A Regression Study on the Impact of Socioeconomic Factors on Youth Crime Rate in the United States of America

Prepared by Ivy Osei, York University

1. Introduction

This study seeks to determine if a relationship exists between crime rate in the United States among youth who are young males aged 18-24, and a set of socioeconomic factors. The data is excluded to juvenile males due to the smaller number of persistent or serious female offenders in the survey. These factors include youth unemployment, police expenditures and educational background among other variables. The outcome of studies such as this may help identify which factors policy makers should consider when working on improving safety standards as it relates to crime among the young and impressionable.

Criminal justice reform remains one of the most hotly debated national issues in the U.S. There is recurring evidence that juveniles commit crime for reasons such as not doing well in school, truanting from school[1], poor parental discipline and supervision, peer pressure, low family income, unemployment, and high percentage of youth in the community[2]. According to the World Health Organization, homicide is the fourth leading cause of death among youth, and 84% of the victims of these homicides are males[3]. Although youth violence is not unique to America, studies have shown that there is a higher proportion of U.S. youths committing violent acts, illustrating the dire need for criminal justice reform in America[4].

This study does not seek to explain or justify any factors believed to contribute to crime rate among U.S. youth. Rather, it simply attempts to investigate if a relationship exists between crime rate and a set of socio-economic factors often mentioned in the press. The method of analysis is multiple linear regression.

---

[1] "Preventing Involvement in Crime." *Nidirect*, 16 Jan. 2018, www.nidirect.gov.uk/articles/preventing-involvement-crime.

[2] "TEEN CRIME RISK FACTORS." *Why Do Youths Commit Crime, Teenage Crime Risk Factors*, www.acs.edu.au/info/psychology/child-development/crime-risk.aspx.

[3]"Youth Violence." *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/youth-violence.

[4] Office of the Surgeon General (US). "Chapter 2 -- the Magnitude of Youth Violence." *Youth Violence: A Report of the Surgeon General.*, U.S. National Library of Medicine, 1 Jan. 1970, www.ncbi.nlm.nih.gov/books/NBK44300/.

2. Data & Methodology

Empirical data for this study was obtained from 47 states in the U.S. The response variable (Y) is crime rate (among American youth) and is measured by the number of offences per million population. It was collected by Katy Dobson, University of Leeds[5].

The following five-variable multiple regression model is specified:

$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + \varepsilon,$
where
X1 = Youth (number of young males aged 18-24 per 1000)
X2 = Education (average number of years schooling up to 25)
X3 = Expenditure (per capita expenditure on police)
X4 = Labor Force (males employed 18-24 per 1000)
X5 = Youth Unemployment (number of males aged 18-24 per 1000)

Pursuant to the purpose of this study, the regression null hypothesis is as follows:

$H_0 = B_1 = B_2 = B_3 = B_4 = B_5 = 0$ [There is no regression relationship]

The dataset for this study is presented in Appendix1. Before the facts, it is expected that X1 and X5 will have a negative impact on crime rate among youth, that is, the sign of their regression coefficients will be positive, meaning that as the value of these independent variables increase, the mean value of crime rate will increase. A state that is highly populated in number of youth (X1) and has a high rate of youth unemployment (X5), will suggest opportunities for struggling youth to commit crimes. In the same vein, more years of schooling (X2), higher expenditure on policing (X3), and more youth in the labor face (X4) are expected to improve quality of life and reduce risk to commit crimes, meaning the signs of their regression coefficients should be negative. It is arguable that these 3 factors ensure the wholesome security of youth, leading them to have higher self-esteem, combat poverty and lead fulfilling lives. Education and employment open many doors and keeps idle minds at bay, which positively impacts youth by keeping them busy, allowing them to earn income, and prevents them from falling into the wrong crowd and leading a destructive life.

I came to select this model by conducting a model comparison in SAS to choose the best 5-variable model. Thus, State Size (in hundred thousands) and Wage (median weekly wage) were excluded from the model since they didn't meet the criterion, all factors considered.

