EMSE 6765

DATA ANALYSIS PROJECT – REGRESSION

Crime Analytics - Analyzing Crime Trends Using Linear Regression

By:

Anbu Ezhilmathi Nambi

G33350186

December 07, 2023

Introduction:

Linear regression is a statistical method that models the relationship between a dependent variable and independent variables. Criminologists are inherently concerned with understanding the dynamics between punishment regimes and crime rates, delving into the intricate relationship between societal consequences and criminal behavior. An examination of patterns and trends reveals the origins of crime, thereby guiding the development of data-driven approaches for contemporary crime prevention and law enforcement. The objectives are as follows:

- To construct a linear regression model for the dataset, using the log of the crime rate, with a focus on identifying the most effective predictor variables.
- To perform a diagnostic analysis on the fitted model.
- To predict the crime rate, along with a 95% confidence interval, the model will used for the following independent variables:
  X1 = 16, X2 = 15, X3 = 6890, X4 = 0.01, X5 = 168, X6 = 12, X7 = 0.14, X8 = 5, X9 = 0.6, X10 = 107, X11 = 27, X12 = 44, X13 = 17.

Dataset Overview:

The analysis utilizes aggregated crime rate data from 1960 across 47 states in the USA. The original dataset is outlined in Table 1 below.

The dependent variables considered are:

- Crime Rate is represented as Y: Number of offenses per 100,000 population in 1960

- The Log of Crime Rate is represented as Log(Y)

The independent variables include:

- Po1(X1): Per capita expenditure in police protection in 1960

- Po1(X2): Per capita expenditure in police protection in 1959

- Wealth(X3): Median value of transferrable assets or family income

- Prob(X4): Probability of imprisonment: ratio of number of commitments to a number of offenses

- Pop(X5): State population in 1960 in hundred thousand

- Ed(X6): Mean years of schooling of the population aged 25 years or over

- U1(X7): An unemployment rate of urban males 14–24

- U2(X8): An unemployment rate of urban males 35–39–24s

- LF(X9): Labor force participation rate of civilian urban male in the age group 14–24

- M.F.(X10): Number of males per 100 females

- Ineq(X11): Income inequality: percentage of families earning below half the median income

- Time(X12): Average time in months served by offenders in state prisons before their first release

- M(X13): Percentage of males aged 14–24 in total state population.

Table 1: Original Performance Data

	Crime (Y)	Log(Crime)	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
Data	Y	Log(Y)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
1	791	2.898	5.8	5.6	3940	0.0846	33	9.1	0.108	4.1	0.51	95	26.1	26.2011	15.1
2	1635	3.214	10.3	9.5	5570	0.0296	13	11.3	0.096	3.6	0.583	101.2	19.4	25.2999	14.3
3	578	2.762	4.5	4.4	3180	0.0834	18	8.9	0.094	3.3	0.533	96.9	25	24.3006	14.2
4	1969	3.294	14.9	14.1	6730	0.0158	157	12.1	0.102	3.9	0.577	99.4	16.7	29.9012	13.6
5	1234	3.091	10.9	10.1	5780	0.0414	18	12.1	0.091	2	0.591	98.5	17.4	21.2998	14.1
6	682	2.834	11.8	11.5	6890	0.0342	25	11	0.084	2.9	0.547	96.4	12.6	20.9995	12.1
7	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
8	1555	3.192	11.5	10.9	4720	0.0401	50	10.9	0.079	3.5	0.542	96.9	20.6	24.5988	13.1
9	856	2.932	6.5	6.2	4210	0.0717	39	9	0.081	2.8	0.553	95.5	23.9	29.4001	15.7
10	705	2.848	7.1	6.8	5260	0.0445	7	11.8	0.1	2.4	0.632	102.9	17.4	19.5994	14
11	1674	3.224	12.1	11.6	6570	0.0162	101	10.5	0.077	3.5	0.58	96.6	17	41.6	12.4
12	849	2.929	7.5	7.1	5800	0.0312	47	10.8	0.083	3.1	0.595	97.2	17.2	34.2984	13.4
13	511	2.708	6.7	6	5070	0.0453	28	11.3	0.077	2.5	0.624	97.2	20.6	36.2993	12.8
14	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
15	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
16	946	2.976	8.1	7.7	4270	0.0521	33	8.8	0.116	4.7	0.497	95.6	24.7	26.0991	14.2
17	539	2.732	6.6	6.3	4870	0.0763	10	11	0.114	3.5	0.537	97.7	16.6	19.1002	14.3
18	929	2.968	12.3	11.5	6310	0.1198	31	10.4	0.089	3.4	0.537	97.8	16.5	18.1996	13.5
19	750	2.875	12.8	12.8	6270	0.0191	51	11.6	0.078	3.4	0.536	93.4	13.5	24.9008	13
20	1225	3.088	11.3	10.5	6260	0.0348	78	10.8	0.13	5.8	0.567	98.5	16.6	26.401	12.5
21	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
22	439	2.642	4.7	4.4	2880	0.0895	22	8.9	0.097	3.4	0.512	96.2	27.6	37.0994	15.7
23	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
24	968	2.986	7.8	7.3	5400	0.0416	7	11.6	0.142	4.2	0.574	103.8	17.6	17.6	13.1
25	523	2.719	6.3	5.7	4860	0.0692	14	11.6	0.07	2.1	0.641	98.4	19.6	21.9003	13
26	1993	3.300	16	14.3	6740	0.0417	3	12.1	0.102	4.1	0.631	107.1	15.2	22.1005	13.1
27	342	2.534	6.9	7.1	5640	0.0361	6	10.9	0.08	2.2	0.54	96.5	13.9	28.4999	13.5
28	1216	3.085	8.2	7.6	5370	0.0382	10	11.2	0.103	2.8	0.571	101.8	21.5	25.8006	15.2
29	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
30	696	2.843	5.8	5.4	3960	0.0753	46	8.9	0.072	2.6	0.521	97.3	23.7	28.3011	16.6
31	373	2.572	5.5	5.4	4530	0.042	6	9.3	0.135	4	0.535	104.5	20	21.7998	14
32	754	2.877	9	8.1	6170	0.0427	97	10.9	0.105	4.3	0.586	96.4	16.3	30.9014	12.5
33	1072	3.030	6.3	6.4	4620	0.0495	23	10.4	0.076	2.4	0.56	97.2	23.3	25.5005	14.7
34	923	2.965	9.7	9.7	5890	0.0408	18	11.8	0.102	3.5	0.542	99	16.6	21.6997	12.6
35	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
36	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
37	831	2.920	5.8	5.6	3820	0.0452	24	8.7	0.076	2.8	0.638	97.4	25.4	31.6995	17.7
38	566	2.753	5.1	4.7	4250	0.054	7	10.4	0.099	2.7	0.599	102.4	22.5	16.6999	13.3
39	826	2.917	6.1	5.4	3950	0.0471	36	8.8	0.086	3.5	0.515	95.3	25.1	27.3004	14.9
40	1151	3.061	8.2	7.4	4880	0.0388	96	10.4	0.088	3.1	0.56	98.1	22.8	29.3004	14.5
41	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
42	542	2.734	5.6	5.4	4890	0.0889	4	10.9	0.107	3.7	0.523	96.8	17	12.1996	14.1
43	823	2.915	7.5	7	4960	0.0549	40	9.9	0.073	2.7	0.522	99.6	22.4	31.9989	16.2
44	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
45	455	2.658	4.6	4.1	4570	0.0562	19	8.8	0.135	5.3	0.48	96.8	24.9	32.5996	13.9
46	508	2.706	10.6	9.7	5930	0.0466	40	10.4	0.078	2.5	0.599	98.9	17.1	16.6999	12.6
47	849	2.929	9	9.1	5880	0.0528	3	12.1	0.113	4	0.623	104.9	16	16.0997	13

While analyzing the relationship between the dependent variable (Y) and its corresponding independent variables (X1-X13) in the original performance data. The observations are as follows:

- The distribution of the dependent variable is not normal. As illustrated in Figure 1 below, the distribution of the dependent variable (Y) shows right skewness.
- The probability plot for the data, presented in Figure 2, highlights deviations of some data points from the trend line, and it shows a noticeable standard deviation.

These insights are visually depicted in the histogram and probability plot generated by Minitab, as shown below:

Figure 1: Histogram of Crime Rate (Y)

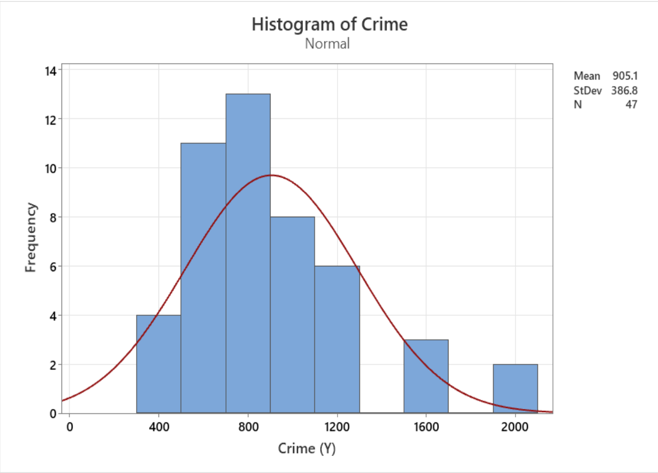
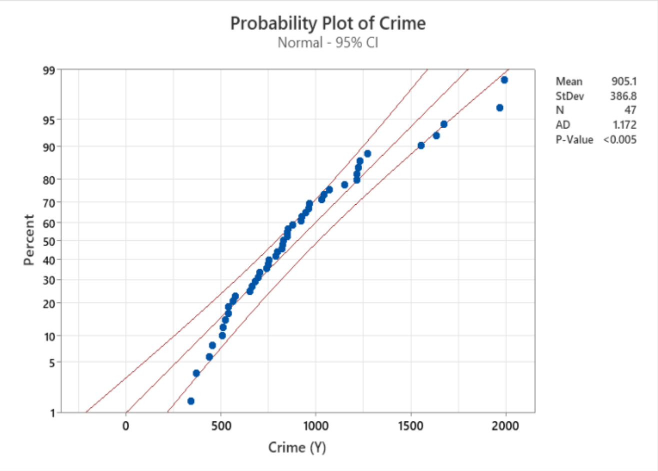


Figure 2: Probability Plot of Crime Rate (Y)



In the goal of achieving a normalized distribution for the dependent variable, log transformation is a commonly applied technique to modify the scale of the dependent variable (Y). Log(Y) displays a better normal distribution compared to the original dependent variable (Y). The analysis of Log(Y) revealed the following observations:

- The distribution of the dependent variable is normal. As illustrated in Figure 3 below.
- The probability plot for the data, presented in Figure 4, highlights that the data points perfectly fits the trendline.
- There is a significant reduction in the standard deviation of the Log(Crime rate). A more favorable P-value is evident, signifying an elevated level of normality in the distribution.
- The regression model will use Log(Crime rate) as the dependent variable.

