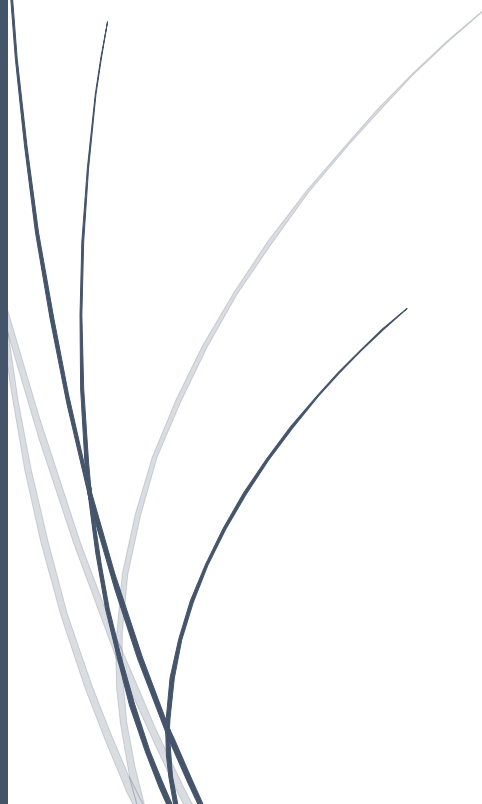


A dark blue vertical bar on the left side of the slide, with a blue arrow pointing right from it.

2017-3-16

# Crime Rate Analysis

CSC 423

Several thin, curved lines in dark blue and light grey originating from the bottom left corner.

Gao Hang, Divyata Patil, Wenyi Yan, Junzhe Yu

# Contents

Non-Technical Summary

## **1.Exploratory data analysis** 2

1.1 Distribution of Y

1.2 Linear Association Exploration

## **2. Modeling**

2.1 Fitting the model

2.2 Selection Method

## **3. Diagnostics**

3.1 Residual Analysis

3.2 Outliers and Influential Points

## **4. Relationship and Associations**

4.1 Standardized Coefficients

## **5.Appendix**

5.1Code

5.2Outliers and Influential points

## Non-Technical Summary

As a part of our data analysis project we are analyzing data pertains to the years 1990 and 1992 to predict the serious crimes using county demographic information for 440 most populous counties in the US. Goal of this analysis is to predict the crimes rate based on the information we have from crimes rate dataset.

Our data set contain around 439 records with and 16 variables in total which will help us to analyze the crimes rates in different regions like north central, north eastern south and west. Variables are nothing but the factors like total population, population between specific age group, income between specific age group, how well people are educated etc. Our goal is to find out associations between these variables against our response variable crimes rate. From regression analysis we are going to find out which factors or variables are having significant impact on crimes rate and how can we predict the crimes rate based on these variables.

After performing the required statistics tests and analyzing results we reached up to the conclusion that there are few variables which re having very significant impact on predicting crimes rate but few variables have absolutely no impact on prediction of crimes rate. Eventually we dropped the variables which does not poses any value in our analysis. Our final model included following variables:

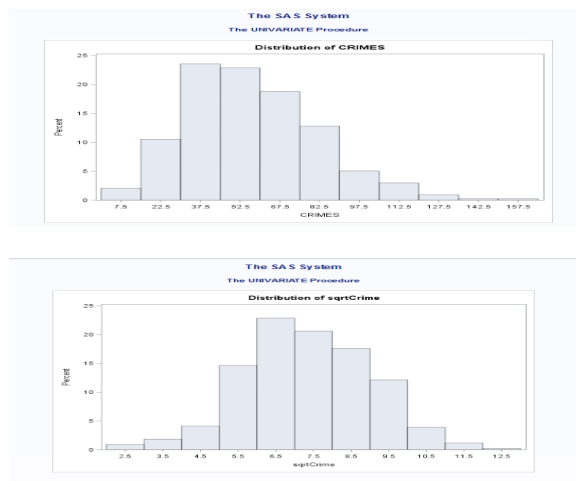
- Estimated total population 1990
- Rate of beds, cribs and bassinets per 1000 population
- Percent of adult population (25 years old or older) with bachelor's degree
- Percent of 1990 population with income below poverty
- Per capita income of 1990 population (dollars)
- geographic region north eastern and north central

among these variables poverty, Rate of beds, cribs and bassinets per 1000 population and per capita income has strong association with crimes rates. It implies that these are the major factors involved in our prediction about crimes rate.