---

[5] Dobson, K. (n.d.). Crime Rate Data. Leeds; University of Leeds. https://www.sheffield.ac.uk/mash/statistics/datasets

6/19/2021 Results: Crime.sas

Number in Model	R-Square	C(p)	MSE	Variables in Model
5	0.5471	4.0961	424.22644	Youth Education Expenditure LabourForce YouthUnemployment
5	0.5459	4.1930	425.27798	Youth Expenditure LabourForce YouthUnemployment Wage_numeric
5	0.5456	4.2199	425.56988	Youth Expenditure LabourForce StateSize YouthUnemployment
5	0.5363	5.0259	434.31602	Youth Education Expenditure LabourForce Wage_numeric
5	0.5355	5.0910	435.02266	Youth Education Expenditure LabourForce StateSize
5	0.5336	5.2547	436.79895	Youth Expenditure LabourForce StateSize Wage_numeric
5	0.5308	5.4987	439.44591	Youth Education Expenditure StateSize YouthUnemployment
5	0.5306	5.5217	439.69567	Youth Education Expenditure YouthUnemployment Wage_numeric
5	0.5254	5.9630	444.48467	Youth Education Expenditure StateSize Wage_numeric
5	0.5233	6.1460	446.47031	Youth Expenditure StateSize YouthUnemployment Wage_numeric
5	0.4598	11.6288	505.96330	Expenditure LabourForce StateSize YouthUnemployment Wage_numeric
5	0.4587	11.7236	506.99142	Education Expenditure LabourForce StateSize Wage_numeric
5	0.4582	11.7695	507.49029	Education Expenditure LabourForce YouthUnemployment Wage_numeric
5	0.4430	13.0787	521.69542	Education Expenditure StateSize YouthUnemployment Wage_numeric
5	0.4304	14.1679	533.51435	Education Expenditure LabourForce StateSize YouthUnemployment
5	0.3346	22.4323	623.19107	Youth LabourForce StateSize YouthUnemployment Wage_numeric
5	0.3330	22.5736	624.72333	Youth Education LabourForce StateSize Wage_numeric
5	0.3256	23.2136	631.66798	Youth Education StateSize YouthUnemployment Wage_numeric
5	0.2815	27.0200	672.97164	Youth Education LabourForce YouthUnemployment Wage_numeric
5	0.2326	31.2448	718.81438	Education LabourForce StateSize YouthUnemployment Wage_numeric
5	0.1507	38.3096	795.47311	Youth Education LabourForce StateSize YouthUnemployment
6	0.5482	6.0011	433.77590	Youth Education Expenditure LabourForce YouthUnemployment Wage_numeric
6	0.5472	6.0883	434.74570	Youth Education Expenditure LabourForce StateSize YouthUnemployment
6	0.5460	6.1915	435.89350	Youth Expenditure LabourForce StateSize YouthUnemployment Wage_numeric
6	0.5363	7.0235	445.14765	Youth Education Expenditure LabourForce StateSize Wage_numeric
6	0.5309	7.4886	450.31957	Youth Education Expenditure StateSize YouthUnemployment Wage_numeric
6	0.4599	13.6186	518.49842	Education Expenditure LabourForce StateSize YouthUnemployment Wage_numeric
6	0.3365	24.2751	637.02223	Youth Education LabourForce StateSize YouthUnemployment Wage_numeric
7	0.5482	8.0000	444.88580	Youth Education Expenditure LabourForce StateSize YouthUnemployment Wage_numeric

According to the image above, the model I selected as the best 5-variable model has the highest R-squared and the lowest MSRes and C statistic. In addition, its C statistic of 4.0961 is less than and close to k=6, making it a highly desirable model in comparison to the other ones.