The observations are visually illustrated in the histogram and probability plot generated by Minitab, as presented below:

Figure 3: Histogram of Log Crime Rate (Log(Y))

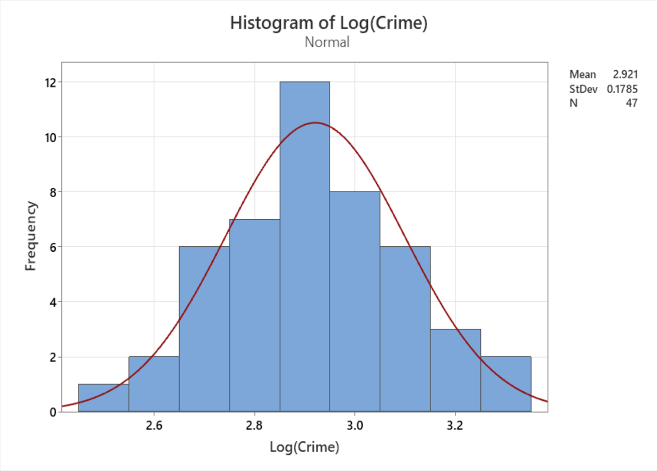
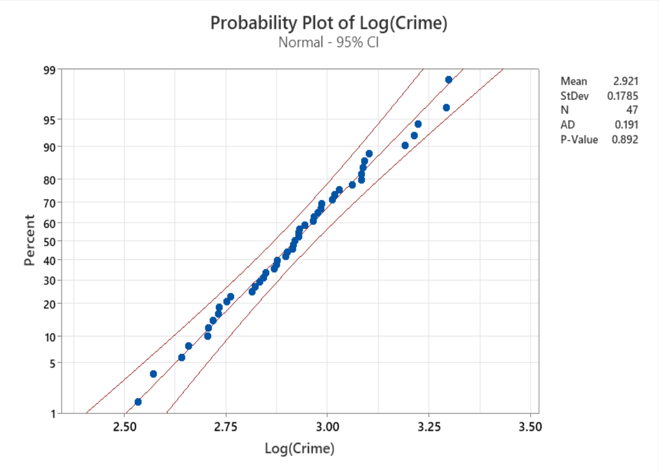


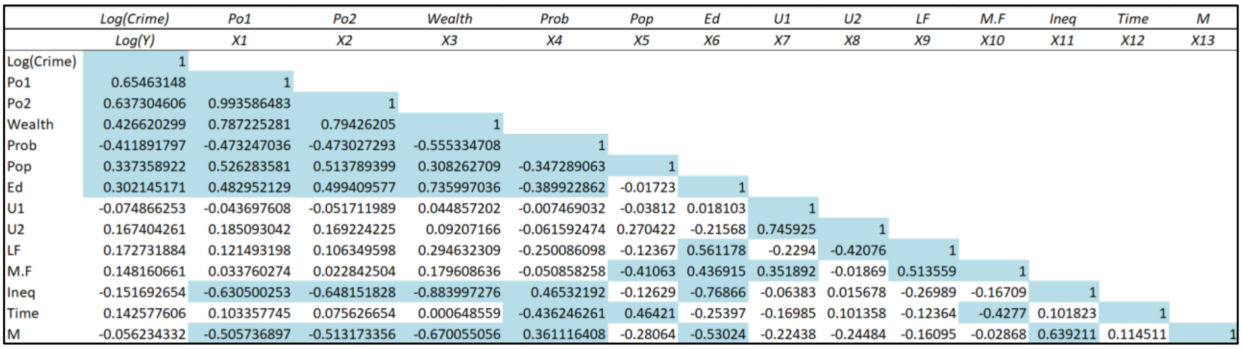
Figure 4: Probability Plot of Log Crime Rate (Log(Y))



Correlation Analysis:

A correlation analysis was conducted on the independent variables (X1-X13) and the dependent variable Log(Y) to identify the most effective predictors for the regression model. The result of the analysis is illustrated in Table 2 below, which shows the correlations between the dependent and independent variables.

Table 2: Correlation between Log(Crime) and X1-X13



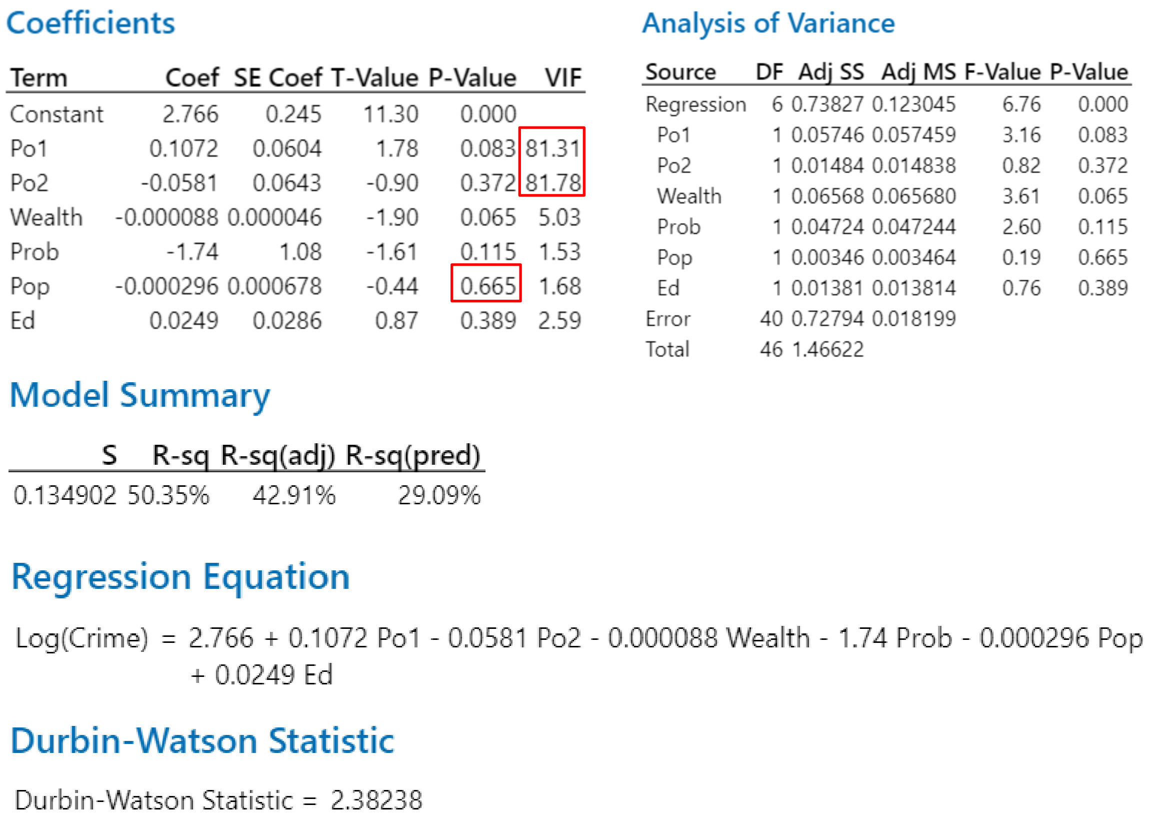
	Log(Crime)	*Po1*	*Po2*	*Wealth*	*Prob*	*Pop*	*Ed*	*U1*	*U2*	*LF*	*M.F*	*Ineq*	*Time*	*M*
	Log(Y)	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*X8*	*X9*	*X10*	*X11*	*X12*	*X13*
Log(Crime)	1													
Po1	0.65463148	1												
Po2	0.637304606	0.993586483	1											
Wealth	0.426620299	0.787225281	0.79426205	1										
Prob	-0.411891797	-0.473247036	-0.473027293	-0.555334708	1									
Pop	0.337358922	0.526283581	0.513789399	0.308262709	-0.347289063	1								
Ed	0.302145171	0.482952129	0.499409577	0.735997036	-0.389922862	-0.01723	1							
U1	-0.074866253	-0.043697608	-0.051711989	0.044857202	-0.007469032	-0.03812	0.018103	1						
U2	0.167404261	0.185093042	0.169224225	0.09207166	-0.061592474	0.270422	-0.21568	0.745925	1					
LF	0.172731884	0.121493198	0.106349598	0.294632309	-0.250086098	-0.12367	0.561178	-0.2294	-0.42076	1				
M.F	0.148160661	0.033760274	0.022842504	0.179608636	-0.050858258	-0.41063	0.436915	0.351892	-0.01869	0.513559	1			
Ineq	-0.151692654	-0.630500253	-0.648151828	-0.883997276	0.46532192	-0.12629	-0.76866	-0.06383	0.015678	-0.26989	-0.16709	1		
Time	0.142577606	0.103357745	0.075626654	0.000648559	-0.436246261	0.46421	-0.25397	-0.16985	0.101358	-0.12364	-0.4277	0.101823	1	
M	-0.056234332	-0.505736897	-0.513173356	-0.670055056	0.361116408	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.639211	0.114511	1

The chosen threshold to indicate significant correlation is set at 0.3, which includes values between -0.3 and 0.3. Highlighted values within this range are considered to demonstrate a significant correlation. Based on this criterion, the following independent variables are considered to be the best predictors for the regression model: Po1 (X1), Po2 (X2), Wealth (X3), Prob (X4), Pop (X5), and Ed (X6).

Initial Model Analysis (Model 0):

Based on the above decision, moving forward with regression analysis using the selected predictors from the correlation analysis. The results of the initial regression analysis are displayed in Figure 5 shown below.

Figure 5: Initial Regression Analysis for Log(Y) and X1-X6



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.766	0.245	11.30	0.000	
Po1	0.1072	0.0604	1.78	0.083	81.31
Po2	-0.0581	0.0643	-0.90	0.372	81.78
Wealth	-0.000088	0.000046	-1.90	0.065	5.03
Prob	-1.74	1.08	-1.61	0.115	1.53
Pop	-0.000296	0.000678	-0.44	0.665	1.68
Ed	0.0249	0.0286	0.87	0.389	2.59

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	0.73827	0.123045	6.76	0.000
Po1	1	0.05746	0.057459	3.16	0.083
Po2	1	0.01484	0.014838	0.82	0.372
Wealth	1	0.06568	0.065680	3.61	0.065
Prob	1	0.04724	0.047244	2.60	0.115
Pop	1	0.00346	0.003464	0.19	0.665
Ed	1	0.01381	0.013814	0.76	0.389
Error	40	0.72794	0.018199		
Total	46	1.46622			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.134902	50.35%	42.91%	29.09%

Regression Equation

Log(Crime) = 2.766 + 0.1072 Po1 - 0.0581 Po2 - 0.000088 Wealth - 1.74 Prob - 0.000296 Pop + 0.0249 Ed

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.38238

Observations:

The observations made from the initial regression analysis in Figure 5 are as follows:

- The VIFs obtained from Minitab indicate a high correlation between Po1 (X1) and Po2 (X2). Therefore, Po2 (X2) is excluded from the next model.
- Although the VIF for Wealth (X3) is slightly above 5, it is retained for the next model due to its moderate collinearity. The VIF for the remaining variables is below 5, ensuring that collinearity is not a significant concern in the model.
- Pop (X5) has the highest P-value among the independent variables and should be excluded for the next model.
- The F-value, which measures the overall fit of the regression model, is reasonably high at 6.76. A higher F-value indicates a better fit.

- The R-squared value of the current model is 50.35%, indicating the model can be improved for a higher R-squared value.
- The current model has an Adj R-squared of 42.91%. A higher Adj R-squared value is considered better.
- The Durbin-Watson statistic is 2.38238, close to the ideal value of 2, suggesting independence of residuals and minimal autocorrelation as shown in Figure 6 of the fitted value vs residuals.
- The residual histogram of the initial model shown in Figure 6 is more left-skewed than the normal distribution.
- The residual's probability plot in Figure 7 displays some deviation from normality, and the residuals exhibit an imperfect fit to the trend line. One outlier is identified within the probability plot for the residuals.

