Final Project Report Part 5

US 1960 Crime Rate Analysis

Using Linear Regression

By: Arsema Demeke

EMSE 6765

December 7, 2023

Table of Contents **Page**

Introduction

Criminologists recently proposed an idea to study how punishment regimes affect crime rates in the United States of America. Data on the US crime rate in 1960 which included information from 47 of the 50 states. This report conducts data analysis on the crime rates and performs linear regression to extract information on the relationship between crime rates and punishment regimes. This is accomplished by:

1. Developing a linear regression model on the crime rate datasets using Log(Y)
2. Performing a detailed diagnostic analysis of the model and interpreting its results
3. Forecasting the crime rate with a 95% prediction interval and the expected crime rate with a 95% confidence interval

The original crime rate data includes variables such as:

The Dependent Variable:

- **Crime Rate (Y)**: Number of offenses per 100,000 population in 1960
- **Log of Crime Rate (Log (Y))**: Log of the number of offenses per 100,000 in 1960

The Independent Variables:

- **Po1**: Per capita expenditure in police protection in 1960
- **Po2**: Per capita expenditure in police protection in 1959
- **Wealth**: Median value of transferrable assets or family income
- **Prob**: Probability of imprisonment
- **Pop:** State population in 1960 in hundred thousands
- **Ed:** mean years of schooling of the population aged 25 years or over
- **U1:** unemployment rate of urban males 14-24
- **U2:** unemployment rate of urban males 39-24
- **LF:** labor force participation rate of civilian urban males in the age group 14-24
- **M.F:** number of males per 100 females
- **Ineq:** Income inequality: percentage of families earning below half the median income
- **Time:** average time in months served by offenders in state prisons before their first
- **M:** percentage of males aged 14-24 in the total state population

The crime rate data and these variables can be analyzed in Table 1 below:

Table 1: Original Crime Rate Data

Crime (Y)	Log(Crime)	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
791	2.898	5.8	5.6	3940	0.084602	33	9.1	0.108	4.1	0.51	95	26.1	26.2011	15.1
1635	3.214	10.3	9.5	5570	0.029599	13	11.3	0.096	3.6	0.583	101.2	19.4	25.2999	14.3
578	2.762	4.5	4.4	3180	0.083401	18	8.9	0.094	3.3	0.533	96.9	25	24.3006	14.2
1969	3.294	14.9	14.1	6730	0.015801	157	12.1	0.102	3.9	0.577	99.4	16.7	29.9012	13.6
1234	3.091	10.9	10.1	5780	0.041399	18	12.1	0.091	2	0.591	98.5	17.4	21.2998	14.1
682	2.834	11.8	11.5	6890	0.034201	25	11	0.084	2.9	0.547	96.4	12.6	20.9995	12.1
963	2.984	8.2	7.9	6200	0.0421	4	11.1	0.097	3.8	0.519	98.2	16.8	20.6993	12.7
1555	3.192	11.5	10.9	4720	0.040099	50	10.9	0.079	3.5	0.542	96.9	20.6	24.5988	13.1
856	2.932	6.5	6.2	4210	0.071697	39	9	0.081	2.8	0.553	95.5	23.9	29.4001	15.7
705	2.848	7.1	6.8	5260	0.044498	7	11.8	0.1	2.4	0.632	102.9	17.4	19.5994	14
1674	3.224	12.1	11.6	6570	0.016201	101	10.5	0.077	3.5	0.58	96.6	17	41.6	12.4
849	2.929	7.5	7.1	5800	0.031201	47	10.8	0.083	3.1	0.595	97.2	17.2	34.2984	13.4
511	2.708	6.7	6	5070	0.045302	28	11.3	0.077	2.5	0.624	97.2	20.6	36.2993	12.8
664	2.822	6.2	6.1	5290	0.0532	22	11.7	0.077	2.7	0.595	98.6	19	21.501	13.5
798	2.902	5.7	5.3	4050	0.0691	30	8.7	0.092	4.3	0.53	98.6	26.4	22.7008	15.2
946	2.976	8.1	7.7	4270	0.052099	33	8.8	0.116	4.7	0.497	95.6	24.7	26.0991	14.2
539	2.732	6.6	6.3	4870	0.076299	10	11	0.114	3.5	0.537	97.7	16.6	19.1002	14.3
929	2.968	12.3	11.5	6310	0.119804	31	10.4	0.089	3.4	0.537	97.8	16.5	18.1996	13.5
750	2.875	12.8	12.8	6270	0.019099	51	11.6	0.078	3.4	0.536	93.4	13.5	24.9008	13
1225	3.088	11.3	10.5	6260	0.034801	78	10.8	0.13	5.8	0.567	98.5	16.6	26.401	12.5
742	2.870	7.4	6.7	5570	0.0228	34	10.8	0.102	3.3	0.602	98.4	19.5	37.5998	12.6
439	2.642	4.7	4.4	2880	0.089502	22	8.9	0.097	3.4	0.512	96.2	27.6	37.0994	15.7
1216	3.085	8.7	8.3	5130	0.0307	43	9.6	0.083	3.2	0.564	95.3	22.7	25.1989	13.2
968	2.986	7.8	7.3	5400	0.041598	7	11.6	0.142	4.2	0.574	103.8	17.6	17.6	13.1
523	2.719	6.3	5.7	4860	0.069197	14	11.6	0.07	2.1	0.641	98.4	19.6	21.9003	13
1993	3.300	16	14.3	6740	0.041698	3	12.1	0.102	4.1	0.631	107.1	15.2	22.1005	13.1
342	2.534	6.9	7.1	5640	0.036099	6	10.9	0.08	2.2	0.54	96.5	13.9	28.4999	13.5
1216	3.085	8.2	7.6	5370	0.038201	10	11.2	0.103	2.8	0.571	101.8	21.5	25.8006	15.2
1043	3.018	16.6	15.7	6370	0.0234	168	10.7	0.092	3.6	0.521	93.8	15.4	36.7009	11.9
696	2.843	5.8	5.4	3960	0.075298	46	8.9	0.072	2.6	0.521	97.3	23.7	28.3011	16.6
373	2.572	5.5	5.4	4530	0.041999	6	9.3	0.135	4	0.535	104.5	20	21.7998	14
754	2.877	9	8.1	6170	0.042698	97	10.9	0.105	4.3	0.586	96.4	16.3	30.9014	12.5
1072	3.030	6.3	6.4	4620	0.049499	23	10.4	0.076	2.4	0.56	97.2	23.3	25.5005	14.7
923	2.965	9.7	9.7	5890	0.040799	18	11.8	0.102	3.5	0.542	99	16.6	21.6997	12.6
653	2.815	9.7	8.7	5720	0.0207	113	10.2	0.124	5	0.526	94.8	15.8	37.4011	12.3
1272	3.104	10.9	9.8	5590	0.0069	9	10	0.087	3.8	0.531	96.4	15.3	44.0004	15
831	2.920	5.8	5.6	3820	0.045198	24	8.7	0.076	2.8	0.638	97.4	25.4	31.6995	17.7
566	2.753	5.1	4.7	4250	0.053998	7	10.4	0.099	2.7	0.599	102.4	22.5	16.6999	13.3
826	2.917	6.1	5.4	3950	0.047099	36	8.8	0.086	3.5	0.515	95.3	25.1	27.3004	14.9
1151	3.061	8.2	7.4	4880	0.038801	96	10.4	0.088	3.1	0.56	98.1	22.8	29.3004	14.5
880	2.944	7.2	6.6	5900	0.0251	9	12.2	0.084	2	0.601	99.8	14.4	30.0001	14.8
542	2.734	5.6	5.4	4890	0.088904	4	10.9	0.107	3.7	0.523	96.8	17	12.1996	14.1
823	2.915	7.5	7	4960	0.054902	40	9.9	0.073	2.7	0.522	99.6	22.4	31.9989	16.2
1030	3.013	9.5	9.6	6220	0.0281	29	12.1	0.111	3.7	0.574	101.2	16.2	30.0001	13.6
455	2.658	4.6	4.1	4570	0.056202	19	8.8	0.135	5.3	0.48	96.8	24.9	32.5996	13.9
508	2.706	10.6	9.7	5930	0.046598	40	10.4	0.078	2.5	0.599	98.9	17.1	16.6999	12.6
849	2.929	9	9.1	5880	0.052802	3	12.1	0.113	4	0.623	104.9	16	16.0997	13

Analysis of Crime(Y) Dependent Variable:

The Crime(Y) dependent variable was further analyzed using a histogram and a probability plot generated through Minitab. Firstly, an analysis of normality was conducted on the Crime (Y) dependent variable using a histogram. Below, Figure 1 illustrates the resulting histogram of Crime (Y). Next, a probability plot of Crme(Y) was constructed as represented in Figure 2 below. Through these depictions, the following were concluded:

- The histogram reveals a left-ward skew of the Crime (Y) data when fitted into a normal distribution graph
- The Crime(Y) data has a large standard deviation from the mean, as supported by both the histogram and probability plot, which shows high variability around the mean
- The probability plot reveals that the data points don't form a straight line with a few points located outside of the boundaries
- A small p-value of less than 0.005 is observed through the probability plot which indicates non-normality and asymmetry

This analysis is supported by the Minitab depiction of Figure 1 and Figure 2 below:

Figure 1: Histogram of Crime (Y)

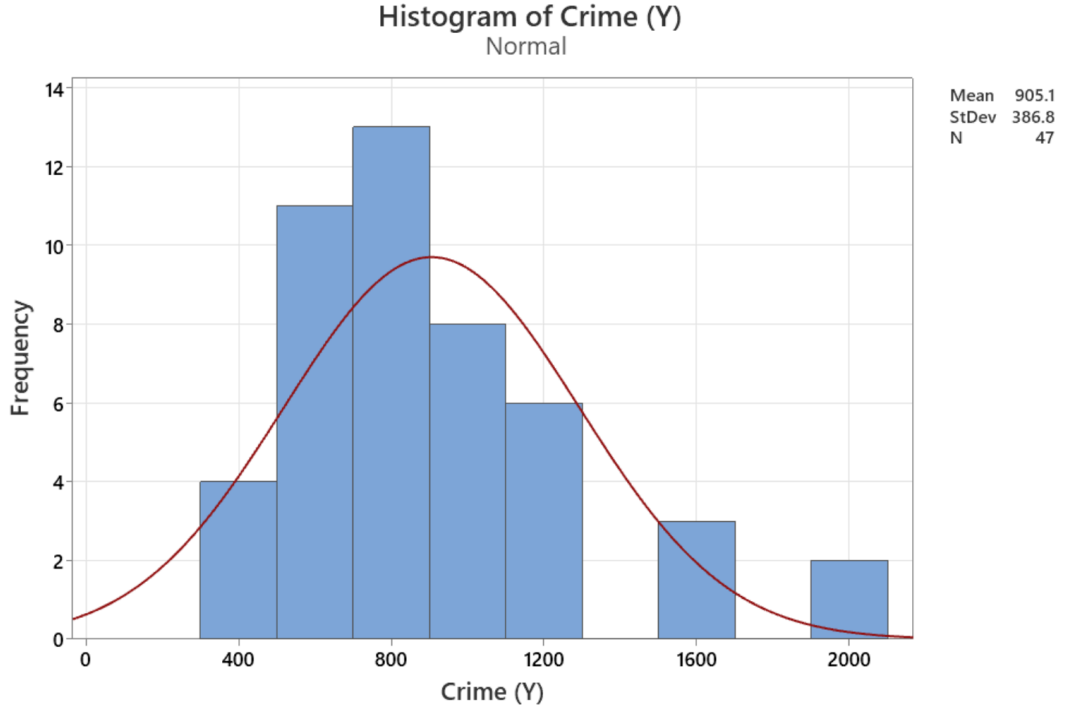
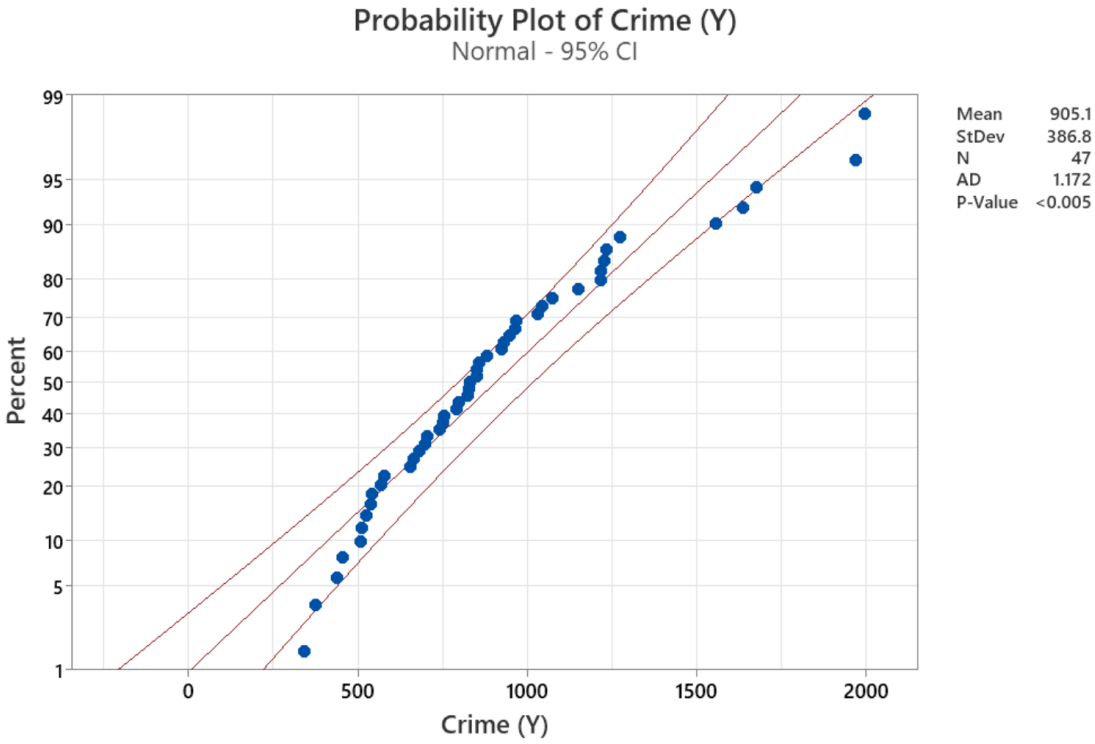


Figure 2: Probability Plot of Crime (Y)



Motivating the Use of Log(Y):

Although it is not required for the dependent variable to be normally distributed, it is preferred. This is because when conducting a linear regression analysis, an assumption is made about the residuals of the dependent variables; with the mean of the residuals equaling zero, the residuals of the model have to be normally distributed. This observance is facilitated by having a dependent variable that is also normally distributed. To overcome this concern and achieve a normally distributed dependent variable, the logarithm of Crime (Y), is used instead of Crime (Y). When the histogram and probability plot of Log(Crime) or Log(Y) is analyzed, the following observations are made:

- The histogram reveals a bell-shaped curve with one peak and symmetry centered around the mean; no skewness is observed and fits well into the fitted normal distribution graph
- The Log(Y) data has a small standard deviation from the mean, as supported by both the histogram and probability plot, which shows low variability around the mean
- The probability plot reveals that the data points are close to forming a straight line with no points located outside of the boundaries
- A large p-value of less than 0.893 is observed through the probability plot which indicates normality and symmetry

Because of these observations, using Log(Y) as the dependent variable is justified.

Below, Figure 3 and Figure 4 display the resulting histogram and probability plot of Log(Y):

Figure 3: Histogram of Log(Y)