# 1.Exploratory data analysis

## 1.1 Distribution of Y

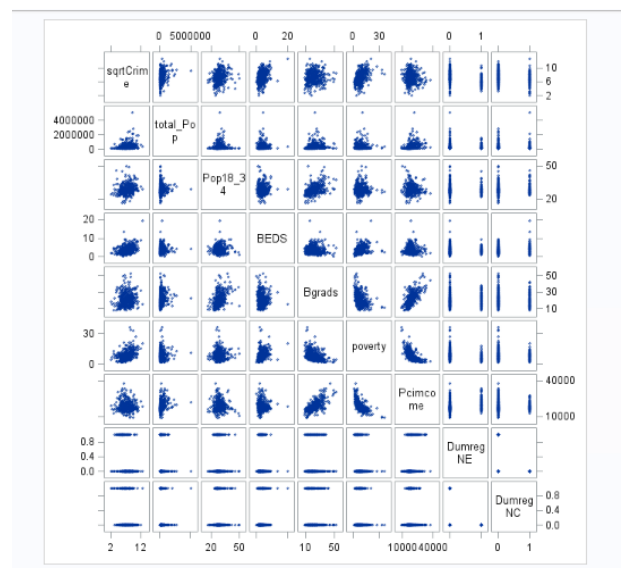
First, from the data results, we interpret that the distribution of Y variable is not normally distributed. (Skewed on the left side, indicating a necessity for a transformation.)



Accordingly, we transformed the Y variable using **sqrt(crimes)**. Histogram below is the transformed Y, which has a symmetric display and normal distribution.

## 1.2 Linear Association Exploration

As we can see from the scatterplot below, the linear relationship between dependent variable(crimes) with the independent variables is not clear. None of the chart show any relationship we could dig in. As result, our data analysis of crimes continues with other mathematic method instead of observation of data.



## 2. Modeling

## 2.1 Fitting the model

After transformation, first, We did regression analysis to the full model to find if there are some problems:

The SAS System

The REG Procedure  
Model: MODEL1  
Dependent Variable: sqrtCrime

Number of Observations Read	439
Number of Observations Used	439

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	761.55940	50.77063	38.22	<.0001
Error	423	561.96615	1.32853		
Corrected Total	438	1323.52555			

Root MSE	1.15262	R-Square	0.5754
Dependent Mean	7.36376	Adj R-Sq	0.5603
Coeff Var	15.65255		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	-1.37740	1.82307	-0.76	0.4503	0
Land	1	-0.00010324	0.00004397	-2.35	0.0193	1.52136
total_Pop	1	0.00000629	8.165432E-7	7.70	<.0001	43.75354
Pop18_34	1	0.08799	0.02124	4.14	<.0001	2.61374
Pop65plus	1	0.01173	0.01982	0.59	0.5541	2.06664
DOCS	1	-0.00132	0.06579	-0.02	0.9840	3.35983
BEDS	1	0.18710	0.05171	3.62	0.0003	3.53748
Hsgrads	1	0.01463	0.01737	0.84	0.4002	4.89624
Bgrads	1	-0.03639	0.01927	-1.89	0.0597	7.19046
poverty	1	0.13244	0.02499	5.30	<.0001	4.46932
unemp	1	0.03163	0.03448	0.92	0.3595	2.14525
Pcincome	1	0.00021571	0.00003468	6.22	<.0001	6.54461
Pers_income	1	-0.00025895	0.00003951	-6.55	<.0001	48.99099
DumregNE	1	-1.74843	0.21774	-8.03	<.0001	2.81344
DumregNC	1	-1.07506	0.21050	-5.11	<.0001	2.71589
DumregS	1	0.10589	0.20391	0.52	0.6038	3.11016

According to the result, we can see that VIF values of Total\_pop and Pers\_income higher than 10, which means they have multicollinearity problem. So, we decide to remove them one by one to see if VIF would be normal. Thus, we g