Figure 6: Residual Plots for Log(Y)

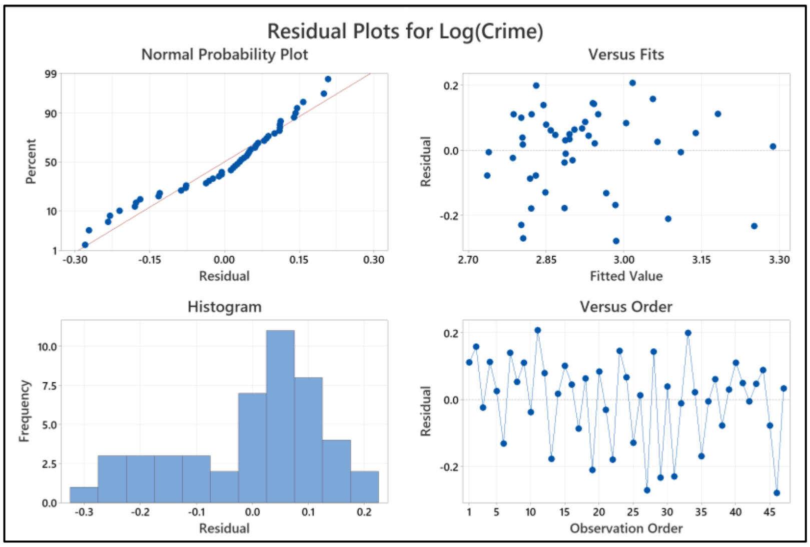
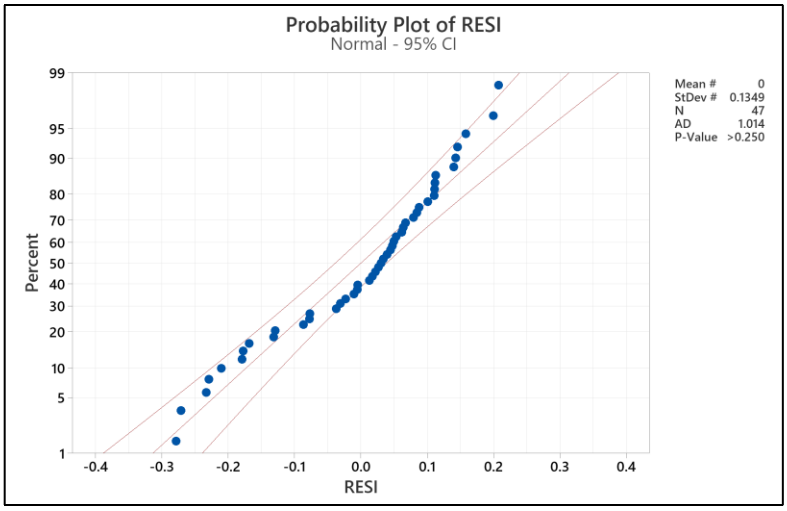


Figure 7: Probability Plots for Residuals



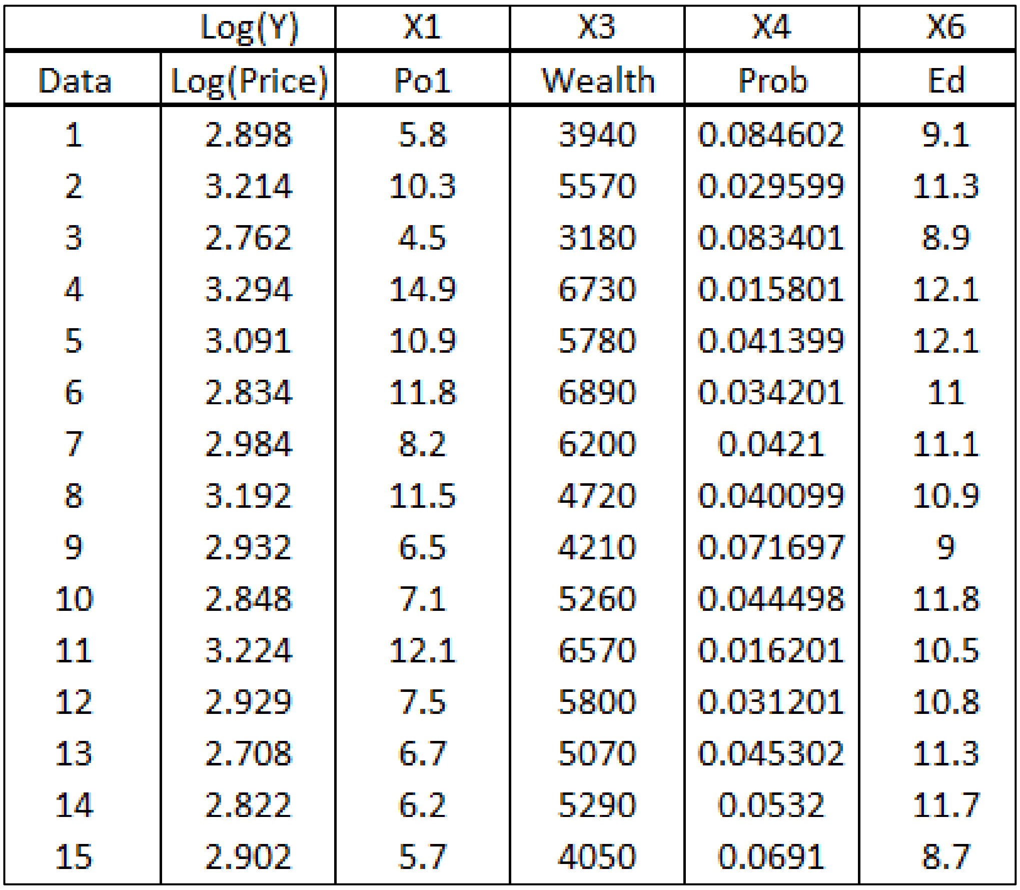
Inference:

The initial analysis of the model has revealed the necessity for improvements. Specifically, the issue of collinearity needs to be addressed, and the fit of the residuals must be improved. One potential way to achieve this is by modifying the model by excluding certain variables. By making these adjustments, it is possible to obtain a more accurate and improved regression model with higher R-squared and Adj R-squared values. Therefore, improving the model is important to increase the predictive performance and overall fit.

Adjusted Model Analysis (Model 1):

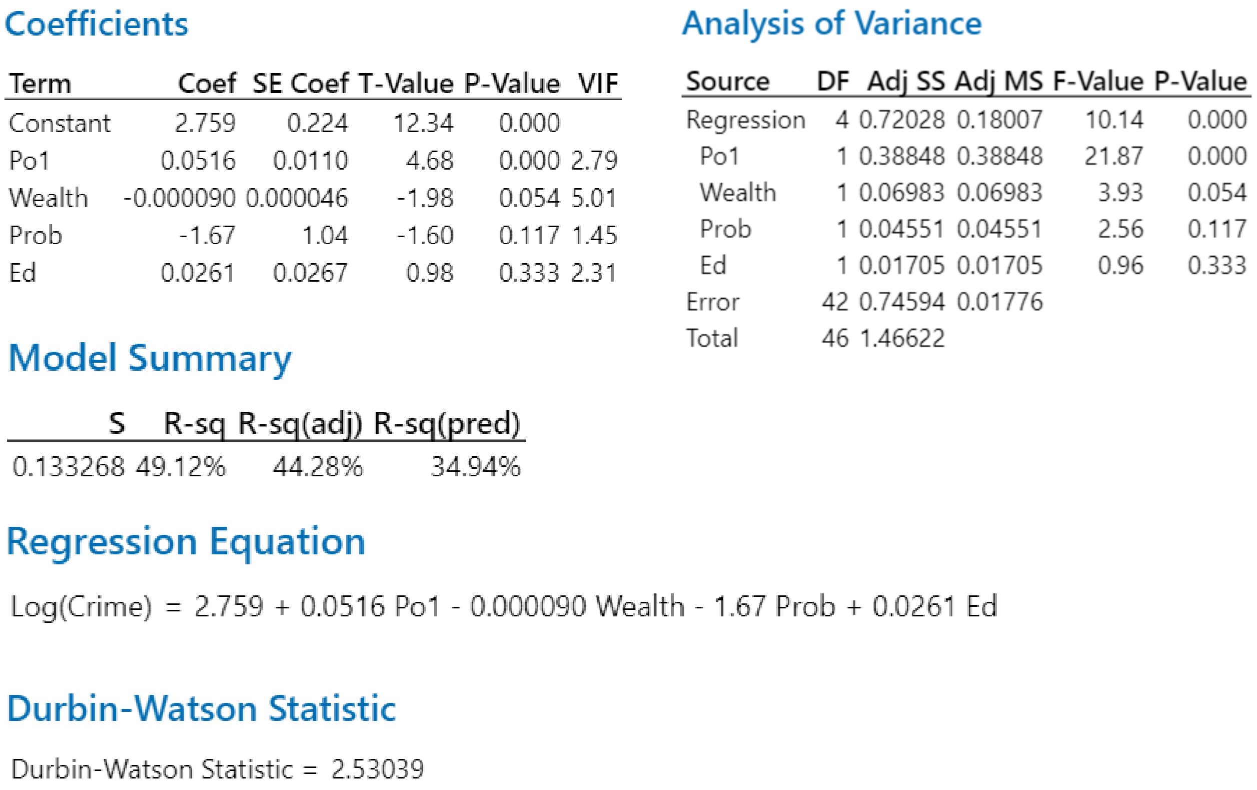
The four independent variables—Po1 (X1), Wealth (X3), Prob (X4), and Ed (X6) - are retained based on the threshold correlation of 0.3 between the dependent variable and independent variables. Given the unfavorable nature of the initial model, a modification is implemented, resulting in an adjusted model now referred to as Model 1. The relevant data for Model 1 is presented in Table 3.

Table 3 : Adjusted Model Table for Model 1



	Log(Y)	X1	X3	X4	X6
Data	Log(Price)	Po1	Wealth	Prob	Ed
1	2.898	5.8	3940	0.084602	9.1
2	3.214	10.3	5570	0.029599	11.3
3	2.762	4.5	3180	0.083401	8.9
4	3.294	14.9	6730	0.015801	12.1
5	3.091	10.9	5780	0.041399	12.1
6	2.834	11.8	6890	0.034201	11
7	2.984	8.2	6200	0.0421	11.1
8	3.192	11.5	4720	0.040099	10.9
9	2.932	6.5	4210	0.071697	9
10	2.848	7.1	5260	0.044498	11.8
11	3.224	12.1	6570	0.016201	10.5
12	2.929	7.5	5800	0.031201	10.8
13	2.708	6.7	5070	0.045302	11.3
14	2.822	6.2	5290	0.0532	11.7
15	2.902	5.7	4050	0.0691	8.7

Figure 8: Model 1 Regression Analysis for Log(Y) and X1, X3, X4 and X6



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.759	0.224	12.34	0.000	
Po1	0.0516	0.0110	4.68	0.000	2.79
Wealth	-0.000090	0.000046	-1.98	0.054	5.01
Prob	-1.67	1.04	-1.60	0.117	1.45
Ed	0.0261	0.0267	0.98	0.333	2.31

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.72028	0.18007	10.14	0.000
Po1	1	0.38848	0.38848	21.87	0.000
Wealth	1	0.06983	0.06983	3.93	0.054
Prob	1	0.04551	0.04551	2.56	0.117
Ed	1	0.01705	0.01705	0.96	0.333
Error	42	0.74594	0.01776		
Total	46	1.46622			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.133268	49.12%	44.28%	34.94%

Regression Equation

Log(Crime) = 2.759 + 0.0516 Po1 - 0.000090 Wealth - 1.67 Prob + 0.0261 Ed

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.53039

Observations:

The observations made from Model 1 regression analysis in Figure 8 are as follows:

- All variables have a VIF below 5, indicating that collinearity is not a significant concern in the model. However, Wealth (X3) shows moderate collinearity.
- The F-value in Model 1 is higher than the Initial Model, which is preferred.
- The P-value for the individual variables is significantly less.
- Although the R-squared drops to 49.12% compared to the Initial Model, the adjusted R-squared value has increased to 44.28%, which is a favorable outcome.
- The Durbin-Waston statistic is equal to 2.53039, slightly above 2.5, suggesting slight positive autocorrelation in the residuals as shown in Figure 9 of the fitted value vs residuals.
- In Figure 9, a four-in-one plot is presented for analyzing the behavior of residuals. The histogram indicates a left skew, and the normal probability plot of data points shows a slight curve, suggesting that the trend line fit is not perfect.
- The residual's probability plot in Figure 10 indicates a better fit to the trend line. However, one outlier is identified within the probability plot for the residuals.