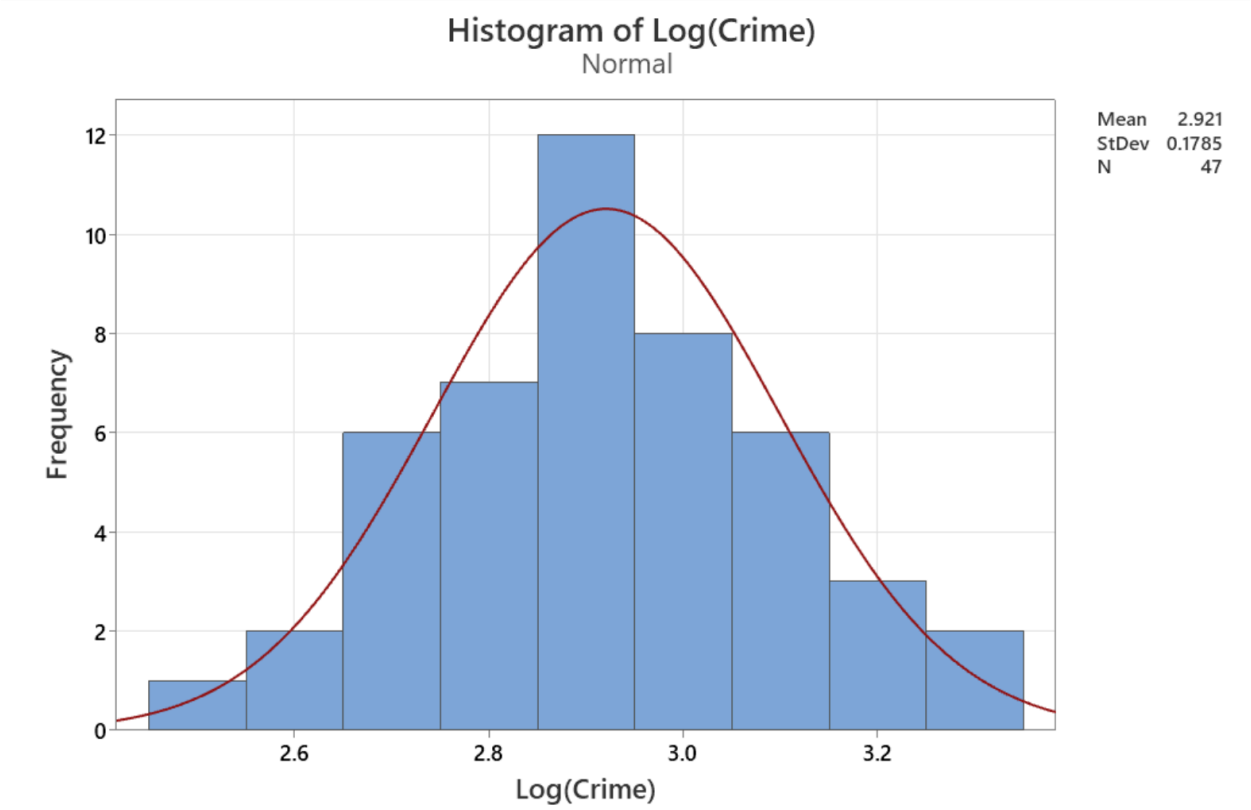
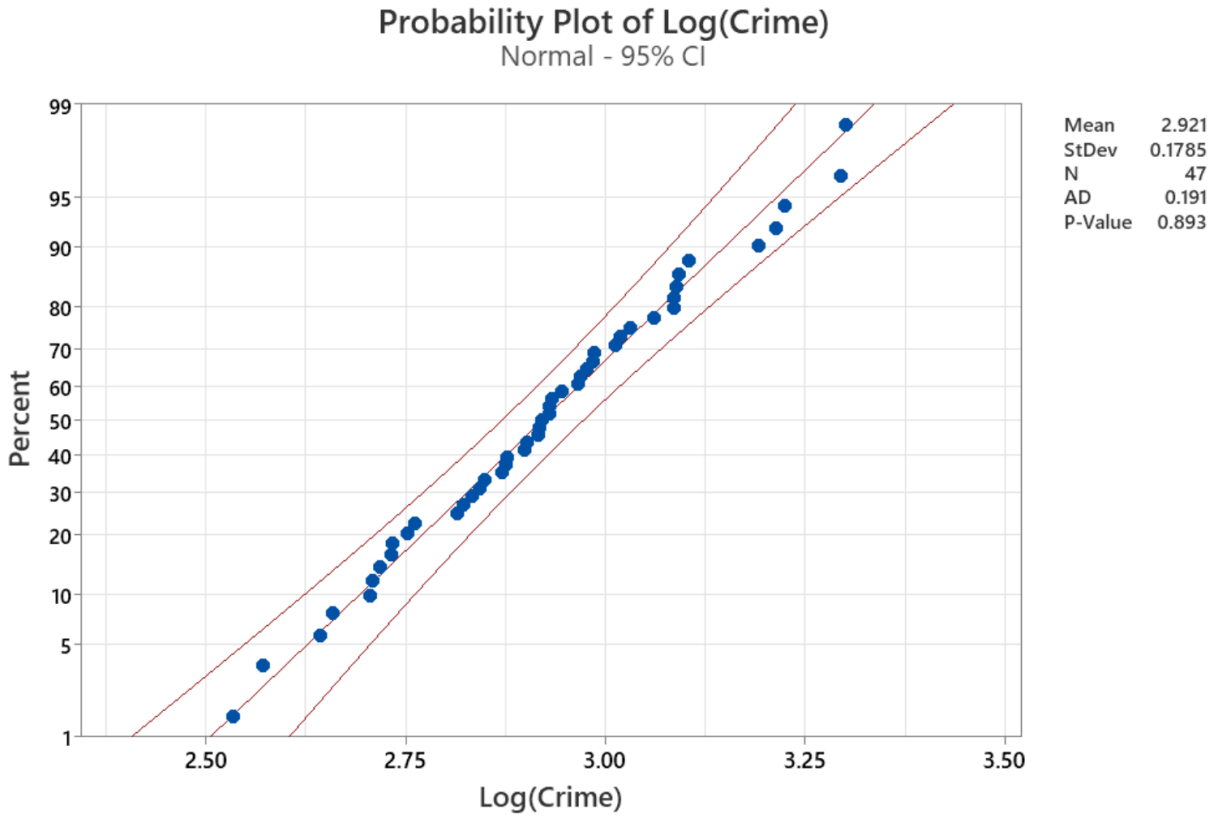


Figure 4: Probability Plot of Log(Y)



Correlation Analysis:

After identifying which dependent variable should be used, the necessary next step is to identify and select the proper explanatory variables that will be used to create and study linear regression. To achieve this, a correlation analysis of the relationship between the dependent variable and the independent variables will be conducted. Specifically, a correlation matrix will be constructed to compare and analyze the strength of the relationship between the dependent variable Log(Y) and the independent variables that exist within the data. The correlation coefficients range from -1 to 1 with values closer to these two numbers indicating stronger relationships between the variables.

Figure 5 below depicts the correlation matrix conducted in Excel between these variables:

Figure 5: Correlation Matrix between Log(Crime) and Independent Varibles

	Log(Crime)	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
Log(Crime	1													
Po1	0.65463148	1												
Po2	0.63730461	0.993586	1											
Wealth	0.4266203	0.787225	0.794262	1										
Prob	-0.4118918	-0.47325	-0.47303	-0.55533	1									
Pop	0.33735892	0.526284	0.513789	0.308263	-0.34729	1								
Ed	0.30214517	0.482952	0.49941	0.735997	-0.38992	-0.01723	1							
U1	-0.0748663	-0.0437	-0.05171	0.044857	-0.00747	-0.03812	0.018103	1						
U2	0.16740426	0.185093	0.169224	0.092072	-0.06159	0.270422	-0.21568	0.745925	1					
LF	0.17273188	0.121493	0.10635	0.294632	-0.25009	-0.12367	0.561178	-0.2294	-0.42076	1				
M.F	0.14816066	0.03376	0.022843	0.179609	-0.05086	-0.41063	0.436915	0.351892	-0.01869	0.513559	1			
Ineq	-0.1516927	-0.6305	-0.64815	-0.884	0.465322	-0.12629	-0.76866	-0.06383	0.015678	-0.26989	-0.16709	1		
Time	0.14257761	0.103358	0.075627	0.000649	-0.43625	0.46421	-0.25397	-0.16985	0.101358	-0.12364	-0.4277	0.101823	1	
M	-0.0562343	-0.50574	-0.51317	-0.67006	0.361116	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.639211	0.114511	1

In Figure 5's Correlation matrix, the threshold of 0.4 was chosen because it allows for the use of 4 exploratory variables of the 13 that are provided, allowing for a more detailed analysis. Next, a pattern is observed in the correlation between the different variables according to this threshold. More importantly, column one shows the correlation coefficients between Log (Y) and the independent variables. The highlighted correlation coefficients in this column identify the independent variables with the strongest correlation and linear relationship with Log (Y). Through this visualization, Po1, Po2, Wealth, and Prob can be selected as the explanatory variables to be used to study crime rates in the US in 1960 as they have the highest correlation with Log (Crime) according to the chosen 0.4 thresholds.