3. Analysis and Results

Regression results are summarized below. The F statistic of 9.90 and the p-value <.0001 in comparison to alpha = 0.05, indicates that the regression is statistically significant, and a linear relationship exists between the dependent and independent variables. The coefficient of determination (R^2) suggests that about 55% of the variation in crime rate among youth is explained by the five independent variables, combined. The prediction model is as follows:

$$\hat{Y} = -197.77 + 1.0137X_1 + 1.1044X_2 + 0.8225X_3 + 0.1045X_4 + 0.1872X_5$$

Model: MODEL1
Dependent Variable: CrimeRate_numeric

Number of Observations Read	47
Number of Observations Used	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	21008	4201.69450	9.90	<.0001
Error	41	17393	424.22644		
Corrected Total	46	38402			

Root MSE	20.59676	**R-Square**	0.5471
Dependent Mean	102.80851	**Adj R-Sq**	0.4918
Coeff Var	20.03410		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-197.76562	84.55234	-2.34	0.0243
Youth	1	1.01373	0.31141	3.26	0.0023
Education	1	1.10436	3.00438	0.37	0.7151
Expenditure	1	0.82249	0.12063	6.82	<.0001
LabourForce	1	0.10446	0.08531	1.22	0.2278
YouthUnemployment	1	0.18715	0.18321	1.02	0.3130

The test of significance of the independent variables – measured by the t statistics – shows that not all the explanatory variables are statistically significant, namely education (X2), labour force (X4), and youth unemployment (X5). In other words, only two variables, number of male youth (X1) and police expenditure (X3), contribute meaningful information in the prediction of youth crime rate. The signs of the coefficients for the independent variables are consistent with prior expectations except for X2, X3, and X5, which were expected to be negative, of which X2 and X5 are not significant. Taking the value of the coefficient for X1 for example, it suggests that the average crime rate increases by about 1.01 for every percentage increase in number of youths per 1000.

4. Test for Multicollinearity

Often, when collinearity exists, the sign of the regression coefficient is the opposite of what logic suggests. Also, the t-value may not be significant even when the F(model) is significant. To confirm that the lack of statistical significance and illogical sign associated with X2, X4, and X5, being education, labour force and youth unemployment, respectively, is not due to multicollinearity, the correlations between the independent variables are examined. The correlation half matrix is presented below.