**M1(remove Total\_pop):**

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	0.30679	1.93036	0.16	0.8738	0
Land	1	-0.00003743	0.00004600	-0.81	0.4163	1.46391
Pop18_34	1	0.08289	0.02264	3.66	0.0003	2.61120
Pop65plus	1	0.00299	0.02110	0.14	0.8874	2.05987
DOCS	1	-0.00447	0.07017	-0.06	0.9492	3.35970
BEDS	1	0.21334	0.05503	3.88	0.0001	3.52212
Hsgrads	1	0.01943	0.01852	1.05	0.2947	4.88995
Bgrads	1	-0.03482	0.02055	-1.69	0.0910	7.18967
poverty	1	0.13497	0.02665	5.07	<.0001	4.46855
unemp	1	0.03828	0.03676	1.04	0.2983	2.14390
Pcincome	1	0.00010665	0.00003377	3.16	0.0017	5.45360
Pers_income	1	0.00004080	0.00000722	5.65	<.0001	1.43943
DumregNE	1	-1.65136	0.23185	-7.12	<.0001	2.80401
DumregNC	1	-1.00552	0.22430	-4.48	<.0001	2.71089
DumregS	1	0.20776	0.21703	0.96	0.3390	3.09707

**M2(remove Pers\_income):**

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	1	0.28910	1.89249	0.15	0.8787	0
Land	1	-0.00005515	0.00004545	-1.21	0.2256	1.47901
total_Pop	1	0.00000102	1.46727E-7	6.92	<.0001	1.28554
Pop18_34	1	0.07997	0.02223	3.60	0.0004	2.60507
Pop65plus	1	0.00287	0.02072	0.14	0.8901	2.05701
DOCS	1	-0.00982	0.06895	-0.14	0.8868	3.35853
BEDS	1	0.21564	0.05402	3.99	<.0001	3.51239
Hsgrads	1	0.01929	0.01820	1.06	0.2899	4.88806
Bgrads	1	-0.03275	0.02019	-1.62	0.1056	7.18451
poverty	1	0.12920	0.02619	4.93	<.0001	4.46757
unemp	1	0.04035	0.03612	1.12	0.2646	2.14206
Pcincome	1	0.00010987	0.00003218	3.41	0.0007	5.12604
DumregNE	1	-1.65797	0.22781	-7.28	<.0001	2.80214
DumregNC	1	-1.01894	0.22049	-4.62	<.0001	2.71139
DumregS	1	0.20268	0.21321	0.95	0.3423	3.09385

We can see that VIF values are normal in these two models. So next, we use Backward selection method to both of them and compare which one is better.

## **2.2 Selection Method**

After analyzing forward and backward Cp selection methods we decided to chose backward selection as it does matches results with our initial analysis and it fulfills all statistic requirements.

## S1(Selection result of M1):

Backward Elimination: Step 6

Variable DumregS Removed: R-Square = 0.5118 and C(p) = 6.5986

Bounds on condition number: 4.6103, 144.38

All variables left in the model are significant at the 0.1000 level.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	677.31950	84.66494	56.34	<.0001
Error	430	646.20604	1.50280		
Corrected Total	438	1323.52555			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.13085	0.77709	11.29967	7.52	0.0064
Pop18_34	0.08213	0.01919	27.51814	18.31	<.0001
BEDS	0.21329	0.03357	60.65107	40.36	<.0001
Bgrads	-0.02979	0.01510	5.84944	3.89	0.0491
poverty	0.12814	0.01899	68.40409	45.52	<.0001
Pcincome	0.00010816	0.00003096	18.34302	12.21	0.0005
Pers_income	0.00003839	0.00000701	45.04664	29.98	<.0001
DumregNE	-1.74140	0.16217	173.29325	115.31	<.0001
DumregNC	-1.07834	0.15505	72.68816	48.37	<.0001

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DOCS	13	0.0000	0.5159	13.0041	0.00	0.9492
2	Pop65plus	12	0.0000	0.5158	11.0245	0.02	0.8864
3	Land	11	0.0007	0.5151	9.6765	0.66	0.4187
4	unemp	10	0.0012	0.5139	8.7047	1.03	0.3099
5	Hsgrads	9	0.0009	0.5130	7.5029	0.80	0.3709
6	DumregS	8	0.0013	0.5118	6.5986	1.10	0.2944