Initial Regression Analysis:

Next, an initial regressional analysis is conducted on the explanatory variables and Log (Crime). The results of this analysis are depicted below in Figure 6:

Figure 6: Initial Regression Analysis for Log(Y) and Four Explanatory Variables

Regression Statistics	
Multiple R	0.698028554
R Square	0.487243863
Adjusted R Squ	0.438409945
Standard Error	0.133782263
Observations	47

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	0.714302014	0.178576	9.977571	9.0905E-06
Residual	42	0.75170314	0.017898		
Total	46	1.466005154			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	VIF
Intercept	2.892264251	0.167897367	17.22638	1.19E-20	2.55343365	3.23109485	2.55343365	3.231094854	
Po1	0.094633166	0.058779966	1.609956	0.114899	-0.0239896	0.21325594	-0.0239896	0.21325594	78.43
Po2	-0.04988119	0.063379336	-0.78703	0.435685	-0.1777859	0.07802349	-0.1777859	0.078023491	80.72
Wealth	-5.6972E-05	3.57479E-05	-1.59371	0.118499	-0.0001291	1.517E-05	-0.0001291	1.51702E-05	3.06
Prob	-1.62461163	1.046317922	-1.55269	0.128	-3.7361667	0.48694342	-3.7361667	0.486943422	1.45

Regression Equation

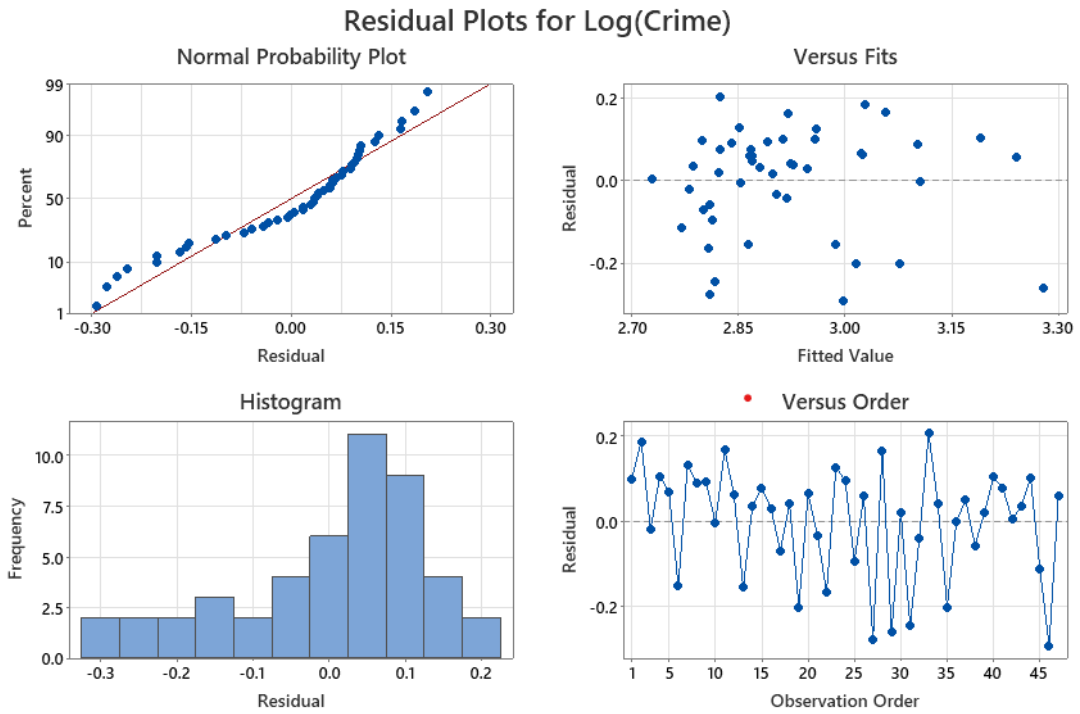
Log(Crime)=2.892 + 0.0946 Po1 - 0.0499 Po2 - 0.000057 Wealth - 1.62 Prob

Through Figure 6, we can make several observations:

- The relationship between Log(Y) and the four explanatory variables produced an R^2 value of 0.49 and an adjusted R^2 value of 0.44 which are values that are not high and thus prove that only some of the explanatory variables are statistically significant and correlated
- The F-statistic value is relatively large with a value of 10 and the small value for the significance of F proves that there is a general relationship between the dependent variable and exploratory variables but is not high enough to prove a strong relationship exists
- The p-values listed for the explanatory variables vary with many being above the level of significance of 0.05 from the 95% confidence interval being used, especially Po2, proving that the coefficients of the variables are not statistically significant nor significantly different from 0
- There are extremely high Variance Inflation Factor values for Po1 and Po2 which indicates that a high multi-colinearity and a high inflation exists in or between them
- The Durbin-Watson statistic value here is 2.3 which is a beneficial finding because its close to a value of 2 since its between the range 1.5-2.5 which indicates little sign of autocorrelation

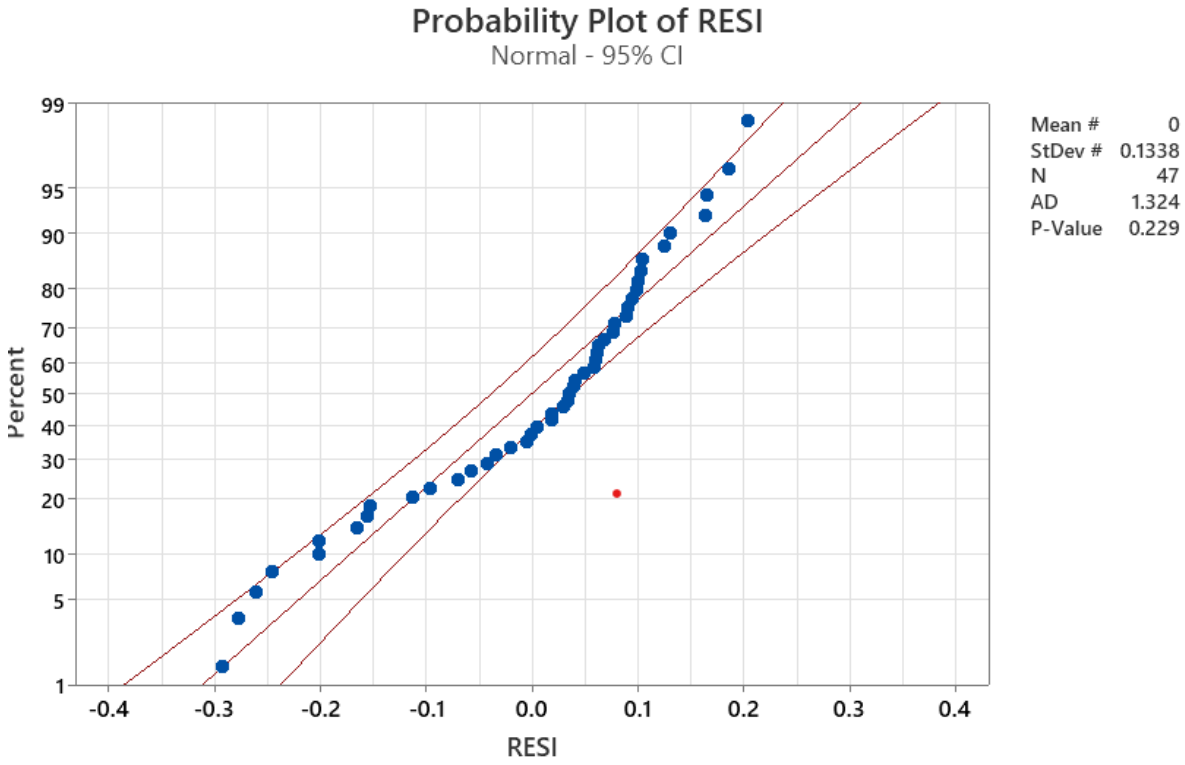
The following plots conducted in Minitab and illustrated in Figure 7 below further support these observations:

Figure 7: Residual Plots for Initial Regression Analysis for Log(Y)



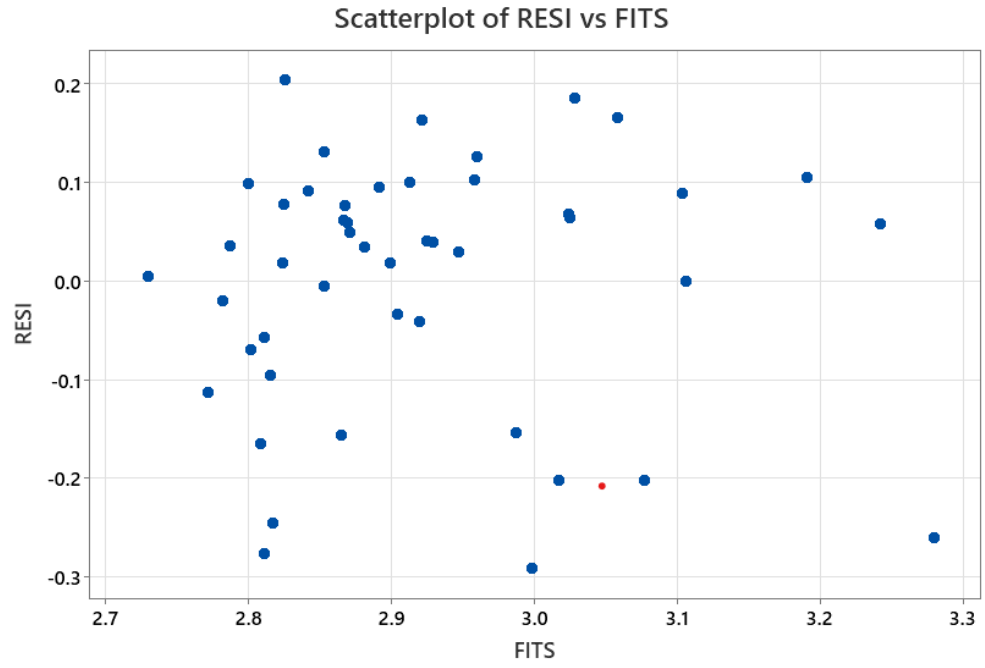
A probability plot of the initial regression residuals is depicted in Figure 8 below:

Figure 8: Probability Plot of Initial Regression Residuals



An initial regression residual versus fitted values plot is also depicted in Figure 9 below:

Figure 9: Initial Regression's Residual versus Fitted Values for Log(Y)



Diagnostic Analysis for Initial Regression Analysis:

The plots represented by the figures above provide graphical evidence of the earlier observations. In Figure 8, we can observe that the Residual of the initial regression analysis is not normally distributed as its 0.229 p-value is below 0.250 and some data points are located outside of the boundaries; furthermore, the four-in-one residual plots in Figure 7 show that most of the data points lie below the residual's zero line and are not as chaotic as preferred; the histogram further supports the probability plot which proves the residual is not normally distributed because of its right-skewness from zero. Lastly, Figure 9 unfortunately doesn't depict the data points being randomly distributed horizontally along the residual line of 0 and might indicate heteroskedasticity.

These observations raise concerns regarding the strength of the relationship between Log(Y) and the chosen explanatory variables and show that they don't have a strong relationship and make it difficult to predict the values of Log(Crime) using the exploratory values. Thus, a new model should be created that can improve the model and better help predict the value of Log(Y) and Y using the exploratory variables.

Updated Regression Analysis:

To construct an improved model that reflects a strong relationship between the dependent and independent variables, explanatory variables should be eliminated. Of the four variables used, Figure 6 above reveals that the Po1 and Po2 variables have high Variance of Inflation values which indicates their multicollinearity. Thus, one must be eliminated to allow for the accurate assessment of each exploratory variable's effect on the dependent variable. Since Po2 also has the highest p-value in the regression analysis, it should be eliminated.

The results of this updated analysis are depicted below in Figure 10:

Figure 10: Updated Regression Analysis for Log(Y) and Three Explanatory Variables

Regression Statistics	
Multiple R	0.69259065
R Square	0.479681809
Adjusted R Square	0.44338054
Standard Error	0.133188898
Observations	47

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0.703216004	0.234405	13.21391	2.99583E-06
Residual	43	0.76278915	0.017739		
Total	46	1.466005154			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	VIF
Intercept	2.905567182	0.166303478	17.47148	3.79E-21	2.570184256	3.240950109	2.57018426	3.24095011	
Po1	0.049157873	0.01074194	4.576256	4E-05	0.027494686	0.07082106	0.02749469	0.07082106	2.64
Wealth	-6.18681E-05	3.50463E-05	-1.76532	0.08461	-0.00013255	8.80956E-06	-0.00013255	8.8096E-06	2.97
Prob	-1.651295952	1.041130141	-1.58606	0.120053	-3.75093499	0.448343081	-3.75093499	0.44834308	1.45

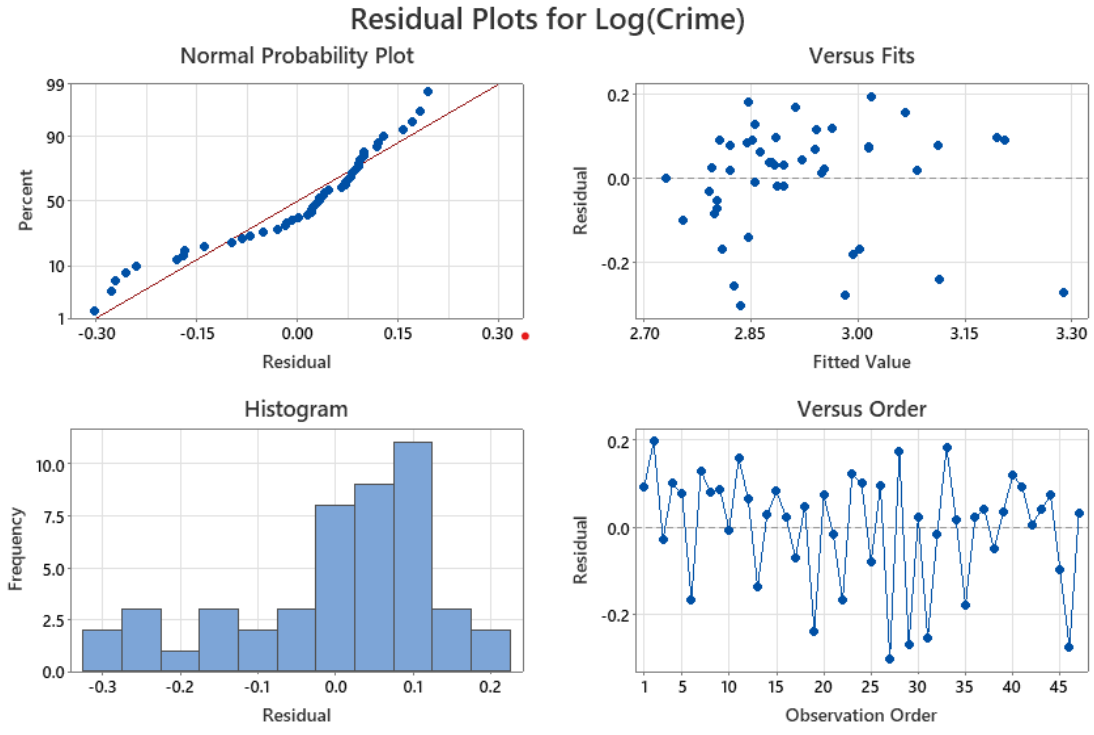
Regression Equation: Log(Crime)=2.906 + 0.0492 Po1 - 0.000062 Wealth - 1.65 Prob

Through Figure 10, we can make several observations:

- The relationship between Log(Y) and the three explanatory variables produced an R^2 value of 0.48 and an adjusted R^2 value of 0.44 which went up from the initial analysis, thus indicating that eliminating Po2 was a good decision
- The F-statistic value increased from 10 to 13.2 and the p-value dramatically decreased which proves that there is a stronger relationship between the dependent variable and exploratory variables and indicates again that the removal of the Po2 variable was a good decision
- The p-values listed for the explanatory variables decreased from the previous analysis, indicating a higher significance
- The Variance Inflation Factor values for most of the variables declined with Po1 experiencing a sharp decrease with a value now between 1 and 5, indicating that the variables are no longer multi-correlated which is a positive discovery
- The Durbin-Watson statistic value here is 2.44 which indicates little to no sign of autocorrelation since it's within the range 1.5-2.5 and is a positive discovery

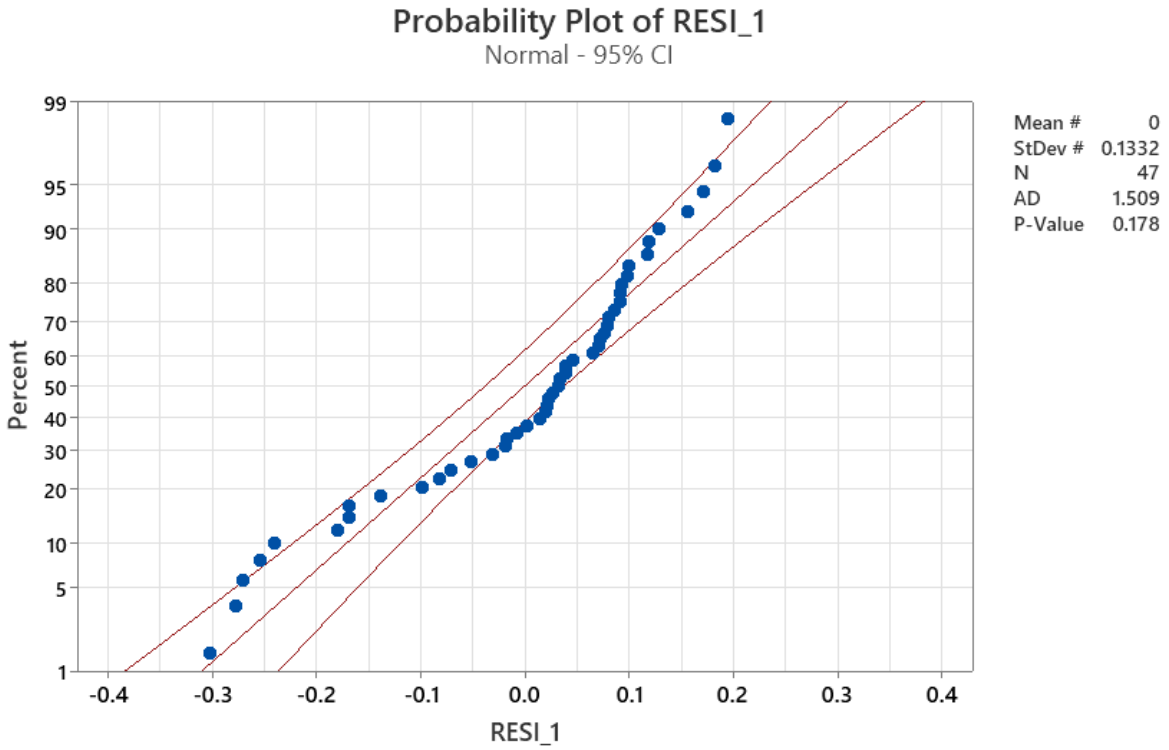
The following plots conducted in Minitab and illustrated in Figure 11 below further support these observations:

Figure 11: Residual Plots for Updated Regression Analysis for Log(Y)



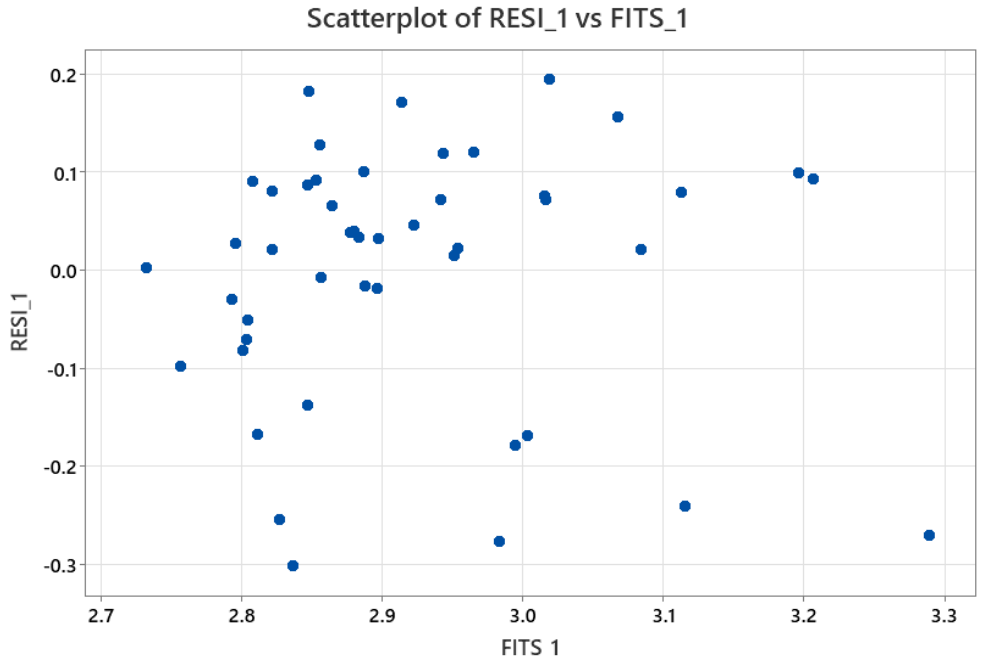
A probability plot of the updated regression residuals is depicted in Figure 12 below:

Figure 12: Probability Plot of Updated Regression Residuals



The updated regression residual versus fitted values plot is also depicted in Figure 13 below:

Figure 13: Updated Regression's Residual versus Fitted Values for Log(Y)



Diagnostic Analysis for Updated Regression Analysis:

The plots represented by the figures above provide graphical evidence of the earlier observations made about the updated regression analysis. In Figure 12, it can be observed that the Residual of the updated regression analysis is not normally distributed as its 0.178 p-value is still below 0.250 and some data points are located outside of the boundaries; furthermore, the four-in-one residual plots in Figure 11 show that most of the data points lie below the residual's zero line and are not as chaotic as preferred; however, the histogram seems more symmetric around the 0 point than the initial regression analysis. Even so, the probability plot continues to prove that the residual is not normally distributed because of its maintained right-skewness from zero. Lastly, Figure 13 unfortunately doesn't depict the data points being randomly distributed horizontally along the residual line of 0 and might indicate heteroskedasticity.

Although the updated regression analysis improved on the initial analysis, as seen through the lower p-values, lower VIF, and a higher R^2 adjusted value, this model can further be improved by adding an interaction term to the model. This will allow us to plot the interaction between the exploratory variables against Log (Y) and create a stronger relationship between them.

Adjusted Regression Analysis:

To construct an improved model that reflects a strong relationship between the dependent and independent variables, interaction terms will be added to the model. After various trial and error and conducting some exploratory data analysis using the Crime Rate data and the correlation matrix from Figure 5, two reasonable interaction terms were chosen.

The table below illustrates the values of the first interaction term added to the model:

Table 2: Crime Rate Data with Interaction Term 1

Crime (Y)	Log(Crime)	Po1	Wealth	Prob	Pop	Po1*Pop
791	2.898	5.8	3940	0.084602	33	191.4
1635	3.214	10.3	5570	0.029599	13	133.9
578	2.762	4.5	3180	0.083401	18	81
1969	3.294	14.9	6730	0.015801	157	2339.3
1234	3.091	10.9	5780	0.041399	18	196.2
682	2.834	11.8	6890	0.034201	25	295
963	2.984	8.2	6200	0.0421	4	32.8
1555	3.192	11.5	4720	0.040099	50	575
856	2.932	6.5	4210	0.071697	39	253.5
705	2.848	7.1	5260	0.044498	7	49.7
1674	3.224	12.1	6570	0.016201	101	1222.1
849	2.929	7.5	5800	0.031201	47	352.5
511	2.708	6.7	5070	0.045302	28	187.6
664	2.822	6.2	5290	0.0532	22	136.4
798	2.902	5.7	4050	0.0691	30	171
946	2.976	8.1	4270	0.052099	33	267.3
539	2.732	6.6	4870	0.076299	10	66
929	2.968	12.3	6310	0.119804	31	381.3
750	2.875	12.8	6270	0.019099	51	652.8
1225	3.088	11.3	6260	0.034801	78	881.4
742	2.870	7.4	5570	0.0228	34	251.6
439	2.642	4.7	2880	0.089502	22	103.4
1216	3.085	8.7	5130	0.0307	43	374.1
968	2.986	7.8	5400	0.041598	7	54.6
523	2.719	6.3	4860	0.069197	14	88.2
1993	3.300	16	6740	0.041698	3	48
342	2.534	6.9	5640	0.036099	6	41.4
1216	3.085	8.2	5370	0.038201	10	82
1043	3.018	16.6	6370	0.0234	168	2788.8
696	2.843	5.8	3960	0.075298	46	266.8
373	2.572	5.5	4530	0.041999	6	33
754	2.877	9	6170	0.042698	97	873
1072	3.030	6.3	4620	0.049499	23	144.9
923	2.965	9.7	5890	0.040799	18	174.6
653	2.815	9.7	5720	0.0207	113	1096.1
1272	3.104	10.9	5590	0.0069	9	98.1
831	2.920	5.8	3820	0.045198	24	139.2
566	2.753	5.1	4250	0.053998	7	35.7
826	2.917	6.1	3950	0.047099	36	219.6
1151	3.061	8.2	4880	0.038801	96	787.2
880	2.944	7.2	5900	0.0251	9	64.8
542	2.734	5.6	4890	0.088904	4	22.4
823	2.915	7.5	4960	0.054902	40	300
1030	3.013	9.5	6220	0.0281	29	275.5
455	2.658	4.6	4570	0.056202	19	87.4
508	2.706	10.6	5930	0.046598	40	424

This first interaction term is created in Table 2 using the Po1 and Pop variables.
The regression model that results from the incorporation of this first interaction term is illustrated in Figure 14 below:

Figure 14: Adjusted Regression Analysis with Interaction Term 1

Regression Statistics	
Multiple R	0.723864757
R Square	0.523980187
Adjusted R Square	0.46592899
Standard Error	0.130463293
Observations	47

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	0.768157655	0.153632	9.026174	7.5852E-06
Residual	41	0.697847499	0.017021		
Total	46	1.466005154			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	VIF
Intercept	2.835544172	0.174269819	16.271	1.72E-19	2.48359913	3.18748921	2.48359913	3.18748921	
Po1	0.064154899	0.013402803	4.786678	2.22E-05	0.03708739	0.09122241	0.03708739	0.09122241	4.29
Wealth	-7.25122E-05	3.56542E-05	-2.03376	0.048481	-0.00014452	-5.071E-07	-0.00014452	-5.0708E-07	3.2
Prob	-1.778242656	1.044662492	-1.70222	0.096286	-3.88798136	0.33149605	-3.88798136	0.33149605	1.52
Pop	0.002775523	0.00188218	1.474632	0.147949	-0.00102562	0.00657666	-0.00102562	0.00657666	13.88
Po1*Pop	-0.000262995	0.000144912	-1.81487	0.076865	-0.00055565	2.966E-05	-0.00055565	2.966E-05	17.18

