make the
#kurtosis of the normal distribution equal to zero,
#as the kurtosis is 3 for a normal distribution.
```

Mean, Median, Kurtosis data post power transformation of CRIMES variable.

```
Obs. 439
Average 19.94897
Median 19.24661
Std. Deviation 6.378685
Variance 40.68762
Skewness 0.2098744
Kurtosis -0.006795067
Range 39.75604
Min 2.640839
Max 42.39688
Outlier Lower Limit 1.911757
Outlier Upper Limit 37.70893

There are 2 Outliers. 39.04398 42.39688

       0%       25%       50%       75%      100%
 2.640839 15.335698 19.246606 24.284991 42.396883
```

Power transformation output.

```
> bc<-powerTransform(dat_new3bc)
Error in powerTransform(dat_new3bc) : object 'dat_new3bc' not found
>
> bc<-powerTransform(data_new3bc)
>
>
> summary(bc)
bcPower Transformations to Multinormality
          Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Land        -0.0096   0.0302          -0.0689           0.0497
total_Pop   -0.4975   0.0616          -0.6182          -0.3769
Pop18_34    -0.6179   0.1661          -0.9435          -0.2923
Pop65plus    0.0323   0.0787          -0.1219           0.1865
DOCS        -0.1707   0.0454          -0.2597          -0.0817
BEDS         0.3309   0.0512           0.2306           0.4311
CRIMES       0.6629   0.0696           0.5264           0.7994
Hsgrads      3.9935   0.3245           3.3573           4.6296
Bgrads       0.1744   0.0745           0.0283           0.3205
poverty      0.2223   0.0525           0.1194           0.3253
unemp       -0.1032   0.0874          -0.2746           0.0682
Pcimcome    -0.1092   0.1002          -0.3056           0.0871

Likelihood ratio tests about transformation parameters
                                                                      LRT df      pval
LR test, lambda = (0 0 0 0 0 0 0 0 0 0 0 0)                     481.470567 12 0.0000000
LR test, lambda = (1 1 1 1 1 1 1 1 1 1 1 1)                    3720.092990 12 0.0000000
LR test, lambda = (0 -0.5 -0.5 0 -0.17 0.33 0.66 3.99 0.17 0.22 0 0)   3.351849 12 0.9925089
>
```

Initial Model(M1)

```
> summary(M6)

Call:
lm(formula = CRIMES ~ Land + total_Pop + Pop18_34 + Pop65plus +
    DOCS + BEDS + Hsgrads + Bgrads + poverty + unemp + Pcimcome +
    NE + NC + S, data = data_new5)

Residuals:
     Min       1Q   Median       3Q      Max
-13.9420  -2.2048   0.1128   2.3413  14.9314

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.775e+03  3.265e+02  -8.497 3.33e-16 ***
Land        -4.915e-01  2.960e-01  -1.660 0.097580 .
total_Pop    1.271e+03  1.689e+02   7.526 3.17e-13 ***
Pop18_34     6.162e+01  1.833e+01   3.362 0.000845 ***
Pop65plus   -1.700e-01  1.016e+00  -0.167 0.867162
DOCS         7.166e-01  8.435e-01   0.850 0.396077
BEDS         1.113e+00  4.647e-01   2.395 0.017046 *
Hsgrads      1.566e-07  1.454e-07   1.077 0.282154
Bgrads      -1.506e+00  9.158e-01  -1.645 0.100728
poverty      3.433e+00  5.735e-01   5.986 4.59e-09 ***
unemp        1.168e+00  1.048e+00   1.114 0.265762
Pcimcome     2.568e+01  6.792e+00   3.781 0.000179 ***
NE          -5.885e+00  7.895e-01  -7.455 5.12e-13 ***
NC          -2.741e+00  7.510e-01  -3.649 0.000296 ***
S            1.620e+00  7.044e-01   2.299 0.021962 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.932 on 424 degrees of freedom
Multiple R-squared:  0.6322,    Adjusted R-squared:  0.6201
F-statistic: 52.07 on 14 and 424 DF,  p-value: < 2.2e-16
```

Stepwise regression with Adj-R2,

```
> summary(step(Base, scope = list(upper=M6, lower=~1 ), direction = "both", trace=FALSE))

Call:
lm(formula = CRIMES ~ poverty + total_Pop + S + NE + Pcimcome +
    Pop18_34 + NC + BEDS + Land + Bgrads, data = data_new5)

Residuals:
     Min       1Q   Median       3Q      Max
-14.0255  -2.2588   0.1373   2.3200  14.8244

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2848.4022   311.1707  -9.154  < 2e-16 ***
poverty         3.3986     0.4561   7.452 5.13e-13 ***
total_Pop    1308.7303   163.9864   7.981 1.35e-14 ***
S               1.2417     0.6428   1.932 0.054033 .
NE             -6.0355     0.7172  -8.415 5.93e-16 ***
Pcimcome       25.8946     6.0684   4.267 2.44e-05 ***
Pop18_34       60.5780    14.2002   4.266 2.45e-05 ***
NC             -2.8998     0.7202  -4.026 6.70e-05 ***
BEDS            1.2553     0.3213   3.907 0.000109 ***
Land           -0.4993     0.2940  -1.698 0.090142 .
Bgrads         -0.9271     0.5815  -1.594 0.111641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.926 on 428 degrees of freedom
Multiple R-squared:  0.6298,    Adjusted R-squared:  0.6212
F-statistic: 72.82 on 10 and 428 DF,  p-value: < 2.2e-16