## S2(Selection result of M2):

Backward Elimination: Step 6

Bounds on condition number: 4.2836, 140.42

Variable Land Removed: R-Square = 0.5268 and C(p) = 7.9664

All variables left in the model are significant at the 0.1000 level.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	697.22707	87.15338	59.84	<.0001
Error	430	626.29847	1.45651		
Corrected Total	438	1323.52555			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.07096	0.74862	11.14653	7.65	0.0059
total_Pop	9.488373E-7	1.420838E-7	64.95421	44.60	<.0001
Pop18_34	0.07973	0.01888	25.97420	17.83	<.0001
BEDS	0.21406	0.03303	61.17677	42.00	<.0001
Bgrads	-0.02851	0.01485	5.36858	3.69	0.0555
poverty	0.12286	0.01870	62.89309	43.18	<.0001
Pcincome	0.00011216	0.00002938	21.22746	14.57	0.0002
DumregNE	-1.73316	0.15965	171.64595	117.85	<.0001
DumregNC	-1.07604	0.15264	72.37923	49.69	<.0001

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Pop65plus	13	0.0000	0.5323	13.0191	0.02	0.8901
2	DOCS	12	0.0000	0.5322	11.0400	0.02	0.8851
3	DumregS	11	0.0011	0.5311	10.0388	1.00	0.3171
4	Hsgrads	10	0.0007	0.5304	8.6783	0.64	0.4232
5	unemp	9	0.0009	0.5296	7.4586	0.78	0.3763
6	Land	8	0.0028	0.5268	7.9664	2.52	0.1130

Based on the result, we found that S2 has higher R-square value than S1, which means independent variables in S2 can explain more of the variability in Sqrt(crimes) than S1. S2 also has higher F-value, indicating it has a stronger support than S1. As for p values, each variable of these two models is under or not much beyond the .05, which means these variables have significant effects to Sqrt(crimes).

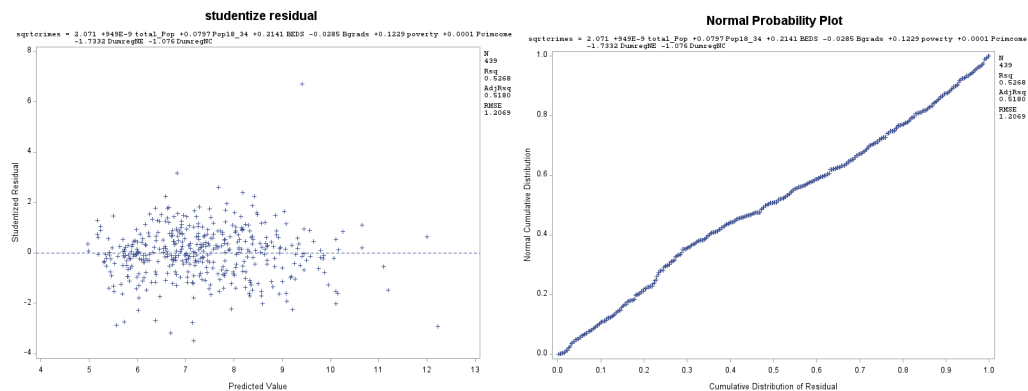
So, we choose S2 as our final model, which is:

$$\begin{aligned} \text{sqrt}(\text{crime}) = & 2.071 + 949\text{E-}9 \text{Total\_pop} + 0.0797 \text{Pop18\_34} + 0.2141 \text{Beds-} \\ & 0.0285 \text{Bgrads} + 0.1229 \text{Poverty} + 0.0001 \text{Pcincome} - 1.7332 \text{DumregNe-} \\ & 1.076 \text{DumregNC} \end{aligned}$$

### 3. Diagnostics

#### 3.1 Residual Analysis

From studentized residual vs predicted values, we can see that points are almost randomly scattered inside a band centered around the horizontal line at zero, so regression model is valid for the data and model assumptions almost hold.



From normal probability plot of residuals, we can see that and points lie close to a line which means the errors can be assumed to be approximately normal, and assumption of normality is satisfied.