Regression Equation
Log(Crime)=2.836 + 0.0642 Po1 - 0.000073 Wealth - 1.78 Prob + 0.00278 Pop- 0.000263 pop*po1

The following can be observed from the adjusted regression analysis of the first interaction term:

- The R^2 value has increased from the initial and updated regressions and is now 0.52
- The adjusted R^2 value increased from the updated regression, indicating that the addition of the interaction term is beneficial to the model
- When compared to the Updated Regression, the F value decreased from 13.2 to 9.03 and the significance of F increased; this indicates that the interaction term should not have been added
- The p-value of the Po1, Wealth, and Prob variables decreased, making them more significant and creating coefficients that will be non-zero
- The p-value of its residual plot was above 0.250 which indicates normality of its residuals
- The VIF value increased but remained between 1 and 5 for the three variables that were used in the updated regression in Figure 10; naturally, the VIF of two variables being used to make the interaction term increased which indicates their multicollinearity
- The Durbin-Watson statistic is 2.23 which indicates that there is little to no sign of autocorrelation between the variables' values over time which is a beneficial discovery

The table below illustrates the values of the second interaction term added to the model:

Table 3: Crime Rate Data with Interaction Term 2

Crime (Y)	Log(Crime)	Po1	Wealth	Prob	M	M*Po1
791	2.898	5.8	3940	0.084602	15.1	87.58
1635	3.214	10.3	5570	0.029599	14.3	147.29
578	2.762	4.5	3180	0.083401	14.2	63.9
1969	3.294	14.9	6730	0.015801	13.6	202.64
1234	3.091	10.9	5780	0.041399	14.1	153.69
682	2.834	11.8	6890	0.034201	12.1	142.78
963	2.984	8.2	6200	0.0421	12.7	104.14
1555	3.192	11.5	4720	0.040099	13.1	150.65
856	2.932	6.5	4210	0.071697	15.7	102.05
705	2.848	7.1	5260	0.044498	14	99.4
1674	3.224	12.1	6570	0.016201	12.4	150.04
849	2.929	7.5	5800	0.031201	13.4	100.5
511	2.708	6.7	5070	0.045302	12.8	85.76
664	2.822	6.2	5290	0.0532	13.5	83.7
798	2.902	5.7	4050	0.0691	15.2	86.64
946	2.976	8.1	4270	0.052099	14.2	115.02
539	2.732	6.6	4870	0.076299	14.3	94.38
929	2.968	12.3	6310	0.119804	13.5	166.05
750	2.875	12.8	6270	0.019099	13	166.4
1225	3.088	11.3	6260	0.034801	12.5	141.25
742	2.870	7.4	5570	0.0228	12.6	93.24
439	2.642	4.7	2880	0.089502	15.7	73.79
1216	3.085	8.7	5130	0.0307	13.2	114.84
968	2.986	7.8	5400	0.041598	13.1	102.18
523	2.719	6.3	4860	0.069197	13	81.9
1993	3.300	16	6740	0.041698	13.1	209.6
342	2.534	6.9	5640	0.036099	13.5	93.15
1216	3.085	8.2	5370	0.038201	15.2	124.64
1043	3.018	16.6	6370	0.0234	11.9	197.54
696	2.843	5.8	3960	0.075298	16.6	96.28
373	2.572	5.5	4530	0.041999	14	77
754	2.877	9	6170	0.042698	12.5	112.5
1072	3.030	6.3	4620	0.049499	14.7	92.61
923	2.965	9.7	5890	0.040799	12.6	122.22
653	2.815	9.7	5720	0.0207	12.3	119.31
1272	3.104	10.9	5590	0.0069	15	163.5
831	2.920	5.8	3820	0.045198	17.7	102.66
566	2.753	5.1	4250	0.053998	13.3	67.83
826	2.917	6.1	3950	0.047099	14.9	90.89
1151	3.061	8.2	4880	0.038801	14.5	118.9
880	2.944	7.2	5900	0.0251	14.8	106.56
542	2.734	5.6	4890	0.088904	14.1	78.96
823	2.915	7.5	4960	0.054902	16.2	121.5
1030	3.013	9.5	6220	0.0281	13.6	129.2
455	2.658	4.6	4570	0.056202	13.9	63.94
508	2.706	10.6	5930	0.046598	12.6	133.56

This second interaction term is created in Table 3 using the M and Po1 variables.
The regression model that results from the incorporation of this second interaction term is illustrated in Figure 15 below:

Figure 15: Adjusted Regression Analysis with Interaction Term 2

Regression Statistics	
Multiple R	0.78549106
R Square	0.6169962
Adjusted R Square	0.57028842
Standard Error	0.11702465
Observations	47

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	0.904519606	0.180904	13.20971	1.1021E-07
Residual	41	0.561485548	0.013695		
Total	46	1.466005154			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	VIF
Intercept	3.65382539	0.801321656	4.559749	4.56E-05	2.03552348	5.27212731	2.03552348	5.27212731	
Po1	-0.153864	0.081895018	-1.8788	0.067395	-0.31925439	0.0115263	-0.31925439	0.0115263	198.97
Wealth	-4.111E-05	3.73932E-05	-1.0993	0.278049	-0.00011662	3.4411E-05	-0.00011662	3.4411E-05	4.37
Prob	-1.3829859	0.919539754	-1.504	0.140247	-3.24003411	0.47406231	-3.24003411	0.47406231	1.47
M	-0.0699921	0.053371327	-1.31142	0.197014	-0.17777767	0.0377935	-0.17777767	0.0377935	15.11
M*Po1	0.01574636	0.006349425	2.479966	0.017338	0.00292344	0.02856928	0.00292344	0.02856928	172.47

Regression Equation

Log(Crime)=3.654 - 0.1539 Po1 - 0.000041 Wealth - 1.383 Prob - 0.0700 M + 0.01575 m*po1

The following can be observed from the adjusted regression analysis of the second interaction term:

- The R^2 value is the highest that its been observed with a value of 0.62 which has increased from the initial, updated, and first adjusted regressions
- The adjusted R^2 value also increased and is the highest observed so far with a value of 0.57
- When compared to the Updated Regression, the F value stayed about the same with an F value of 13.2 but the significance of F decreased significantly; this finding supports the addition of the interaction into the model
- The p-value of the Po1, Wealth, and Prob variables increased from the updated regressions, making them less significant and creating coefficients that will be zero
- The p-value of its residual plot was above 0.250 which indicates normality of its residuals
- The VIF value also increased dramatically for all three variables that were used in the updated regression model; more specifically, high inflation factors are observed in the two variables used to create the second interaction term, M and Po1; this indicates their multicollinearity
- The Durbin-Watson statistic with these variables is 2.05 which is a value really close to 2, thus indicating that there is no autocorrelation between the variables's values over time, hence increasing its accuracy because their values are not autocorrelated over time

Comparing Models:

After observing the outcomes of the two regression models that include the interaction terms, it's unclear which one improves the updated regression model most. While the model with the first interaction term has lower VIF values and low p-values that indicate non-zero coefficients for the variables, the model with the second interaction term has a DW statistic closer to 2 and is less autocorrelated, has higher F values, and higher R^2 and R^2 adjusted values.

Thus, a more rigours comparison analysis should be conducted to test for weather model improvements were made to decide weather the increases in the R^2 values for the models with the interaction terms are statistically significant or due to chance.

Figure 16 below tests for model improvements by comparing the Updated Model with the Regression Model that uses the first interaction term which is made possible by its normally distributed residual:

Figure 16: Best Regression Model Fit between Updated Model & First Interaction Term Regression Model

	Interaction 1 model		
	R Square	52.40%	
	Degrees of Freedom (DF)	35	
	Updated model		
	R Square	47.97%	
	Degrees of Freedom	39	
	Difference R-Squared	4.43%	
	Difference Df	4	
		Value	**Df**
	Numerator	0.0111	4
	Denominator	0.0136	35
	F-Statistic	0.814	
	α	5%	
Method 1	**Critical Value**	2.641	**Conclusion**
	Conclusion	No Model Improvement	Fail to Reject H0
Method 2	**p-value**	52.47%	**Conclusion**
	Conclusion	No Model Improvement	Fail to Reject H0
	H_0:	No Model Improvement	
	H_1:	Model Improvement	

This first model improvement test in Figure 16 was conducted on the regression model that incorporated the first interaction term using two methods:

- Method 1 used the F-statistic value that was calculated using the ratio of the numerator and denominator values and resulted in a value of 0.814. When compared with the F critical value of 2.641, it is observed that the F-statistic doesn't lie within the critical value region since it has a smaller value than the critical value. This concluded in a fail to reject decision on the null hypothesis.
- Method 2 used the p-value, which was calculated to be 52.47%. When compared to the 5% level of significance, the p-value is dramatically greater than than alpha, thus the decision of failing to reject the null hypothesis was made.