> |
```

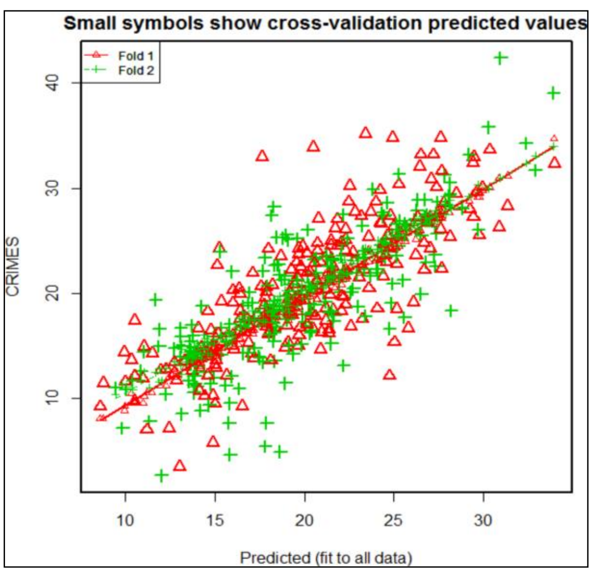
Backward selection,

```
Step:  AIC=1211.62
CRIMES ~ Land + total_Pop + Pop18_34 + BEDS + Bgrads + poverty +
    Pcimcome + NE + NC + S

            Df Sum of Sq    RSS    AIC
<none>                   6596.8 1211.6
- Bgrads     1     39.17 6636.0 1212.2
- Land       1     44.46 6641.3 1212.6
- S          1     57.52 6654.3 1213.4
- BEDS       1    235.23 6832.1 1225.0
- NC         1    249.88 6846.7 1225.9
- Pop18_34   1    280.50 6877.3 1227.9
- Pcimcome   1    280.65 6877.5 1227.9
- poverty    1    855.97 7452.8 1263.2
- total_Pop  1    981.69 7578.5 1270.5
- NE         1   1091.54 7688.4 1276.8

Call:
lm(formula = CRIMES ~ Land + total_Pop + Pop18_34 + BEDS + Bgrads +
    poverty + Pcimcome + NE + NC + S, data = data_new5)

Coefficients:
(Intercept)         Land    total_Pop     Pop18_34         BEDS       Bgrads      poverty     Pcimcome           NE           NC            S
 -2848.4022      -0.4993    1308.7303      60.5780       1.2553      -0.9271       3.3986      25.8946      -6.0355      -2.8998       1.2417
```

Standardized Model,

```
> M7_Sd<-lm(scale(CRIMES)~scale(total_Pop)+scale(Pop18_34)+scale(BEDS)+scale(poverty)+scale(Pcimcome)+NE+NC+S,data=data_new5)
> summary(M7_Sd)

Call:
lm(formula = scale(CRIMES) ~ scale(total_Pop) + scale(Pop18_34) +
    scale(BEDS) + scale(poverty) + scale(Pcimcome) + NE + NC +
    S, data = data_new5)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15338 -0.35772  0.01772  0.38577  2.40441

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.17968    0.07598   2.365 0.018478 *
scale(total_Pop)  0.28306    0.03647   7.762 6.14e-14 ***
scale(Pop18_34)   0.13879    0.02999   4.628 4.90e-06 ***
scale(BEDS)       0.15589    0.03999   3.898 0.000113 ***
scale(poverty)    0.46063    0.05851   7.872 2.86e-14 ***
scale(Pcimcome)   0.26343    0.05503   4.787 2.33e-06 ***
NE               -0.83326    0.10258  -8.123 4.88e-15 ***
NC               -0.33610    0.10172  -3.304 0.001032 **
S                 0.27901    0.09110   3.063 0.002332 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6183 on 430 degrees of freedom
Multiple R-squared:  0.6247,    Adjusted R-squared:  0.6177
F-statistic: 89.48 on 8 and 430 DF,  p-value: < 2.2e-16

> |
```

2-Fold Cross Validation:



Additional Predictions: -

Important factors for the model are as listed below,

1) Region
2) Poverty
3) Total Population
4) Income per Capita
5) Percentage of Beds in a County
6) Percentage of Population (18-34)

Region was chosen as a group since it has huge influence on crime rate more than all other variables. Overall, the top three factors seem to make a lot of sense on how that would play a big part in predicting Crime Rate. However, it's hard to make sense of how the number of beds are important to this analysis. Last, it's surprising that percentage of population age was the least influential on predicting Crime Rate. Especially since it's effect is lower than percentage of beds.