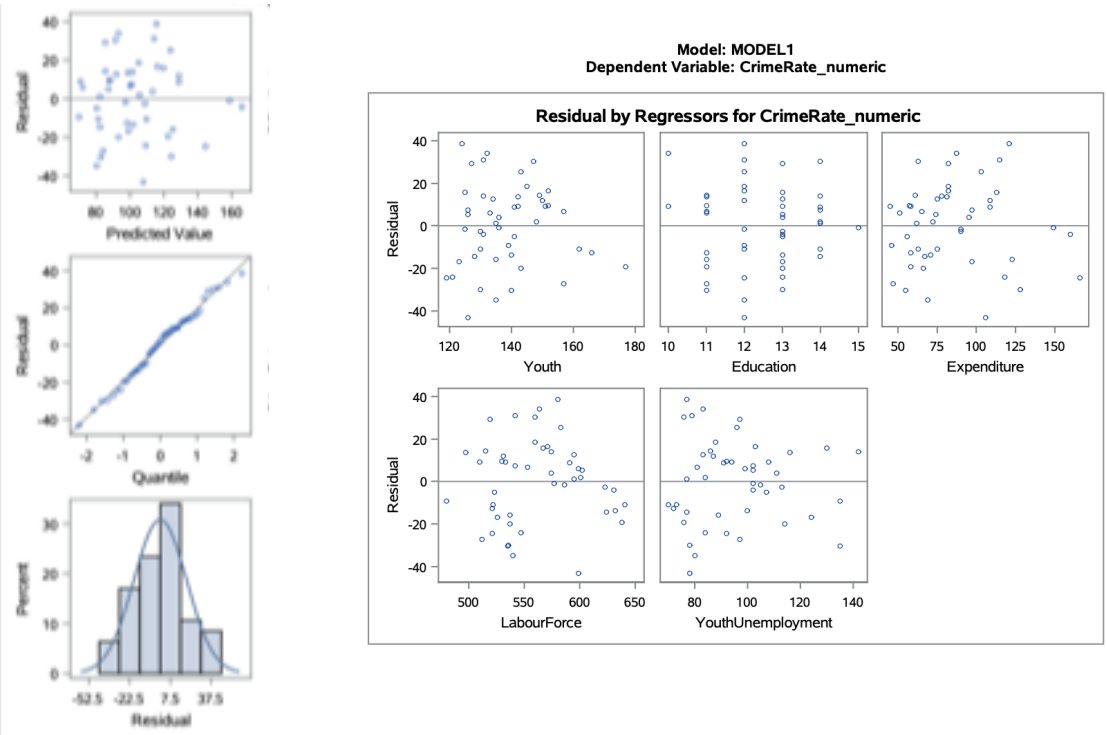
Pearson Correlation Coefficients, N = 47 Prob > |r| under H0: Rho=0						
	Youth	Education	Expenditure	LabourForce	YouthUnemployment	CrimeRate_numeric
Youth	1.00000	-0.40342 0.0049	-0.50574 0.0003	-0.16095 0.2798	-0.22438 0.1295	-0.05500 0.7135
Education	-0.40342 0.0049	1.00000	0.23204 0.1165	0.39923 0.0054	0.04463 0.7658	0.12750 0.3931
Expenditure	-0.50574 0.0003	0.23204 0.1165	1.00000	0.12149 0.4159	-0.04370 0.7706	0.64621 <.0001
LabourForce	-0.16095 0.2798	0.39923 0.0054	0.12149 0.4159	1.00000	-0.22940 0.1209	0.16931 0.2552
YouthUnemployment	-0.22438 0.1295	0.04463 0.7658	-0.04370 0.7706	-0.22940 0.1209	1.00000	-0.05061 0.7355
CrimeRate_numeric	-0.05500 0.7135	0.12750 0.3931	0.64621 <.0001	0.16931 0.2552	-0.05061 0.7355	1.00000

We can't conclude that multicollinearity is present, because the absolute values of the correlation coefficients between dependent and independent variables aren't large. Thus, it is necessary to conduct a more rigorous analysis of multicollinearity based on the calculation of the variance inflation factor (VIF) for each of the variables, defined as follows:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-197.76562	84.55234	-2.34	0.0243	0
Youth	1	1.01373	0.31141	3.26	0.0023	1.66083
Education	1	1.10436	3.00438	0.37	0.7151	1.38980
Expenditure	1	0.82249	0.12063	6.82	<.0001	1.39365
LabourForce	1	0.10446	0.08531	1.22	0.2278	1.28881
YouthUnemployment	1	0.18715	0.18321	1.02	0.3130	1.18303

A high VIF means that the variance of the regression coefficient is large, suggesting that there is an error in $\hat{B}_j$, signifying that the regression coefficient is not desirable. A helpful rule of thumb is that collinearity exists if VIF > 10. As the table above shows, the highest VIF is 1.66083 < 10, which is too small to cause concern. With this finding, one can conclude that collinearity is most likely not a problem in the model.

5. Tests for Normality and Constant Variance



The top plot (residual vs predicted value) can be used to verify the constant variance assumption and to check the pattern of residuals. One can see that there is a problem of violation of the constant variance assumption in the top plot since as the residual decreases, the predicted values increase and they appear to fan inward. Thus, because the pattern doesn't appear to be random as with the residual of the youth, it indicates that the data may be nonlinear and it's necessary to transform the data. However, the middle plot (Q-Q) plot and the bottom plot (histogram of residuals) demonstrate that the data closely adhere to the normality assumption.