Figure 9: Residual Plots for Log(Y)

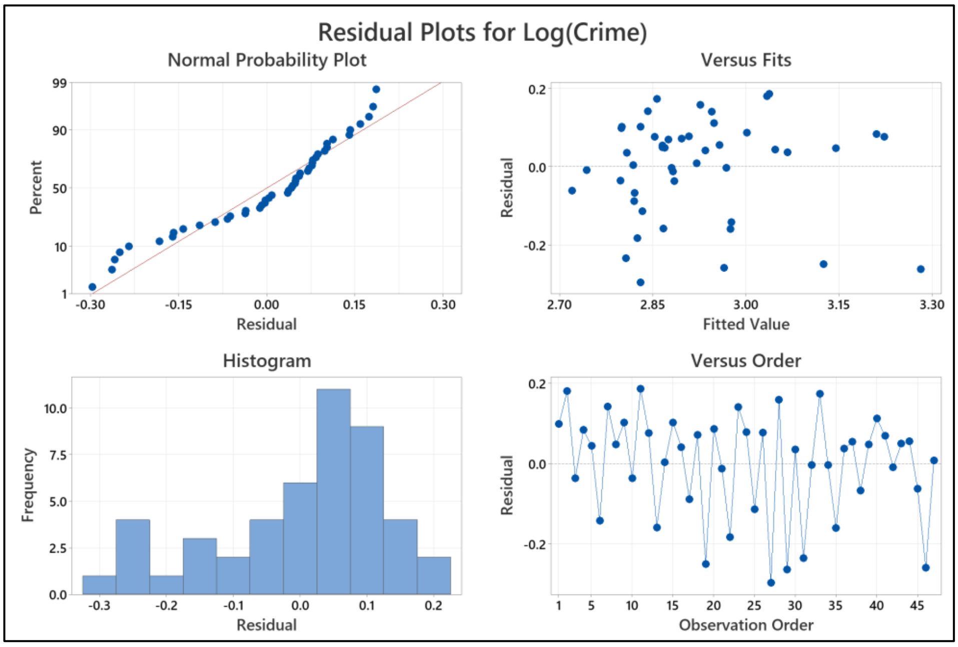
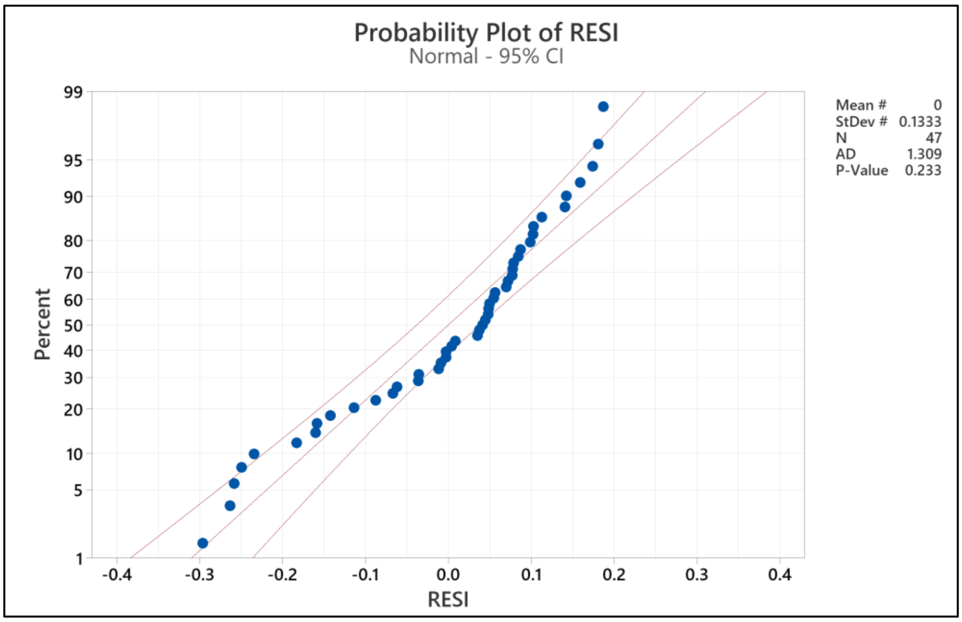


Figure 10: Probability Plots for Residuals



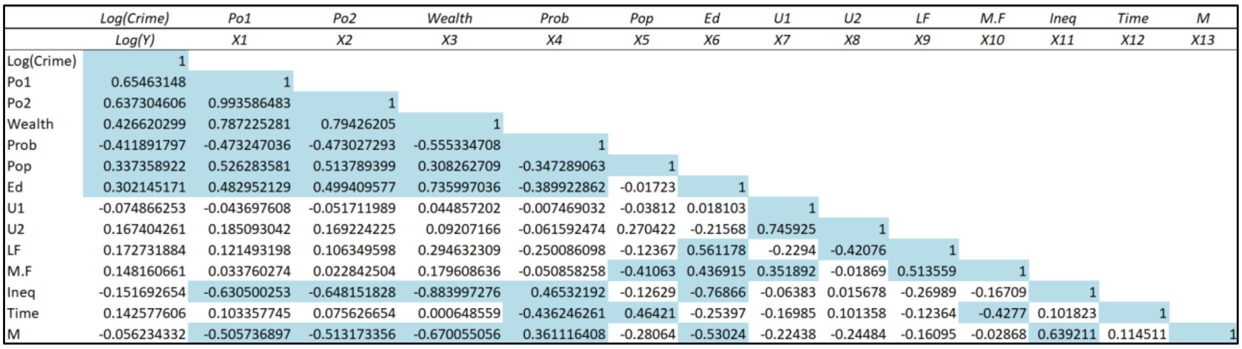
Inference:

The adjusted analysis of Model 1 has identified the need for improvements, particularly in improving the fit of the residuals. While the adjusted R-squared value showed significant improvement after removing some variables when compared to the initial model, further modification is essential. Modifying the model through adjustments can lead to a more accurate and improved regression model with higher R-squared and adjusted R-squared values.

Adjusted Model Analysis (Model 2):

Model 1 with R-squared 49.12% is not enough to output a good prediction value. The next objective is to identify the optimal independent variables to consider. On reviewing the correlation table in Table 2, it is evident that Ineq (X11) is highly correlated with the existing variables X1, X3, X4, and X6.

Table 2: Correlation between Log(Crime) and X1-X13

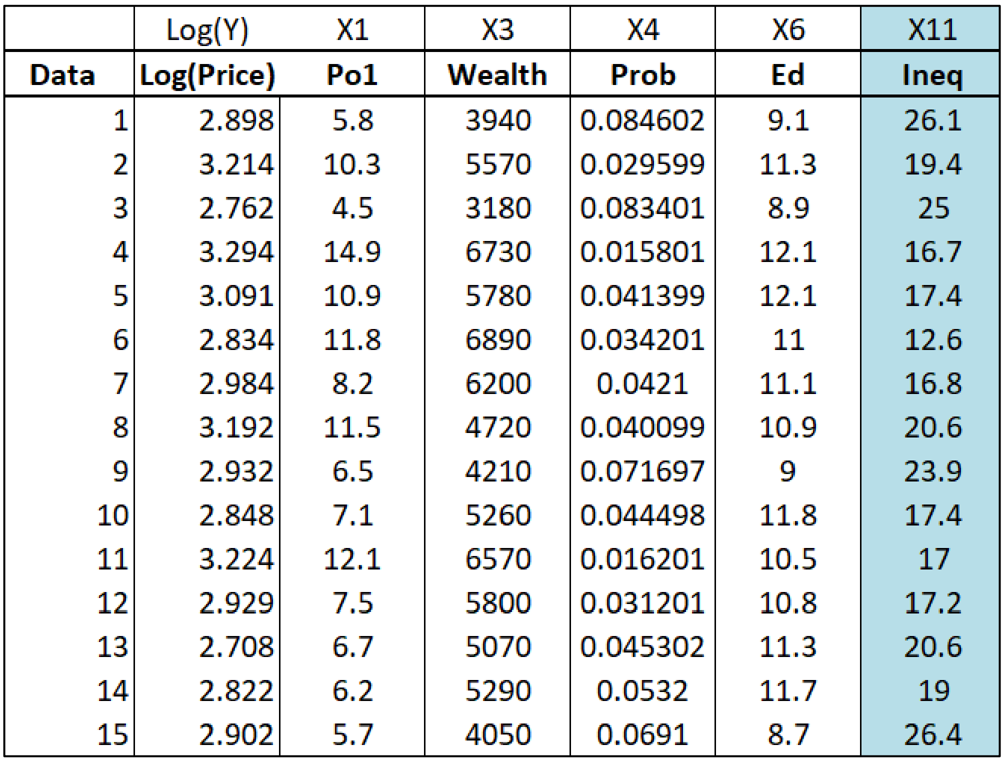


	Log(Crime)	*Po1*	*Po2*	*Wealth*	*Prob*	*Pop*	*Ed*	*U1*	*U2*	*LF*	*M.F*	*Ineq*	*Time*	*M*
	Log(Y)	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*X8*	*X9*	*X10*	*X11*	*X12*	*X13*
Log(Crime)	1													
Po1	0.65463148	1												
Po2	0.637304606	0.993586483	1											
Wealth	0.426620299	0.787225281	0.79426205	1										
Prob	-0.411891797	-0.473247036	-0.473027293	-0.555334708	1									
Pop	0.337358922	0.526283581	0.513789399	0.308262709	-0.347289063	1								
Ed	0.302145171	0.482952129	0.499409577	0.735997036	-0.389922862	-0.01723	1							
U1	-0.074866253	-0.043697608	-0.051711989	0.044857202	-0.007469032	-0.03812	0.018103	1						
U2	0.167404261	0.185093042	0.169224225	0.09207166	-0.061592474	0.270422	-0.21568	0.745925	1					
LF	0.172731884	0.121493198	0.106349598	0.294632309	-0.250086098	-0.12367	0.561178	-0.2294	-0.42076	1				
M.F	0.148160661	0.033760274	0.022842504	0.179608636	-0.050858258	-0.41063	0.436915	0.351892	-0.01869	0.513559	1			
Ineq	-0.151692654	-0.630500253	-0.648151828	-0.883997276	0.46532192	-0.12629	-0.76866	-0.06383	0.015678	-0.26989	-0.16709	1		
Time	0.142577606	0.103357745	0.075626654	0.000648559	-0.436246261	0.46421	-0.25397	-0.16985	0.101358	-0.12364	-0.4277	0.101823	1	
M	-0.056234332	-0.505736897	-0.513173356	-0.670055056	0.361116408	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.639211	0.114511	1

Additionally, given the potential significance of income inequality in socioeconomic contexts and its potential impact on crime rates, there is a compelling reason to include this variable.

