Michigan Technological University

“Factors Affecting Crime in Small Cities”

Submitted by:-

Pranav Anand Joshi

Student: Data Science

Pjoshi3@mtu.edu

Submitted to:-

Prof. Min Wang Ph.D.

Faculty: Mathematical Sciences

Statistical Programming & Analysis

Abstract

Over the last few decades, crime increases drastically and becomes a major problem not only in big cities but in small cities as well. This inspires me to analyze on the basis of data sets and find out the major factors that foster crime. The datasets is chosen from

(http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr06.html)

Here, in this project, we are using multiple linear regression analysis between reported overall crime rate and some important predictor variables like violent crime rate, annual police funding etc in order to investigate profoundly the condition and predict the upcoming condition.

Introduction and description of variables

Here, we are providing the brief introduction and describing the each variable of dataset which is really important before the analysis. Data were obtained from the source: “Life in America's Small Cities”, By G.S. Thomas. This data contains 50 observations that include many predictor variables.

Variable Name	Description
Y	Total overall reported crime rate per 1 million
X1	Reported violent crime rate per 100,000 residents
X2	Annual police funding in $/resident
X3	% of people 25 years+ with 4 yrs. of high school
X4	% of 16 to 19 yr not in highschool & not highschool graduates.
X5	% of 18 to 24 year-olds in college
X6	% of people 25 years+ with at least 4 years of college

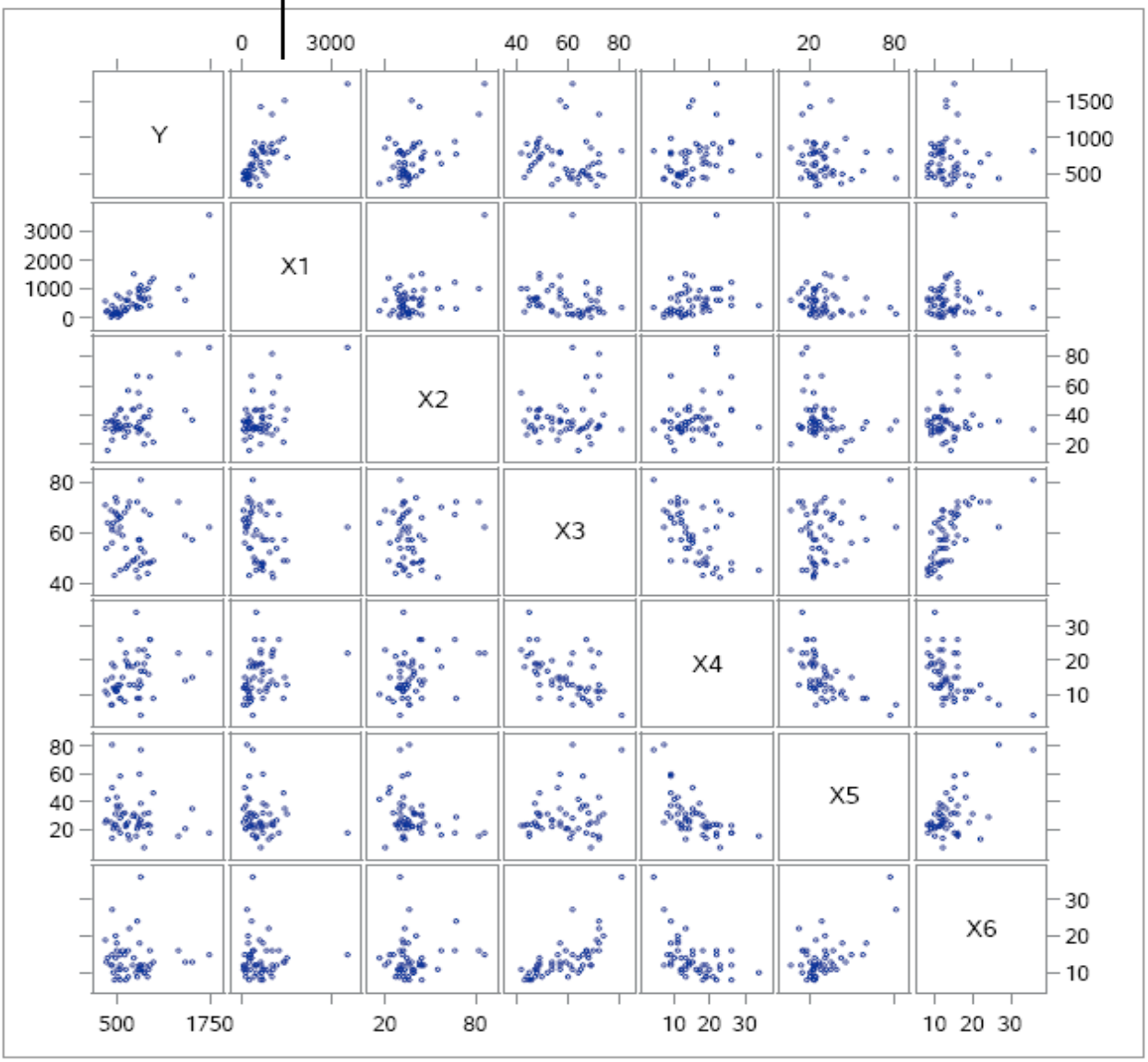
The main objective of this project is to find out the relationship between crime in small cities and the various explanatory variables using multiple regression analysis.