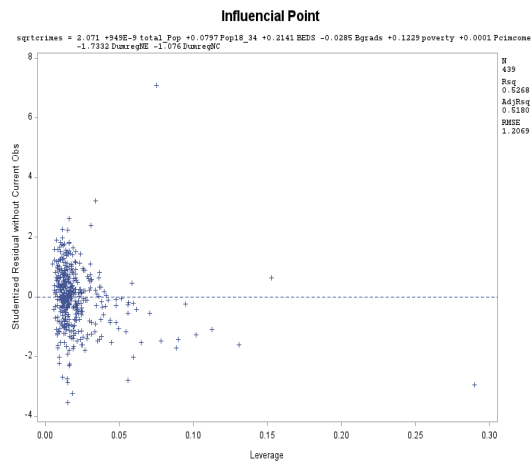
#### 3.2 Outliers and Influential Points

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITs
1	-2.9700	-2.9458	0.2896	1.2004	-1.8810
2	-1.7021	-1.4708	0.0781	1.0586	-0.4280
3	-1.7709	-1.5197	0.0649	1.0405	-0.4002
4	-2.3630	-2.0261	0.0594	0.9965	-0.5094
5	7.7879	7.0825	0.0749	0.4080	2.0156
6	-0.3550	-0.3011	0.0476	1.0702	-0.0673

279	0.8827	0.7362	0.0141	1.0240	0.0879
280	-1.7676	-1.4827	0.0215	0.9967	-0.2200
281	3.1228	2.6266	0.0162	0.8992	0.3371
282	1.8656	1.5593	0.0140	0.9843	0.1856
283	-0.0231	-0.0192	0.0149	1.0366	-0.0024
284	1.5522	1.2951	0.0123	0.9982	0.1445



When we performed test to check influential points we found that there were two influential points. In order to remove these two points we removed 5<sup>th</sup> and 281<sup>st</sup> rows from our data. After removing the influential points we observed significant increase in R value that is 5%. Cooks D value is also less than 1 which indicates we have successfully dealt with influential points.



Basing on the analysis(see 5.2 in appendix for the complete result), Because there are 4 point's studentized residual higher than 3, so there is 4 outlier in this mode.

## 4. Relationship and Associations

### 4.1 Standardized Coefficients

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate
Intercept	1	2.07096	0.74862	2.77	0.0059	0
total_Pop	1	9.488373E-7	1.420838E-7	6.68	<.0001	0.24352
Pop18_34	1	0.07973	0.01888	4.22	<.0001	0.19230
BEDS	1	0.21406	0.03303	6.48	<.0001	0.24668
Bgrads	1	-0.02851	0.01485	-1.92	0.0555	-0.12569
poverty	1	0.12286	0.01870	6.57	<.0001	0.32937
Pcincome	1	0.00011216	0.00002938	3.82	0.0002	0.26211
DumregNE	1	-1.73316	0.15965	-10.86	<.0001	-0.42299
DumregNC	1	-1.07604	0.15264	-7.05	<.0001	-0.26691

We see that Poverty has the highest standardized coefficients value, which means Poverty and Pcincome is predictor that has the strongest influence on Sqrt(crime). In addition, Beds, Total\_Pop also have fairly strong influence on Sqrt(crime).