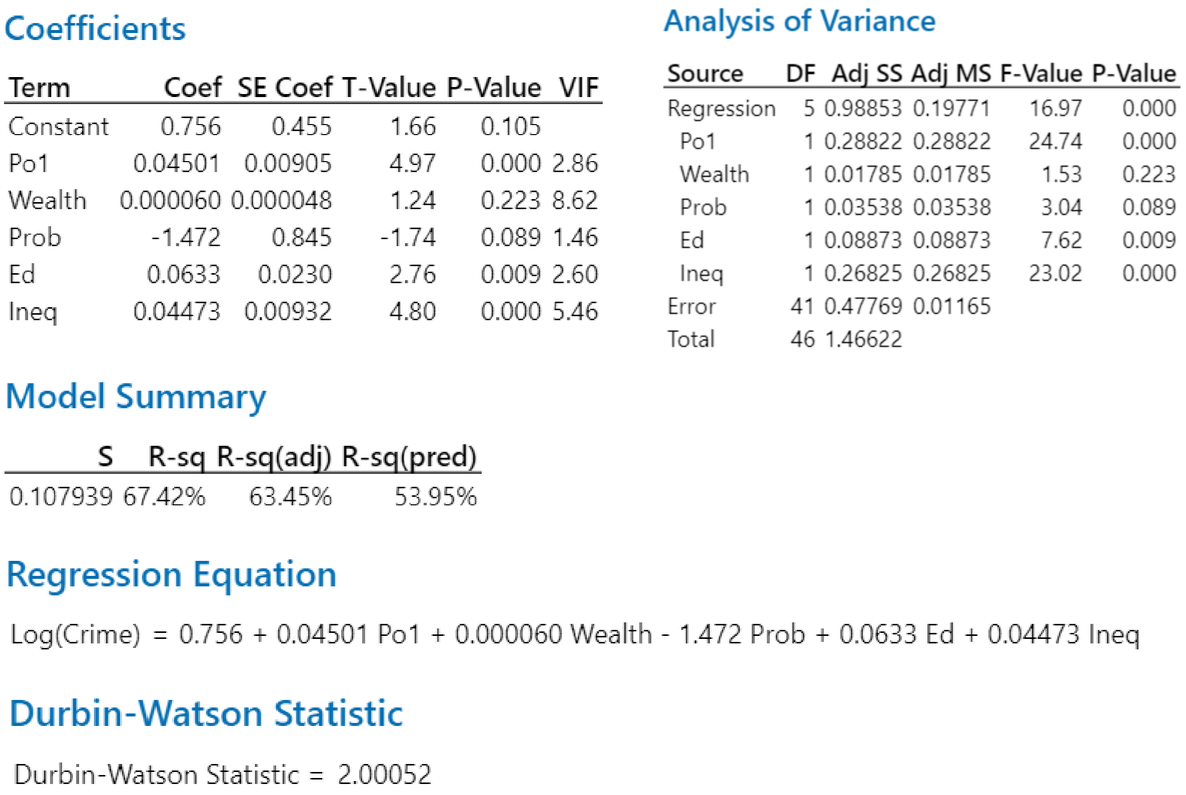
By adding Ineq(X11) has a notable improvement. This modified model, tested with the addition of Ineq (X11), is considered as Model 2. The data of the addition of new independent variables are displayed in Table 4.

Table 4: Adjusted Model Table for Model 2



	Log(Y)	X1	X3	X4	X6	X11
Data	**Log(Price)**	**Po1**	**Wealth**	**Prob**	**Ed**	**Ineq**
1	2.898	5.8	3940	0.084602	9.1	26.1
2	3.214	10.3	5570	0.029599	11.3	19.4
3	2.762	4.5	3180	0.083401	8.9	25
4	3.294	14.9	6730	0.015801	12.1	16.7
5	3.091	10.9	5780	0.041399	12.1	17.4
6	2.834	11.8	6890	0.034201	11	12.6
7	2.984	8.2	6200	0.0421	11.1	16.8
8	3.192	11.5	4720	0.040099	10.9	20.6
9	2.932	6.5	4210	0.071697	9	23.9
10	2.848	7.1	5260	0.044498	11.8	17.4
11	3.224	12.1	6570	0.016201	10.5	17
12	2.929	7.5	5800	0.031201	10.8	17.2
13	2.708	6.7	5070	0.045302	11.3	20.6
14	2.822	6.2	5290	0.0532	11.7	19
15	2.902	5.7	4050	0.0691	8.7	26.4

Figure 11: Model 2 Regression Analysis for Log(Y) and X1, X3, X4, X6 and X11



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.756	0.455	1.66	0.105	
Po1	0.04501	0.00905	4.97	0.000	2.86
Wealth	0.000060	0.000048	1.24	0.223	8.62
Prob	-1.472	0.845	-1.74	0.089	1.46
Ed	0.0633	0.0230	2.76	0.009	2.60
Ineq	0.04473	0.00932	4.80	0.000	5.46

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	0.98853	0.19771	16.97	0.000
Po1	1	0.28822	0.28822	24.74	0.000
Wealth	1	0.01785	0.01785	1.53	0.223
Prob	1	0.03538	0.03538	3.04	0.089
Ed	1	0.08873	0.08873	7.62	0.009
Ineq	1	0.26825	0.26825	23.02	0.000
Error	41	0.47769	0.01165		
Total	46	1.46622			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.107939	67.42%	63.45%	53.95%

Regression Equation

Log(Crime) = 0.756 + 0.04501 Po1 + 0.000060 Wealth - 1.472 Prob + 0.0633 Ed + 0.04473 Ineq

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.00052

Observations:

The observations made from Model 2 in Figure 11 are as follows:

- The VIF for Wealth and Ineq exceeds 5, signifying a moderate level of collinearity between these two variables.
- The F-value in Model 2 is notably high, which is preferred.
- Among the independent variables, Wealth (X3) and Prob (X4) have relatively higher P-values compared to others.
- The R-squared in Model 2, standing at 67.42%, exhibits a significant improvement compared to the previous model. The adjusted R-Square value has increased to 63.45%, indicating a favorable outcome.
- The Durbin-Watson statistic is 2.00052, indicating independence of residuals and minimal autocorrelation.
- The four-in-one plot presented in Figure 12 for analyzing the behavior of residuals. The histogram indicates a left skew, and the normal probability plot of data points shows a slight curve, suggesting that the trend line fit is not perfect.

- The probability plot in Figure 13 shows that the data points exhibit a better fit to the trend line, although one outlier is identified within the probability plot for the residuals.

Figure 12: Residual Plots for Log(Y)

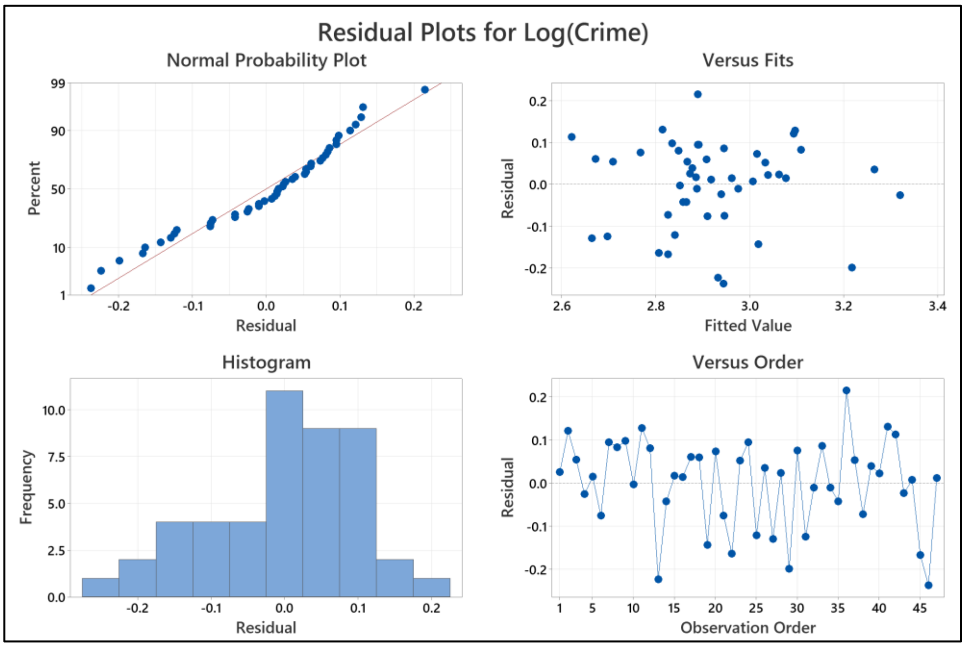
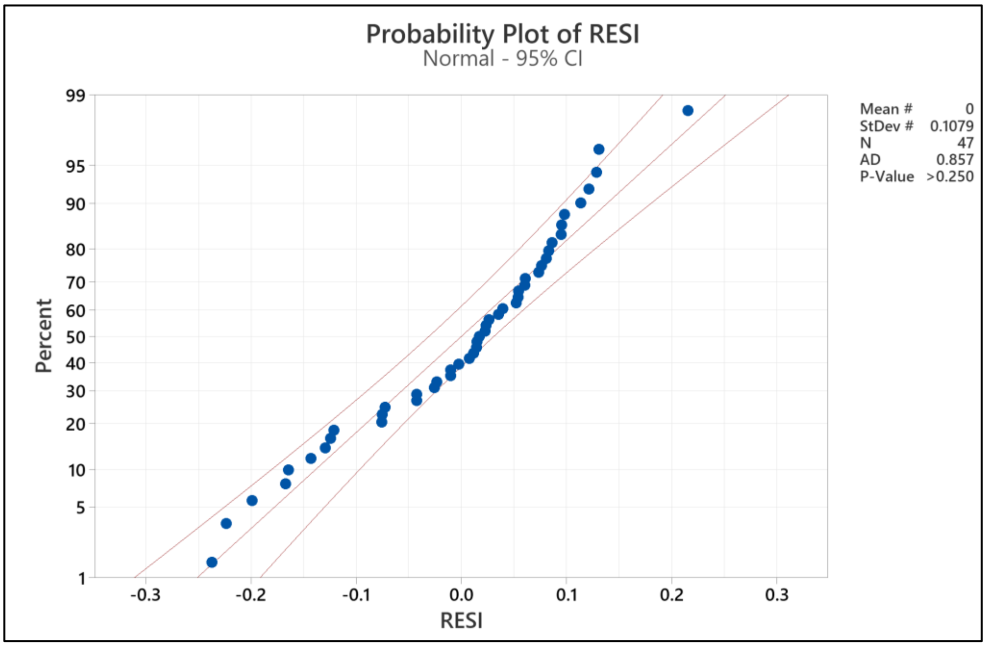


Figure 13: Probability Plots for Residuals



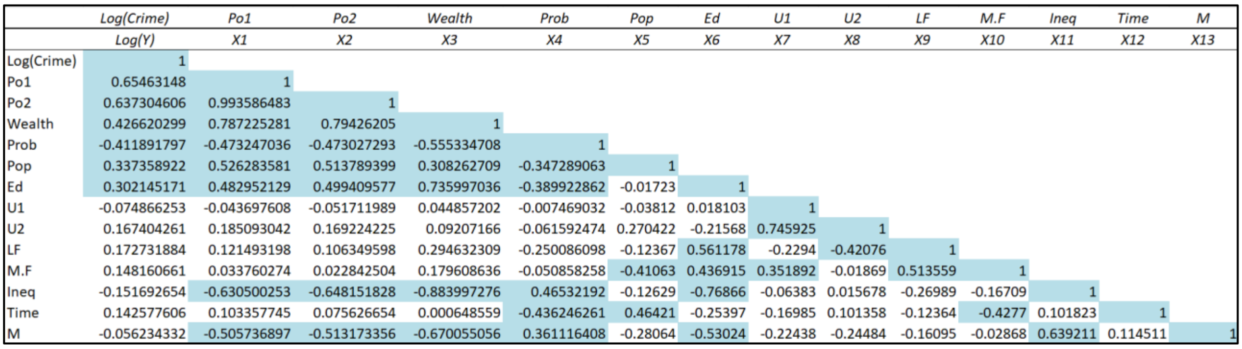
Inference:

The adjusted analysis of Model 2 has identified the need for improvements, particularly in improving the fit of the residuals. While the R-square and the adjusted R-squared values showed significant improvement after adding a new variable when compared to model 1, further modification is essential. Modifying the model through adjustments can lead to a more accurate and improved regression model with higher R-squared and adjusted R-squared values.

Adjusted Model Analysis (Model 3):

Model 2, with an R-squared value of 67.42%, represents an improvement, but a higher R2 value is expected. The next goal is to determine the optimal independent variables to consider. Upon reviewing the correlation table in Table 2, it is evident that M(X13) is highly correlated with the existing variables X1, X3, X4, X6, and X11.