Main Objective:-

"To find the factors that affect crime in small cities and identify the suitable regression model for that which best estimates the crime."

Here, our predictors are X1 to X6 and we are going to predict Y. Now, let's see what the *Scatter plot matrix* says:-

Y & X1 seems linear



As we can see in the above scatter plot, it is clearly shown that there is some linear relationship in between Y and X1. However, the other predictors X2, X3, X4, X5, and X6 don't show any linear relationship with Y. As we have 6 predictors here. So, our model will be:-

$$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4 + B_5 X_5 + B_6 X_6$$

Now, we will check which of this predictor variable are useful in predicting Y. As shown below on parameter estimates, ***Estimated Linear Regression*** model is:-

$$Y = 100.39 + 0.33 X_1 + 4.0 X_2 + 1.86 X_3 + 7.84 X_4 + 2.56 X_5 - 3.23 X_6$$

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	100.39361	370.69317	0.27	0.7878
X1	1	0.33234	0.05962	5.57	<.0001
X2	1	3.99817	2.68248	1.49	0.1434
X3	1	1.85791	5.24087	0.35	0.7247
X4	1	7.83886	7.75987	1.01	0.3181
X5	1	2.55877	3.42695	0.75	0.4593
X6	1	-3.23116	10.71537	-0.30	0.7645

Now, we should look at ***R-squared*** value. Higher the R squared, the more variation is explained by our input variables. But, in case of multiple variable, we look at adjusted R-squared value to check whether the addition of variables is useful or not. R-squared value always increases by adding a variable; however, Adjusted R-square depends on the added variable.

Root MSE	195.15783	**R-Square**	0.6132
Dependent Mean	717.96000	**Adj R-Sq**	0.5592
Coeff Var	27.18227		

Here, we could say that the fit explains the 61.32% of the total variation in the data and the adjusted R-squared is 55.92 %.

Adjusted R-squared provides an adjustment to the R-squared statistic such that an independent variable that has a correlation to Y increases adjusted R-squared and any variable without a strong correlation will make adjusted R-squared decrease. This is the desired property of a goodness-of-fit statistic.

Now, we will calculate overall ***F-test*** and the purpose is to determine if at least one of these variables is useful in predicting Y.