Thus, both methods proved that the regression model didn't make any model improvements to the Updated regression and the increase in R^2 value observed previous is not statistically significant.

Next, the model improvement test was conducted on the regression model that contains the second interaction term.

Figure 17 below tests for model improvements by comparing the Updated Model with the Regression Model that uses the second interaction term which is made possible by its normally distributed residual:

Figure 17: Best Regression Model Fit between Updated Model & Second Interaction Term Regression Model

	Interaction 2 model		
	R Square	61.70%	
	Degrees of Freedom (DF)	35	
	Updated model		
	R Square	47.97%	
	Degrees of Freedom	39	
	Difference R-Squared	13.73%	
	Difference Df	4	
		Value	Df
	Numerator	0.0343	4
	Denominator	0.0109	35
	F-Statistic	3.137	
	α	5%	
Method 1	Critical Value	2.641	Conclusion
	Conclusion	Model Improvement	Reject H0
Method 2	p-value	2.64%	Conclusion
	Conclusion	Model Improvement	Reject H0
	H_0:	No Model Improvement	
	H_1:	Model Improvement	

This second model improvement test was conducted on the regression model that incorporated the second interaction term using two methods:

- Method 1 used the F-statistic value that was calculated using the ratio of the numerator and denominator values and resulted in a value of 3.137. When compared with the F critical value of 2.641, it is observed that the F-statistic lies within the critical value region since its value is greater than the critical value. This concluded in the rejection of the null hypothesis.
- Method 2 used the p-value, which was calculated to be 2.64%. When compared to the 5% level of significance, the p-value is smaller than the alpha, thus resulting in a decision to reject the null hypothesis.

Thus, both methods proved that the second regression model makes model improvements to the Updated regression and the increase in R^2 value observed previously is due to statistical significance. In summary, it can be said with confidence that the best fit model is represented by the Regression equation: Log(Crime)=3.654 - 0.1539 Po1 - 0.000041 Wealth - 1.383 Prob - 0.0700 M + 0.01575 m*po1

Best Model Fit using Prediction Interval:

Next, a prediction analysis was conducted through Minitab between the Updated Model and the regression model chosen above that incorporated the second interaction term.

The value of the dependent variable was predicted using the provided values for the independent variables below:

- Po1=16
- Wealth=6890
- Prob=0.01
- M=17
- M*Po1=272

Figure 18 below shows the resulting Prediction Fits, Standard Error of the Fits, the confidence intervals, and the prediction intervals:

Figure 18: Prediction of Log(Y) using Updated and Interaction Term 2 Regression

Updated Regression:	PFITS_1	PSEFITS_1	CLIM_2	CLIM_3	PLIM_2	PLIM_3	**PI=**	0.582333
	3.249309	0.05573	3.136919	3.361699	2.958142	3.540475		
Interaction 1:	PFITS_2	PSEFITS_2	CLIM_4	CLIM_5	PLIM_4	PLIM_5	**PI=**	1.014087
	3.988092	0.222128	3.539495	4.436688	3.481048	4.495135		

It can be observed through Figure 18 above that the predicted value of Log(Y) from the Updated Regression is 3.25 while it is 3.99 in the second interaction term. When the prediction widths are calculated, a smaller width is calculated for the Updated Regression.

Since the prediction widths account for both the error and uncertainty of the mean and the coefficients, a smaller prediction interval is preferred. Thus, the final decision is to choose the Updated Regression Model for predicting the values of Log(Y). Through this, the Log(Crime) was forecasted using the above provided values for the variables with a 95% prediction interval for Log(Crime) and 95% confidence interval for the expected value of Log(Crime).

However, the values of Crime itself is required and thus, a transformation of the independent variable is in order.

The below table provides data on the transformation of Log(Y) to Y:

Figure 19: Updated Model's Prediction Value for Crime

Regression Equation of Log(Crime)	2.906 + 0.0492 Po1 - 0.000062 Wealth - 1.65 Prob
Log(Crime)	3.249309
Crime	$10^{3.249309} \rightarrow 1775.45$

The transformation analyzed in Figure 19 above stems from the following conceptual understanding:

- The median of Log(Crime)=3.25, thus the probability of Log(Crime)<3.25 is 50%
- Then, a calculation can be conducted to eliminate Log by using the Power of 10 and multiplying it to both sides of the inequality
- This will result in Crime<$10^{3.25}$ which equals 1778.28 offenses per 100,000 population in 1960

Thus, crime was forecasted with a 95% prediction interval for Crime and a 95% confidence interval for the expected value of Crime.

The prediction of the Initial Regression Model is then analyzed to compared with the Updated Model. Figure 20 below shows the resulting Prediction Fits, Standard Error of the Fits, the confidence intervals, and the prediction intervals:

Figure 20: Prediction of Log(Y) using Initial Regression Model

Initial Regression:		PFITS_3	PSEFITS_3	CLIM_6	CLIM_7	PLIM_6	PLIM_7			PI=	0.585331
		3.249394	0.055978	3.136425	3.362363	2.956728	3.542059				

When this is compared with the initial regression model, it results in the prediction of Crime provided by Figure 21 below:

Figure 21: Intial Model's Prediction Value for Crime

Regression Equation of Log(Crime)	2.892 + 0.0946 Po1 - 0.0499 Po2 - 0.000057 Wealth - 1.62 Prob
Log(Crime)	3.249394
Crime	$10^{3.249394} \rightarrow 1775.80$

When comparing the prediction values of the Initial and Updated Regression models using Figure 19 and 21, similar and almost identical values are observed for their Crime and Log(Crime) values. However, the Updated Model has a smaller prediction width than the Initial model as observed in Figure 20. Thus, the best prediction model is the Updated Regression Model.

After the transformation, the confidence and credibility intervals of Crime is calculated and is represented in Figure 22 below:

Figure 22: Prediction and Confidence Intervals of Updated Regression Model

	x0	b-hat	μ = LOG(CRIME) - hat		3.249		Standard Error Residuals	0.133189
Intercept	1	2.9055672	MEDIAN[CRIME]	1775.45				
Po1	16	0.0491579	E[CRIME]	1860.94			σ^2 = Var[Y]=Var[Log(Crime)]	0.017739
Wealth	6890	-6.19E-05					σ = Standard Deviation [Log(Crime)	0.133189
Prob	0.01	-1.651296						
			95% Confidence Interval			95% Prediction Interval (or Credibility Interval)		
			LB E[LOG(CRIME)]	3.13692		LB LOG(CRIME)	2.958142	
			UB E[LOG(CRIME)]	3.36170		UB LOG(CRIME)	3.540475	
			Approximate 95% Confidence Interval			95% Prediction Interval (or Credibility Interval)		
			LB E[CRIME]	1370.62		CRIME	908.12	
			UB E[CRIME]	2299.85		CRIME	3471.17	

From Figure 22 above, the following observations were made:

- Confidence Interval of Crime= (1370.62, 2299.85)
- The median of Crime= 1775.45 which is also supported by Figure 19 above
- Credability/Prediction Interval of Crime= (908.12, 3471.17)
- The mean of Crime or expected value of crime= 1860.94

In these observations, both the median of Crime and the mean of Crime lie within their respected upper and lower intervals; the median of Crime lies inside the 95% confidence interval and the mean of Crime lies inside the 95% credability interval.

Conclusion and Recommendation: The mean and median Crime Rates that was predicted are extremely large for the population parameter. When contextualizing the data to be in 1960 and from the 47 states, it further brings into question the extremity of these values. Nonetheless, it is also observed that the transformation of the independent varible from Log(Crime) to Crime dismissed the previous normal distribution that the mean and median values shared. When compared to the Initial Regression model, the Updated model has very similar Crime rate findings. However, it has a smaller prediction interval and is thus assumed to better predict the value of the dependent variable. Despite these similar prediction values, there was a dramatic increase in the R^2 value in the Updated Model than the Initial Model and an increase in the adjusted R^2 value was observed. Furthermore, the Updated Model had lower p-values and lower Variance Inflation Factors which indicates that the results are more accurate. T

Therefore, the Updated Model is the recommended model for predicting the Crime Rate in this dataset.