Table 2: Correlation between Log(Crime) and X1-X13

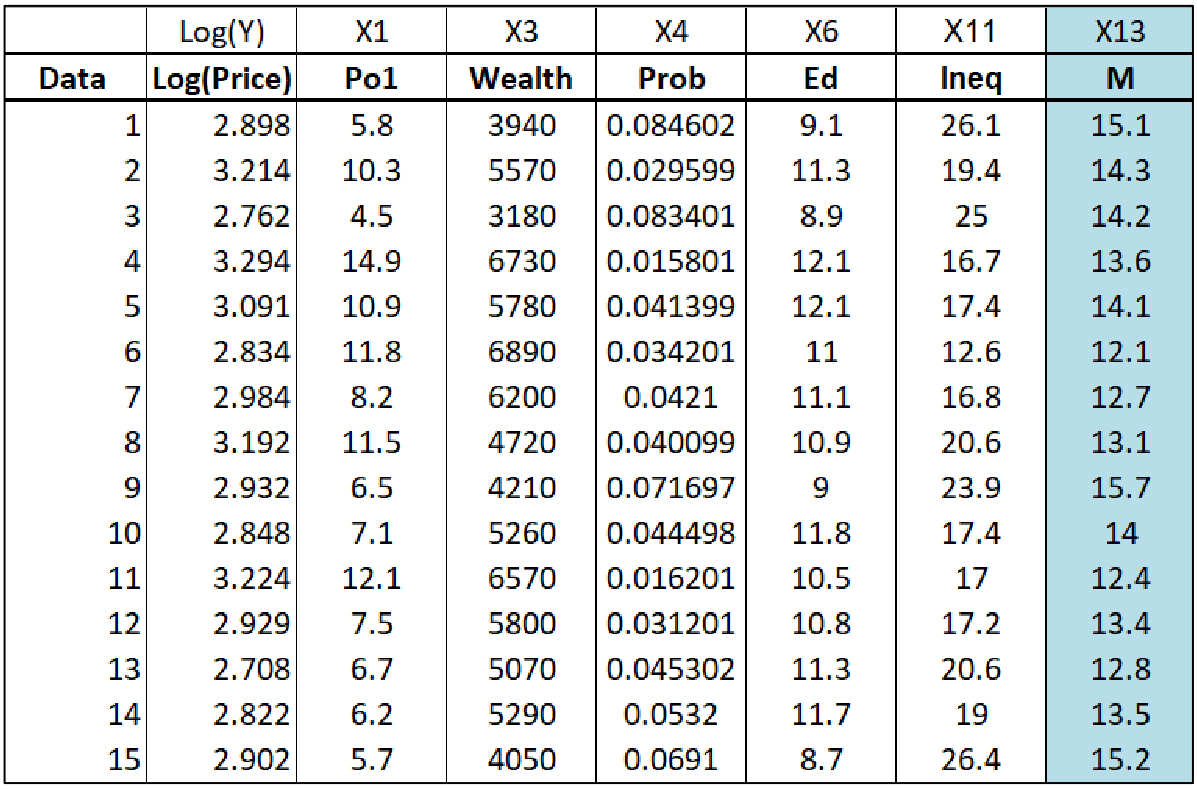


	Log(Crime)	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
	Log(Y)	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
Log(Crime)	1													
Po1	0.65463148	1												
Po2	0.637304606	0.993586483	1											
Wealth	0.426620299	0.787225281	0.79426205	1										
Prob	-0.411891797	-0.473247036	-0.473027293	-0.555334708	1									
Pop	0.337358922	0.526283581	0.513789399	0.308262709	-0.347289063	1								
Ed	0.302145171	0.482952129	0.499409577	0.735997036	-0.389922862	-0.01723	1							
U1	-0.074866253	-0.043697608	-0.051711989	0.044857202	-0.007469032	-0.03812	0.018103	1						
U2	0.167404261	0.185093042	0.169224225	0.09207166	-0.061592474	0.270422	-0.21568	0.745925	1					
LF	0.172731884	0.121493198	0.106349598	0.294632309	-0.250086098	-0.12367	0.561178	-0.2294	-0.42076	1				
M.F	0.148160661	0.033760274	0.022842504	0.179608636	-0.050858258	-0.41063	0.436915	0.351892	-0.01869	0.513559	1			
Ineq	-0.151692654	-0.630500253	-0.648151828	-0.883997276	0.46532192	-0.12629	-0.76866	-0.06383	0.015678	-0.26989	-0.16709	1		
Time	0.142577606	0.103357745	0.075626654	0.000648559	-0.436246261	0.46421	-0.25397	-0.16985	0.101358	-0.12364	-0.4277	0.101823	1	
M	-0.056234332	-0.505736897	-0.513173356	-0.670055056	0.361116408	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.639211	0.114511	1

Moreover, the percentage of males aged 14-24 is a demographic group often associated with higher rates of criminal activity. Including this variable helps to understand the transition from adolescence to adulthood.

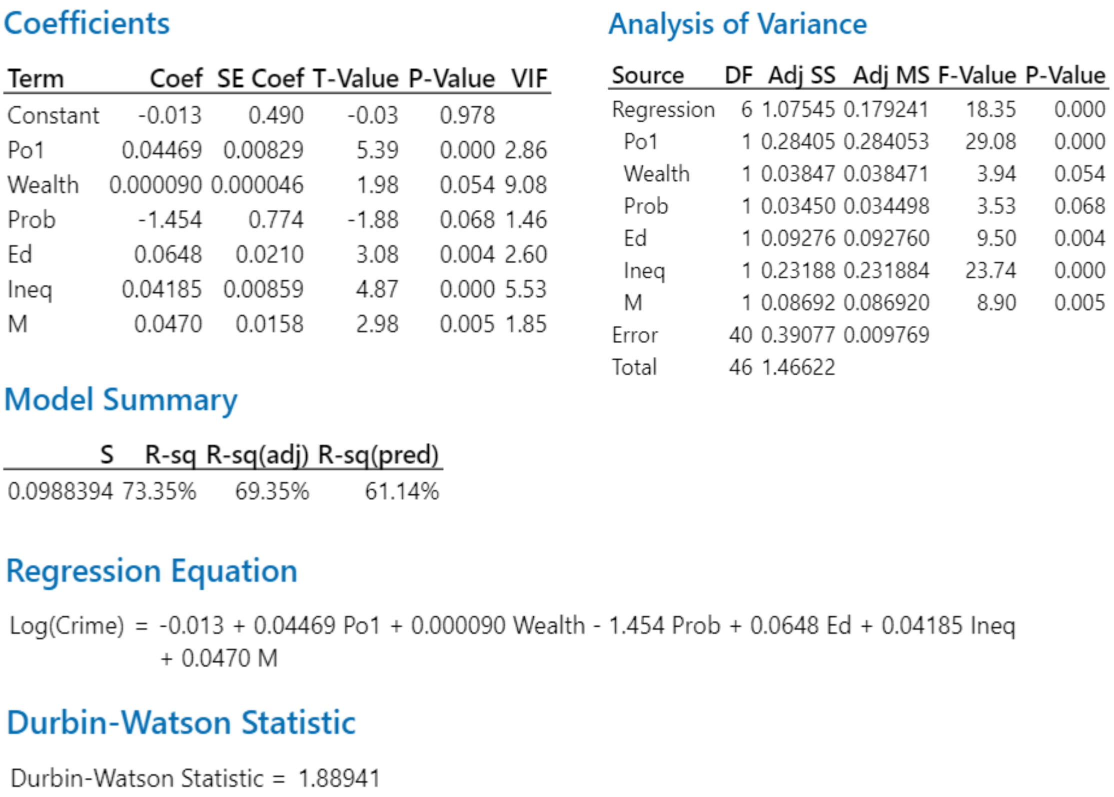
This modified model, tested with the inclusion of M(X13), is referred to as Model 3. The relevant data for the introduction of new independent variables is presented in Table 5.

Table 5: Adjusted Model Table for Model 3



	Log(Y)	X1	X3	X4	X6	X11	X13
Data	**Log(Price)**	**Po1**	**Wealth**	**Prob**	**Ed**	**Ineq**	**M**
1	2.898	5.8	3940	0.084602	9.1	26.1	15.1
2	3.214	10.3	5570	0.029599	11.3	19.4	14.3
3	2.762	4.5	3180	0.083401	8.9	25	14.2
4	3.294	14.9	6730	0.015801	12.1	16.7	13.6
5	3.091	10.9	5780	0.041399	12.1	17.4	14.1
6	2.834	11.8	6890	0.034201	11	12.6	12.1
7	2.984	8.2	6200	0.0421	11.1	16.8	12.7
8	3.192	11.5	4720	0.040099	10.9	20.6	13.1
9	2.932	6.5	4210	0.071697	9	23.9	15.7
10	2.848	7.1	5260	0.044498	11.8	17.4	14
11	3.224	12.1	6570	0.016201	10.5	17	12.4
12	2.929	7.5	5800	0.031201	10.8	17.2	13.4
13	2.708	6.7	5070	0.045302	11.3	20.6	12.8
14	2.822	6.2	5290	0.0532	11.7	19	13.5
15	2.902	5.7	4050	0.0691	8.7	26.4	15.2

Figure 14: Model 3 Regression Analysis for Log(Y) and X1, X3, X4, X6, X11 and X13



Coefficients

```
Term           Coef  SE Coef T-Value P-Value  VIF
Constant     -0.013    0.490   -0.03   0.978
Po1         0.04469  0.00829    5.39   0.000 2.86
Wealth     0.000090 0.000046    1.98   0.054 9.08
Prob         -1.454    0.774   -1.88   0.068 1.46
Ed           0.0648   0.0210    3.08   0.004 2.60
Ineq        0.04185  0.00859    4.87   0.000 5.53
M            0.0470   0.0158    2.98   0.005 1.85
```

Model Summary

```
        S   R-sq R-sq(adj) R-sq(pred)
0.0988394 73.35%    69.35%     61.14%
```

Analysis of Variance

```
Source      DF  Adj SS   Adj MS F-Value P-Value
Regression   6 1.07545 0.179241   18.35   0.000
  Po1        1 0.28405 0.284053   29.08   0.000
  Wealth     1 0.03847 0.038471    3.94   0.054
  Prob       1 0.03450 0.034498    3.53   0.068
  Ed         1 0.09276 0.092760    9.50   0.004
  Ineq       1 0.23188 0.231884   23.74   0.000
  M          1 0.08692 0.086920    8.90   0.005
Error       40 0.39077 0.009769
Total       46 1.46622
```

Regression Equation

```
Log(Crime) = -0.013 + 0.04469 Po1 + 0.000090 Wealth - 1.454 Prob + 0.0648 Ed + 0.04185 Ineq
             + 0.0470 M
```

Durbin-Watson Statistic

```
Durbin-Watson Statistic = 1.88941
```

Observations:

The observations made from Model 3 in Figure 14 are as follows:

- The VIF for Wealth and Ineq exceeds 5, indicating moderate collinearity between these two variables.
- Wealth and Prop have relatively higher P-values compared to other independent variables.
- The F-value of Model 3 is notably high, indicating a robust model.
- Model 3 has an R-squared of 73.53%, which is a significant improvement over the previous model. The adjusted R-Square value has also increased to 69.55%, which is a favorable outcome.
- Based on the Durbin-Watson statistic of 1.88941, it is concluded that the residuals are independent.
- Figure 15 presents a four-in-one plot for analyzing the behavior of residuals. The histogram shows a left skew, while the normal probability plot of data points shows a slight curve, suggesting that the trend line fit may not be perfect.
- The data points fit better to the trend line, although an outlier is identified within the probability plot for the residuals in Figure 16.