While the scatter of residuals appear to fan inwards, there is no definitive indication that nonconstant variance is present. Nevertheless, a popular and often effective variance stabilizing transformation is the square root transformation of only the values of Y, as follows:

$$\sqrt{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5$$

The regression results of this transformed model are summarized below. These results do not show a marked improvement over the original model. The coefficients have the same manner of statisitical significance as before. Also, the value of the coefficient of determination decreased from 0.5471 to 0.4901, indicating that the transformed data did not result in a better model. In

addition, the residual plot vs predicted value has the same pattern as the original model. Thus, the original model remains the most effective 5-variable model of the dataset.

The REG Procedure
Model: MODEL1
Dependent Variable: yt

Number of Observations Read	47
Number of Observations Used	47

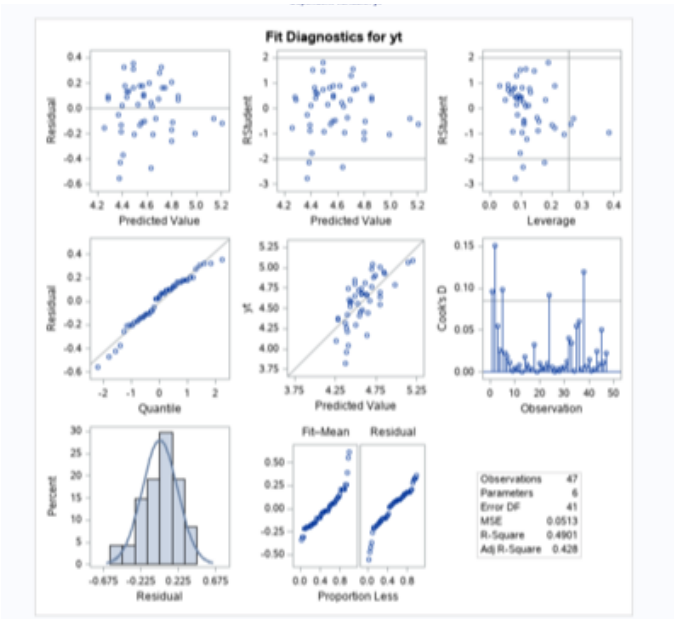
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.02197	0.40439	7.88	<.0001
Error	41	2.10325	0.05130		
Corrected Total	46	4.12522			

Root MSE	0.22649	R-Square	0.4901
Dependent Mean	4.59121	Adj R-Sq	0.4280
Coeff Var	4.93317		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.55833	0.92978	1.68	0.1013
Youth	1	0.01038	0.00342	3.03	0.0042
Education	1	0.01383	0.03304	0.42	0.6778
Expenditure	1	0.00804	0.00133	6.06	<.0001
LabourForce	1	0.00104	0.00093812	1.11	0.2727
YouthUnemployment	1	0.00162	0.00201	0.80	0.4262



6. Conclusions, Limitations and Future Improvements

This study examined the relationship between crime rate among youth and a set of socioeconomic factors. Crime rate is measured by number of offences per million population in the United States.

Results show that while number of male youth (X1), and police expenditure (X3) affect the youth crime rate, education (X2), labour force (X4) and the youth unemployment rate (X5) are ineffective predictor variables. This means that because of their high p-values, changes in these predictor variables are unassociated with changes in the response variable. However, the original model accounts for about 55% of changes in the mean youth crime rate in the United States which is an indicator that this is a capable model. More specifically, there is evidence that the crime rate in America rises in states that are highly populated with youth and quite surprisingly, states that spend more on police expenditure.

Although the coefficient of determination is reasonably high, the insignificance of three of the independent variables indicates that this model is not an accuracte predictor of the U.S. youth crime rate. The best predictor of the crime rate would be determined from the number of male youth (X1) and police expenditure (X3) parameter estimates.

A limitation to this study is the few number of observations involved and many independent variables, which could have lead to a few of the regression coefficients being insignificant. However, the nature of this study being that it examines crime rates among youth in different states, prohibits one from obtainining more observations, as there are only 50 states in the U.S. In addition, the inability of a few parameter estimates to predict the response variable illustrates the variability and spontaneity of human nature, which is sometimes unpredictable. Nevertheless, interventions such as after-school programs for youth, and community and problem-oriented policing may reduce the crime rate in the United States, and help reduce the effect of more police expenditure and high youth population as predictor variables affecting crime rate.