### 4.2 Correlation Values

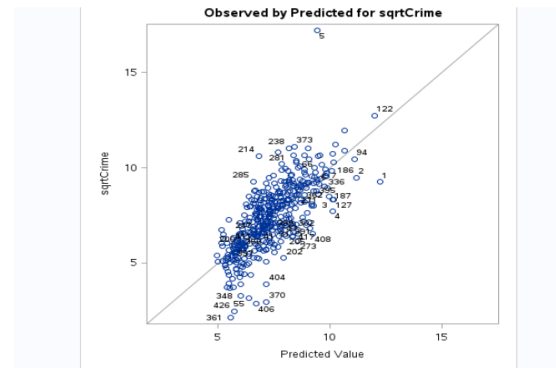
Pearson Correlation Coefficients, N = 429 Prob >  r  under H0: Rho=0								
	sqrtrcrimes	total_Pop	Pop18_34	BEDS	Bgrades	poverty	Pcincome	DumregNE
sqrtrcrimes	1.00000	0.32547 <.0001	0.19620 <.0001	0.35721 <.0001	0.04944 <.0001	0.46328 <.0001	-0.07986 <.0001	-0.37654 <.0001
total_Pop	0.32547 <.0001	1.00000	0.06936 <.0001	0.03879 <.0001	0.19142 <.0001	0.02454 <.0001	0.29459 <.0001	0.02748 <.0001
Pop18_34	0.19620 <.0001	0.06936 <.0001	1.00000	0.03005 <.0001	0.45617 <.0001	0.03283 <.0001	-0.03274 <.0001	-0.07820 <.0001
BEDS	0.35721 <.0001	0.03879 <.0001	0.03005 <.0001	1.00000	-0.04533 <.0001	0.37196 <.0001	-0.05324 <.0001	-0.04297 <.0001
Bgrades	0.04944 <.0001	0.19142 <.0001	0.45617 <.0001	-0.04533 <.0001	1.00000	-0.40884 <.0001	0.69542 <.0001	0.05108 <.0001
poverty	0.46328 <.0001	0.02454 <.0001	0.03283 <.0001	0.37196 <.0001	-0.40884 <.0001	1.00000	-0.00297 <.0001	-0.26002 <.0001
Pcincome	-0.07986 <.0001	0.29459 <.0001	-0.03274 <.0001	-0.05324 <.0001	0.69542 <.0001	-0.00297 <.0001	1.00000	0.27867 <.0001
DumregNE	-0.37654 <.0001	0.02748 <.0001	-0.07820 <.0001	-0.04297 <.0001	0.05108 <.0001	-0.26002 <.0001	0.27867 <.0001	1.00000
DumregNC	-0.13854 <.0001	-0.03531 <.0001	-0.06609 <.0001	0.13908 <.0001	-0.09727 <.0001	-0.09545 <.0001	-0.31626 <.0001	0.00000

From above SAS output it is evident that independent variables present in model have pretty good correlation with response variable crime.

## 5. Prediction Analysis:

The following figure shows the predicted value versus observed value using the variable. In this plot, each point is one observation, where the prediction made by the model is on the x-axis, and the accuracy of the prediction is on the y-axis. The distance from the line at 0 is how bad the

prediction was for that value.



The model for predicting crime rate is quite accurate, there's a strong correlation between the model's predictions and its actual results. Although, we have an outlier of observation number 5 which has quite a large prediction error. This large residual error might have resulted from abnormal predictor variables. We also predicted values using a fitted regression model with 95% confidence interval. The following figure shows the prediction values along with their 95% confidence bound.

The REG Procedure  
Model: MODEL1  
Dependent Variable: sqrtCrime

Output Statistics					
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	Residual
1	9.25	12.2214	0.6495	10.9448 13.4980	-2.9700
2	9.48	11.1868	0.3372	10.5240 11.8496	-1.7021
3	8.34	10.1125	0.3074	9.5084 10.7166	-1.7709
4	7.74	10.1060	0.2942	9.5277 10.6843	-2.3630
5	17.20	9.4163	0.3303	8.7670 10.0656	7.7879
6	9.15	9.5031	0.2634	8.9854 10.0208	-0.3550
7	9.58	9.7894	0.2867	9.2259 10.3529	-0.2051
8	11.24	10.2542	0.2306	9.8009 10.7075	0.9857
9	10.75	10.1127	0.2118	9.6965 10.5290	0.6408
10	6.30	8.9994	0.2387	8.5301 9.4686	-0.7025
11	9.11	8.8166	0.1859	8.4511 9.1821	0.2884
12	7.17	9.0794	0.1914	8.7031 9.4556	-1.9084
13	7.65	9.0021	0.1897	8.6292 9.3751	-1.3474
14	7.20	8.8079	0.1918	8.4308 9.1849	-1.6106
15	5.06	7.1735	0.1980	6.7844 7.5626	-2.1121
16	6.13	7.7902	0.2024	7.3923 8.1881	-1.6622
17	7.10	6.9197	0.1812	6.5635 7.2759	0.1850

From the adobe figure we can also see the residual error for observation number is 7.789 which is very high corresponding to other observation.