Figure 15: Residual Plots for Log(Y)

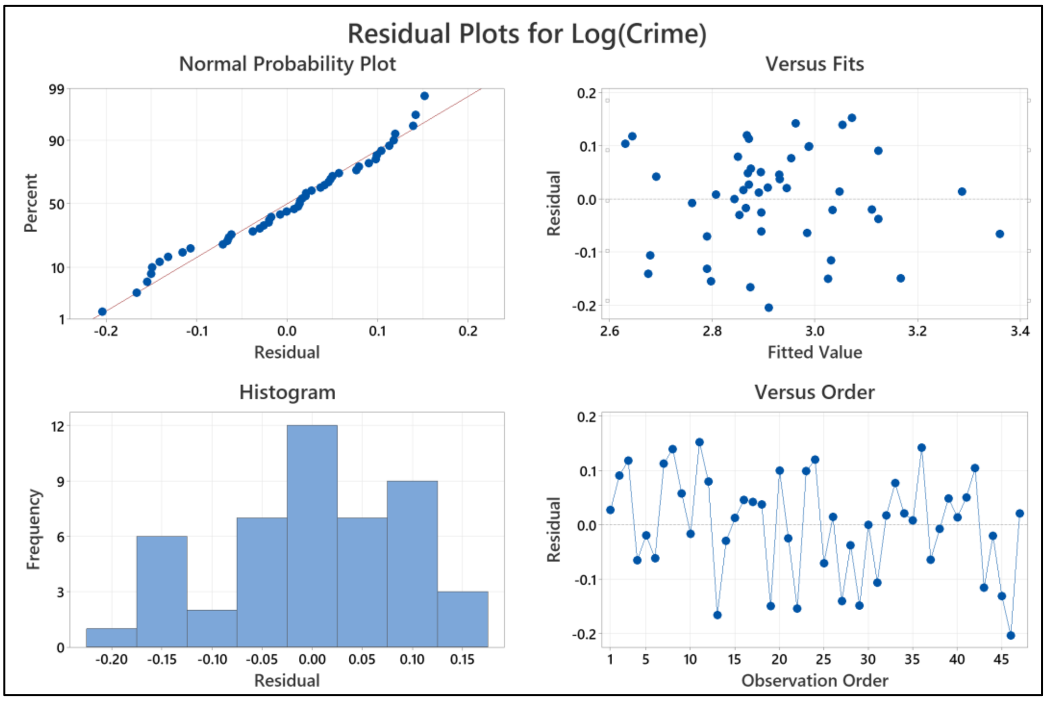
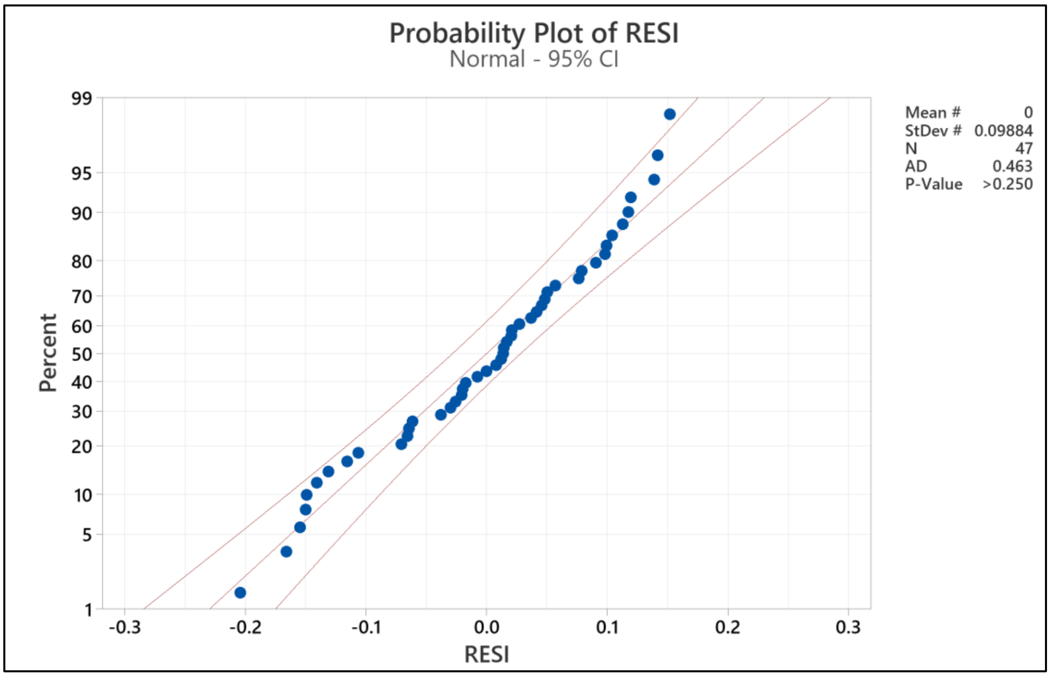


Figure 16: Probability Plots for Residuals



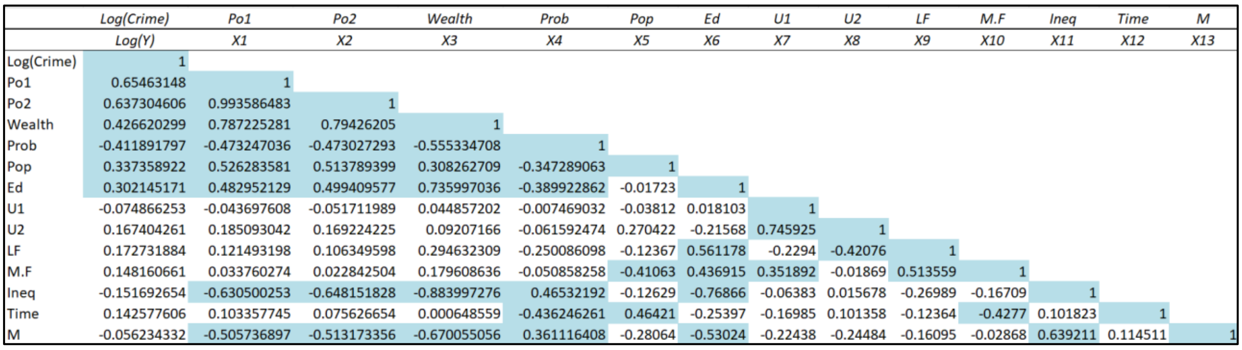
Inference:

The findings for Model 3 are encouraging, as it gives a higher R-squared value, a strong F-value, and a significant contribution from M(X13) towards the explanatory variable. Overall, the results of Model 3 suggest that the explanatory variables have a considerable impact on the dependent variable, but further investigation is needed to identify potential areas of improvement.

Adjusted Model Analysis (Model 4):

To further enhance the model, a thorough examination of the correlation table (Table 2) reveals a pronounced dependence between the independent variables Ineq (X11) and M (X13). The model experiences significant improvement by incorporating both Ineq (X11) and M (X13), which is not achieved by adding either of them individually.

Table 2: Correlation between Log(Crime) and X1-X13



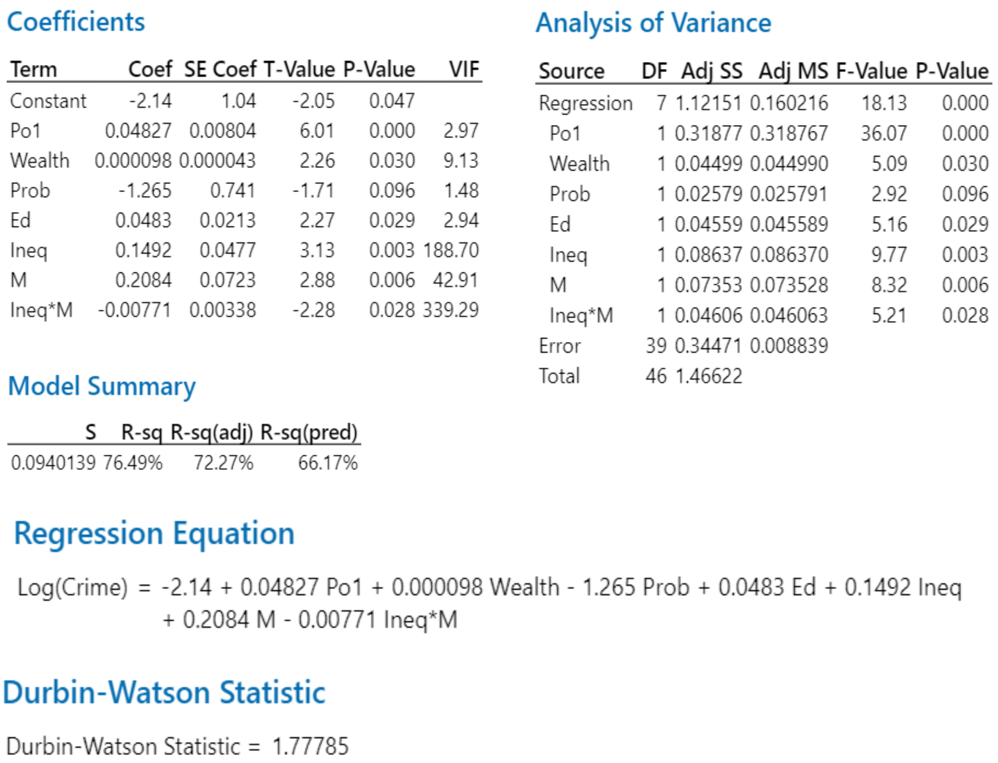
	Log(Crime)	*Po1*	*Po2*	*Wealth*	*Prob*	*Pop*	*Ed*	*U1*	*U2*	*LF*	*M.F*	*Ineq*	*Time*	*M*
	Log(Y)	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*X8*	*X9*	*X10*	*X11*	*X12*	*X13*
Log(Crime)	1													
Po1	0.65463148	1												
Po2	0.637304606	0.993586483	1											
Wealth	0.426620299	0.787225281	0.79426205	1										
Prob	-0.411891797	-0.473247036	-0.473027293	-0.555334708	1									
Pop	0.337358922	0.526283581	0.513789399	0.308262709	-0.347289063	1								
Ed	0.302145171	0.482952129	0.499409577	0.735997036	-0.389922862	-0.01723	1							
U1	-0.074866253	-0.043697608	-0.051711989	0.044857202	-0.007469032	-0.03812	0.018103	1						
U2	0.167404261	0.185093042	0.169224225	0.09207166	-0.061592474	0.270422	-0.21568	0.745925	1					
LF	0.172731884	0.121493198	0.106349598	0.294632309	-0.250086098	-0.12367	0.561178	-0.2294	-0.42076	1				
M.F	0.148160661	0.033760274	0.022842504	0.179608636	-0.050858258	-0.41063	0.436915	0.351892	-0.01869	0.513559	1			
Ineq	-0.151692654	-0.630500253	-0.648151828	-0.883997276	0.46532192	-0.12629	-0.76866	-0.06383	0.015678	-0.26989	-0.16709	1		
Time	0.142577606	0.103357745	0.075626654	0.000648559	-0.436246261	0.46421	-0.25397	-0.16985	0.101358	-0.12364	-0.4277	0.101823	1	
M	-0.056234332	-0.505736897	-0.513173356	-0.670055056	0.361116408	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.639211	0.114511	1

By adding an interactive term, Ineq(X11)*M(X13), is introduced as a new independent variable now referred to as Model 4. The data in Table 6 illustrates the addition of this new interactive term as one of the independent variables.