Hypothesis: - H_0: $B_1 = B_2 = B_3 = B_4 = B_5 = B_6 = 0$

H_1: At least one of $B_1, B_2, B_3, B_4, B_5, B_6$ is not 0

Here, we assume α is equal to 0.05. Now, we can see below in ***ANOVA*** table that p-value is less than 0.0001, which means p-value < α. Therefore, we reject null hypothesis and conclude alternative hypothesis is true and at least one of these B_1, B_2, B_3, B_4, B_5, and B6 is useful in predicting Y.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	2595877	432646	11.36	<.0001
Error	43	1637723	38087		
Corrected Total	49	4233600			

Now, we know that at least one of this predictor is useful but we don't know which one is that. So, for that we look at individual t-tests to test whether the predictor is useful or not when everything is included in the model. As we can see in previous page on parameter estimates table, p-value for X_1, X_2, X_3, X_4, X_5, X_6 are <0.0001, 0.1434, 0.7247, 0.3181, 0.4593, 0.7645. This clearly shows that only for X_1, p-value is less than alpha. Hence, we could say X_1 is still useful or B_1 should be included in the model.

Residual Analysis for Multiple Regression Model

In this case of multiple regression models, we will construct and analyze the following residual plots:

- Residual vs Predicted value
- Residual vs X_1
- Residual vs X_2
- Residual vs X_3
- Residual vs X_4
- Residual vs X_5
- Residual vs X_6

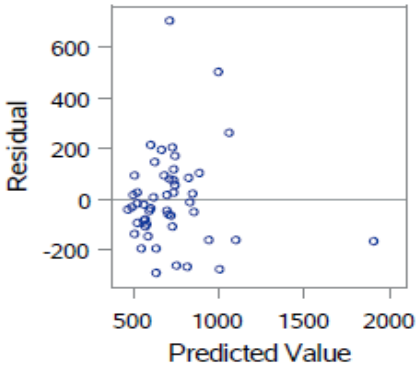
Apart from this, we have also showed ***q-q plot*** here.

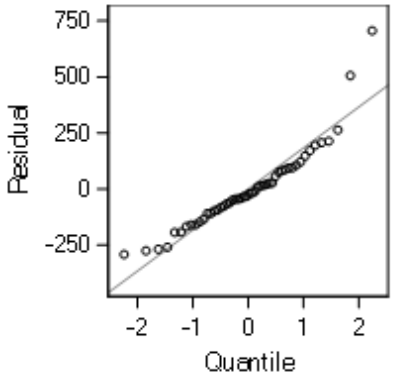
The first residual plot examines the pattern of residuals versus the predicted values of Y. If the residual shows a pattern for predicted values of Y, it means that there is a proof of curvilinear effect in at least one predictor variable or there is need to transform the Y-variable.

The other 6 residual plots involve the independent variables. Patterns in the plots of residual versus independent variable may indicate the existence of curvilinear effect.