Appendix

6/20/2021 Code: Crime.sas

```
data crime;
    infile '/home/u58672568/sasuser.v94/Project/Crime_SPSS.csv' dlm=',' firstobs=2;
    input CrimeRate $ Youth Education Expenditure LabourForce StateSize YouthUnemployment Wage $;
run;


data crime2; set crime;
    CrimeRate_numeric = input(CrimeRate, best5.);
    Wage_numeric = input(Wage, best5.);
run;

proc rsquare cp mse;
model CrimeRate_numeric = Youth Education Expenditure LabourForce StateSize YouthUnemployment Wage_numeric;
run;

data crime_finalmodel;
set WORK.crime2;
drop CrimeRate StateSize Wage_numeric;
keep CrimeRate_numeric Youth Education Expenditure LabourForce YouthUnemployment;
run;

proc corr data = crime_finalmodel;
proc reg data = crime_finalmodel;
model CrimeRate_numeric = Youth Education Expenditure LabourForce YouthUnemployment / p vif;
run;

data crime_transformation; set crime_finalmodel;
    yt = log(CrimeRate_numeric);
run;
proc reg data = crime_transformation;
model CrimeRate_numeric = Youth Education Expenditure LabourForce YouthUnemployment;
plot rstudent.*predicted.;
run;
proc reg data = crime_transformation;
model yt = Youth Education Expenditure LabourForce YouthUnemployment;
plot rstudent.*predicted.;
run;
```

6/20/2021 Code: CrimeRegression.ctk

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '6/20/21, 10:47 PM'
 * Generated by 'u58672568'
 * Generated on server 'ODAWS04-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chr
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-04%253A00&ticket=S
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc reg data=WORK.CRIME2 alpha=0.05 plots(only)=(diagnostics residuals
        observedbypredicted);
    model CrimeRate_numeric=Youth Education Expenditure LabourForce
        YouthUnemployment /;
    run;
quit;
```

i	Y CrimeRate	X1 Youth	X2 Education	X3 Expenditure	X4 LabourForce
1	45.50	135	12	69	540
2	52.30	140	11	55	535
3	56.60	157	11	47	512
4	60.30	139	12	46	480
5	64.20	126	12	106	599
6	67.60	128	14	67	624
7	70.50	130	14	63	641
8	73.20	143	13	66	537
9	75.00	141	13	56	523
10	78.10	133	11	51	599
11	79.80	142	13	45	533
12	82.30	123	13	97	526
13	83.10	135	14	62	595
14	84.90	121	13	118	547
15	85.60	166	11	58	521
16	88.00	140	13	71	632
17	92.30	126	13	74	602
18	94.30	130	13	128	536
19	95.30	125	12	90	586
20	96.80	151	10	58	510
21	97.40	152	11	57	530
22	98.70	162	12	75	522
23	99.90	149	11	61	515
24	103.00	177	11	58	638
25	104.30	134	13	75	595
26	105.90	130	13	90	623
27	106.60	157	11	65	553
28	107.20	148	14	72	601
29	108.30	126	14	97	542
30	109.40	135	11	123	537
31	112.10	142	11	81	497
32	114.30	127	13	82	519
33	115.10	131	14	78	574
34	117.20	136	13	95	574
35	119.70	119	12	166	521
36	121.60	147	14	63	560
37	123.40	145	12	82	560
38	127.20	132	10	87	564
39	132.40	152	12	82	571
40	135.50	125	13	113	567
41	137.80	141	14	109	591
42	140.80	150	12	109	531
43	145.40	131	12	115	542
44	149.30	143	12	103	583
45	154.30	124	12	121	580
46	157.70	136	15	149	577

* The 47th observation is excluded from the image of the table above, due to it exceeding the size of the page*