## 6. Conclusion:

In this project, we did the data analysis to predict the rates of serious crimes using county demographic information for 440 most populous counties in US. First, we checked if distribution output variable follow a normal distribution and therefore transformed output variable by square root transformation. Secondly, we checked the multicollinearity or existence of any dependent variable by checking variance inflation factor and we remove total\_pop and Pers\_income based on variance inflation factor one by one to and get two new models. Then we performed different linear regression model to two models and found the best fitted model based on Backward selection. The final model includes only 7 variables total\_Pop, Pop18\_34, BEDS, Bgrads, poverty, Pcincome, DumregNE and DumregNC. From the diagnostic analysis of this predicted model, we have found that residuals are normally distributed and Square value represents good accuracy of the prediction model. Finally, we checked the prediction analysis to see how our predictive model performs and we found that our final predicted model performed very well with respect to observed value. In conclusion, we found that the regression model, crime rate of a certain area has positive correlation with predictors total population, percent of 1990 population aged 18-34, rate of beds, cribs and bassinets per 1000 population, Per capita income of 1990 population (dollars)and negative correlation with predictors Percent of adult population (25 years old or older) with bachelor's degree, DumregNE and DumregNC.

## 5.Appendix

### 5.1Code

```
*/ histogram before transformation;

proc univariate ;

    histogram crimes;

run;

*/ transformation process;

data crimes;

set crimes;

DumregNE = (region=1);

DumregNC = (region=2);

DumregS = (region=3);

sqrtCrime = sqrt(crimes);

run;

*/ histogram after transformation;

proc univariate ;

    histogram sqrtCrime;

run;

*/ Scatter plot after transformation ;

proc sgscatter;

matrix sqrtCrime  total_Pop  Pop18_34  BEDS

        Bgrads poverty  Pcimcome

*/ regression analysis for full model ;

PROC REG;

MODEL CRIMESnew = Land total_Pop Pop18_34 Pop65plus DOCS BEDS

        Hsgrads Bgrads poverty unemp Pcimcome Pers_income

        DumregNE DUMREGNC DUMREGS/vif;
```

```

RUN;

*/ regression analysis for M1;

PROC REG;

MODEL sqrtrimes = Land Pop18_34 Pop65plus DOCS BEDS

      Hsgrads Bgrads poverty unemp Pcimcome Pers_income

      DumregNE DUMREGNC DUMREGS/vif;

RUN;

*/ regression analysis for M2;

PROC REG;

MODEL sqrtCRIMES = Land total_Pop Pop18_34 Pop65plus DOCS BEDS

      Hsgrads Bgrads poverty unemp Pcimcome

      DumregNE DUMREGNC DUMREGS/vif;

RUN;

*/ S1(backward selection to M1);

PROC REG;

MODEL sqrtCRIMES = Land Pop18_34 Pop65plus DOCS BEDS

      Hsgrads Bgrads poverty unemp Pcimcome Pers_income

      DumregNE DUMREGNC DUMREGS/selection=b;

RUN;

*/ S2(backward selection to M2);

PROC REG;

MODEL sqrtCRIMES = Land total_pop Pop18_34 Pop65plus DOCS BEDS

      Hsgrads Bgrads poverty unemp Pcimcome

      DumregNE DUMREGNC DUMREGS/SELECTION=B;

RUN;

*/ Residual analysis;

plot residual.*predicted.;

plot student.*predicted.;

title "studentize residual";

run;

```

```

plot npp.*residual.;

plot npp.*student.;

title "Normal Probability Plot";

run;

*/ check for outliers and influential points;

model sqrtcrime=total_pop pop18_34 BEDS bgrads poverty pcimcome dumregne dumregnc/influence r;

plot rstudent.*h.;

title "Influencial Point";

run;

*/ standardized coefficients;

proc reg data=crimerates;

model sqrtCrime = total_Pop Pop18_34 BEDS

Bgrads poverty Pcimcome

DumregNE DumregNC /stb;

*/ correlation values;

PROC CORR;

VAR sqrtcrime total_pop Pop18_34 BEDS Bgrads poverty Pcimcome

DumregNE DUMREGNC;

run;

*/ K fold cross validation ;

proc glmselect data = logistic plots = (asePlot criteria);

model sqrtCrime = total_Pop Pop18_34 BEDS

Bgrads poverty Pcimcome

DumregNE

DumregNC /selection = backward (stop = cv) cvmethod = split(5) cvdetails = all ;

run;

*/ Influential calcualtion of Data;

ods graphics on;

proc reg data=crimes

plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);

```

```
model sqrtcrime= total_Pop Pop18_34 BEDS Bgrads poverty Pcincome DumregNE DumregNC/clm;
```

```
run;
```

```
ods graphics off;
```

```
*/ Prediction for all
```

## 5.2 Outliers and Influential points