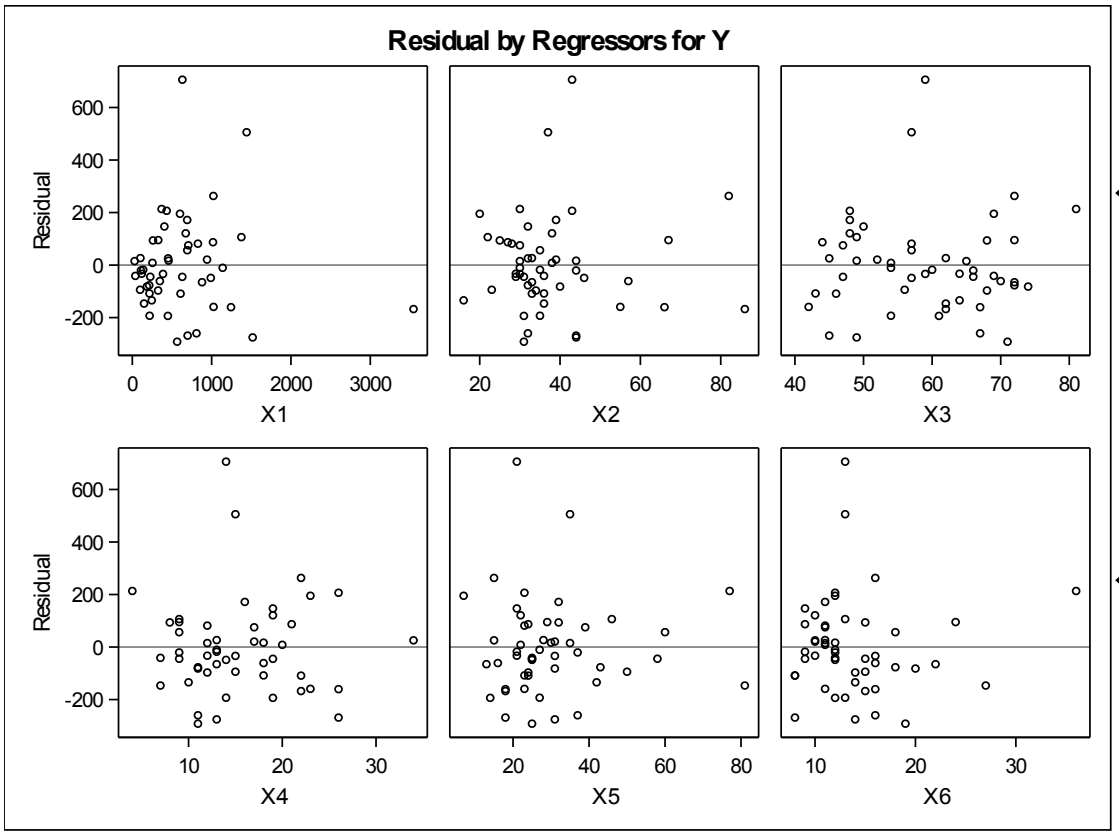
The below figure show the residual plots and q-q plot for this total overall reported crime rate project.

Residual Vs Predicted





Q-Q plot



Residual Vs X1 X2 X3

Residual Vs X4 X5 X6

As we can see above on parameter estimates table (page-3), P-values from X2 to X_6 are greater than alpha which means we fail to reject null hypothesis and these variables are not much useful in predicting Y. So now we remove these 5 variables and then continue regression analysis with only one independent variable i.e. reported violent crime rate per 100,000 residents.

Now after using only one variable X1, following statistics come:-

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**50% Confidence Limits**	
Intercept	1	479.14487	40.52693	11.82	<.0001	451.60134	506.6884
X1	1	0.38757	0.04836	8.01	<.0001	0.35471	0.42044

Now, estimated linear regression equation is $\mathbf{Y = 479.14487 + 0.38757\ X_1}$

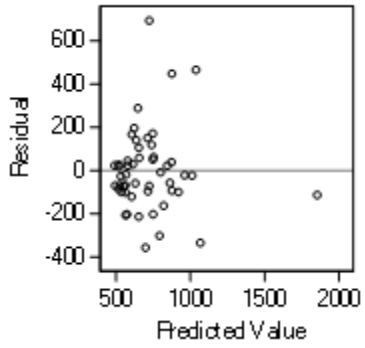
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	2422889	242288	64.23	<.0001
Error	48	1810711	37723		
Corrected Total	49	4233600			

Root MSE	194.2244	**R-Square**	0.572
Dependent Mean	717.9600	**Adj R-Sq**	0.564
Coeff Var	27.05227		

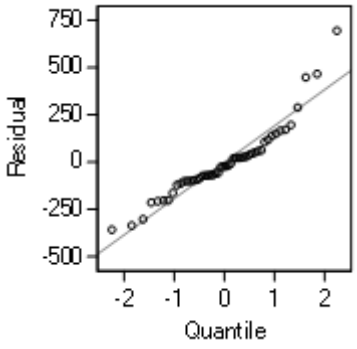
If we see here, F-value = 64.23 is increased here compared to F-value = 11.36 as mentioned on parameter estimates of previous above model. F-statistics is improved 6 times here compared to previous model. Also, if we look at adjusted R-square value i.e. 56.4%; it is better than previous model that shows adjusted value of 55%. Now, we can say this model is better than the previous one.

P-value of X1 is very small here which clearly tells the necessity of reported crime rate per 100,000 residents. This is the best model to predict the crime condition in small cities

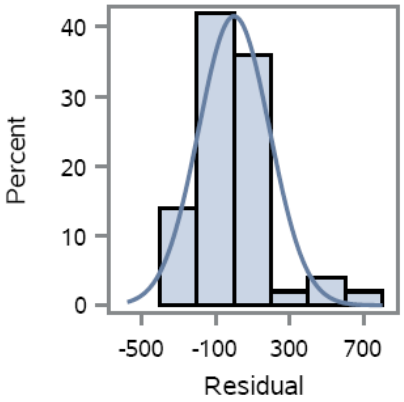
Residual analysis

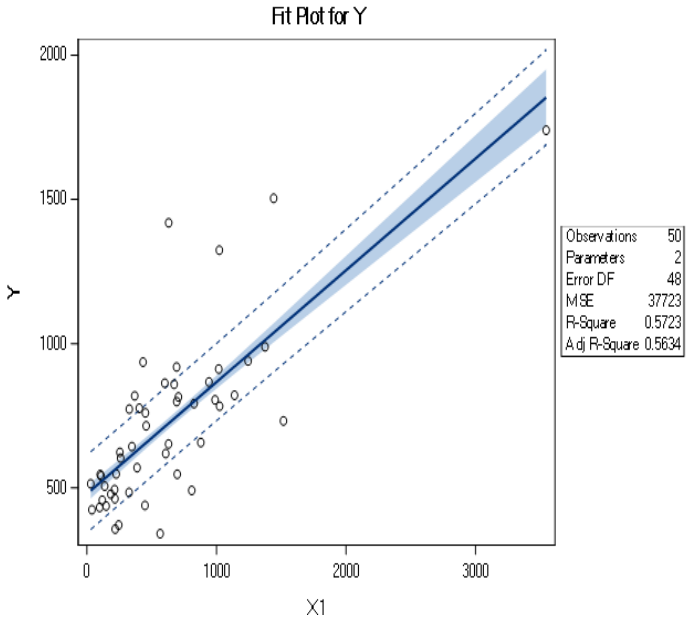


Q-Q plot



Histogram





The regression line appears to be a good fit to the transformed data. As we can see in graph, MSE is decreased and R-square is increased for this transformed data.

Conclusion

The one-variable model, using reported violent crime rate per 100,000 residents, is the best predictors of total overall reported crime rate per 1 million residents. Intuitively thinking, this result is rather reasonable. The total overall reported crime rate per 1 million residents should have a strong linear relationship to the reported violent crime rate of a smaller sample set. Furthermore, although the 6 and 4 variable models had higher R-square values, the superfluous variables made them less efficient than our one-variable model which yielded similar results. At last, the variable X1 show its necessity comparing with the other 5 variables.

Our final model for predicting delinquency rates is as follows:

The one-variable model, using reported violent crime rate per 100,000 residents

$Y = 479.14 + 0.39\ X1$

Appendix

```
/* Inserting a code to get the data*/
data Crime_Data;
infile '/folders/myfolders/Crime_Data.csv' dlm=',' firstobs=2;
input Y X1 X2 X3 X4 X5 X6;
run;

/* Inserting a code to plot the scatter plot*/
proc sgscatter data=Crime_Data;
matrix Y X1 X2 X3 X4 X5 X6;
RUN;

/* Inserting a code for the regression analysis with 6 predictor variables*/
proc reg data=Crime_Data alpha=0.5;
model y = X1 X2 X3 X4 X5 X6 /clb;
run;

/* Inserting a code for the regression analysis with 1 predictor variables*/
proc reg data=Crime_Data alpha=0.5;
model y = X1 /clb;
run;
```