Table 6: Adjusted Model Table for Model 4

	Log(Y)	X_1	X_3	X_4	X_5	X11	X13	X14
Data	Log(Crime)	Po1	Wealth	Prob	Ed	Ineq	M	M*Ineq
1	2.898	5.8	3940	0.084602	9.1	26.1	15.1	394.11
2	3.214	10.3	5570	0.029599	11.3	19.4	14.3	277.42
3	2.762	4.5	3180	0.083401	8.9	25	14.2	355
4	3.294	14.9	6730	0.015801	12.1	16.7	13.6	227.12
5	3.091	10.9	5780	0.041399	12.1	17.4	14.1	245.34
6	2.834	11.8	6890	0.034201	11	12.6	12.1	152.46
7	2.984	8.2	6200	0.0421	11.1	16.8	12.7	213.36
8	3.192	11.5	4720	0.040099	10.9	20.6	13.1	269.86
9	2.932	6.5	4210	0.071697	9	23.9	15.7	375.23
10	2.848	7.1	5260	0.044498	11.8	17.4	14	243.6
11	3.224	12.1	6570	0.016201	10.5	17	12.4	210.8
12	2.929	7.5	5800	0.031201	10.8	17.2	13.4	230.48
13	2.708	6.7	5070	0.045302	11.3	20.6	12.8	263.68
14	2.822	6.2	5290	0.0532	11.7	19	13.5	256.5
15	2.902	5.7	4050	0.0691	8.7	26.4	15.2	401.28

Figure 17: Model 3 Regression Analysis for Log(Y) and X1, X3, X4, X6, X11, X13 and X11*X13



```
Coefficients

Term           Coef  SE Coef  T-Value  P-Value     VIF
Constant      -2.14     1.04    -2.05    0.047
Po1         0.04827  0.00804     6.01    0.000    2.97
Wealth     0.000098 0.000043     2.26    0.030    9.13
Prob         -1.265    0.741    -1.71    0.096    1.48
Ed           0.0483   0.0213     2.27    0.029    2.94
Ineq         0.1492   0.0477     3.13    0.003  188.70
M            0.2084   0.0723     2.88    0.006   42.91
Ineq*M     -0.00771  0.00338    -2.28    0.028  339.29
```

```
Analysis of Variance

Source      DF   Adj SS    Adj MS  F-Value  P-Value
Regression   7  1.12151  0.160216    18.13    0.000
  Po1        1  0.31877  0.318767    36.07    0.000
  Wealth     1  0.04499  0.044990     5.09    0.030
  Prob       1  0.02579  0.025791     2.92    0.096
  Ed         1  0.04559  0.045589     5.16    0.029
  Ineq       1  0.08637  0.086370     9.77    0.003
  M          1  0.07353  0.073528     8.32    0.006
  Ineq*M     1  0.04606  0.046063     5.21    0.028
Error       39  0.34471  0.008839
Total       46  1.46622
```

```
Model Summary

        S   R-sq  R-sq(adj)  R-sq(pred)
0.0940139 76.49%     72.27%      66.17%
```

Regression Equation

Log(Crime) = -2.14 + 0.04827 Po1 + 0.000098 Wealth - 1.265 Prob + 0.0483 Ed + 0.1492 Ineq
+ 0.2084 M - 0.00771 Ineq*M

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.77785

Observations:

Examining the results of the regression analysis for Model 4 in Figure 17 yields the following observations:

- The variables Ineq (X11), M (X13), and Ineq*M (X14) exhibit VIF values greater than 10, indicating potential issues with multicollinearity.
- Model 4 achieves an R-squared value of 76.49%, with an adjusted R-squared value of 72.27%.
- The F-value in Model 4 is the highest among the models, suggesting a robust overall fit.
- The P-value in Model 4 remains significantly low, indicating the model's statistical significance.
- The Durbin-Watson statistic is calculated as 1.77785, falling within the acceptable range of 1.5 to 2.5.
- Residuals versus fit plots in Figure 18 and residuals versus order plots in Model 4 closely resemble those of Model 3.
- No outliers are identified in the probability plot of residuals presented in Figure 19, contributing to the model's stability.

Figure 18: Residual Plots for Log(Y)

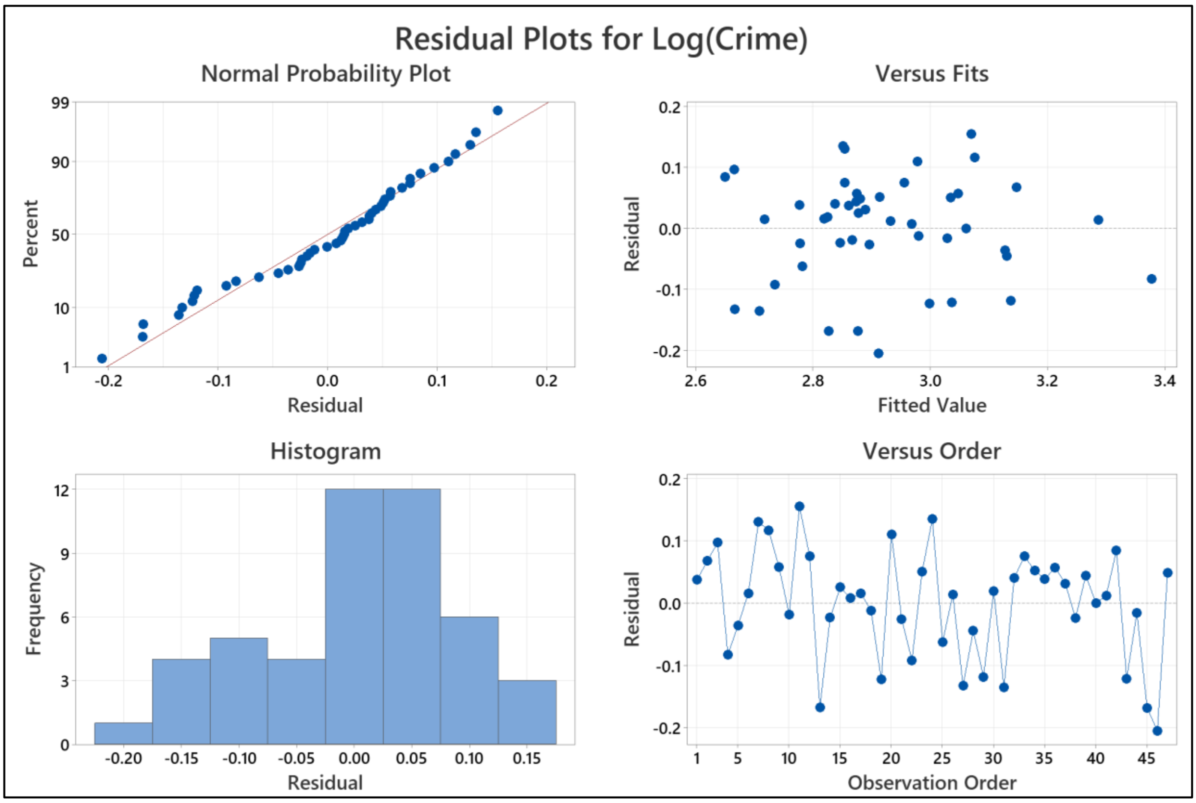
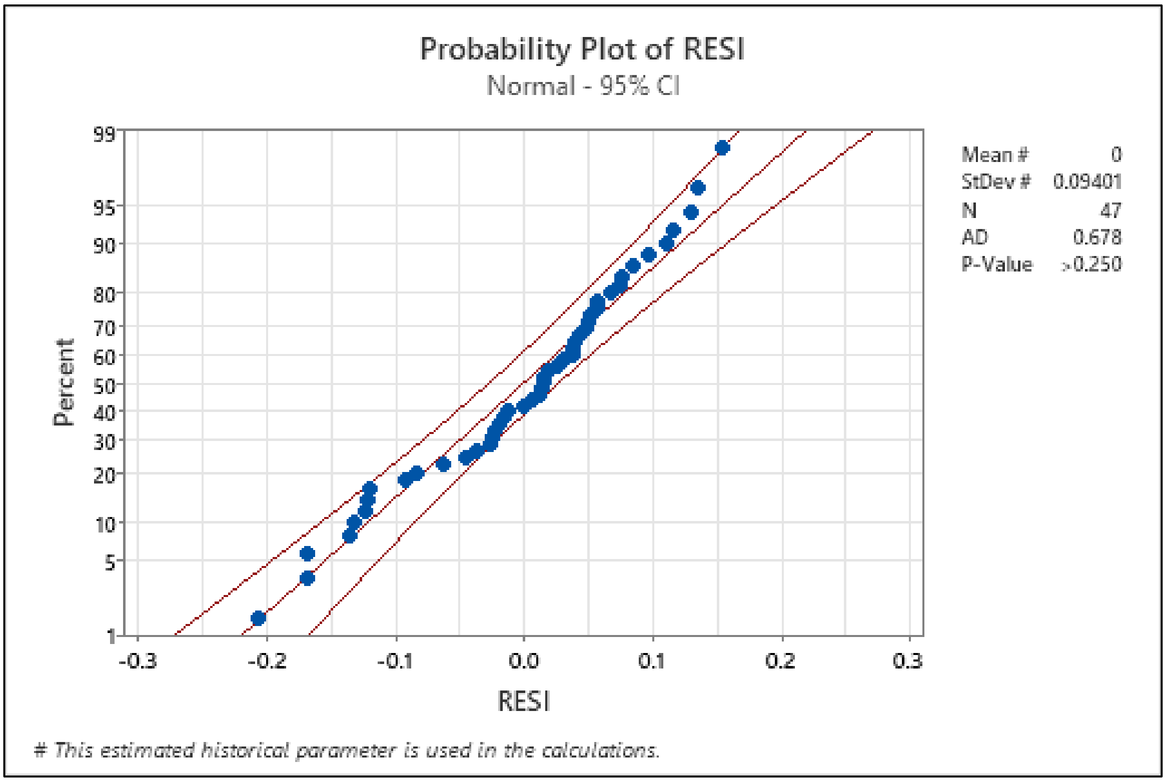


Figure 19: Probability Plots for Residuals



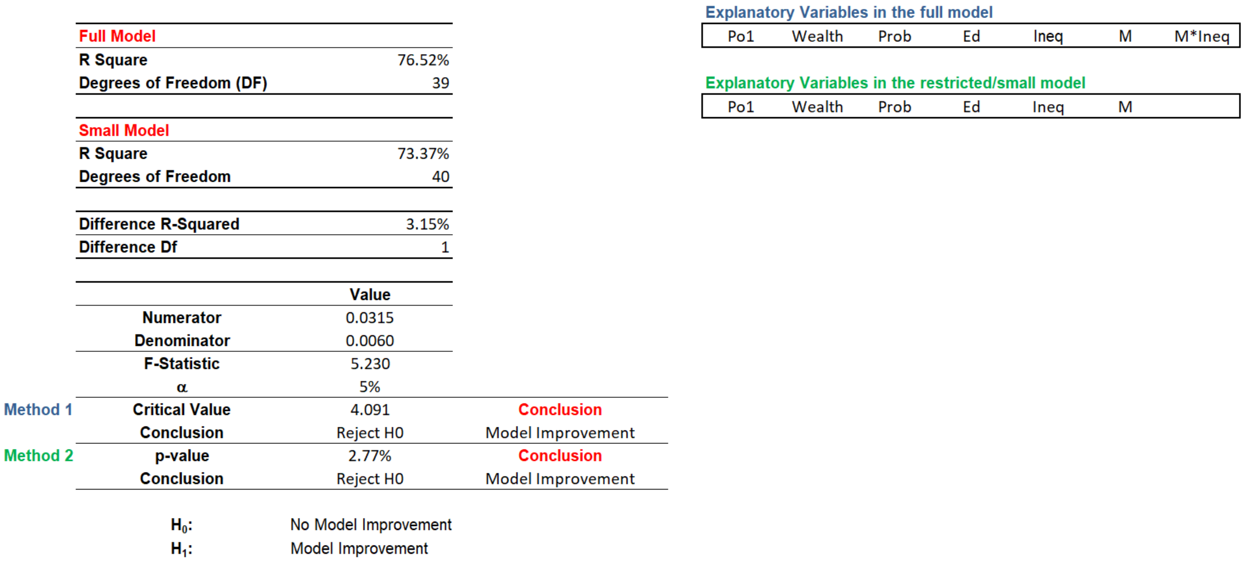
Inference:

Observations from Model 4 show a substantial improvement in explanatory power with high R-squared values. Although elevated VIF suggests challenges in assessing predictors accurately, it's unnecessary to remove variables solely based on this. Despite multicollinearity concerns, Model 4 outperforms Model 3, highlighted by a comparable R-squared, the highest F-value, and a consistently low P-value, indicating statistical significance. In summary, Model 4 effectively captures variable relationships, demonstrating robust performance.

Best Regression Fit:

An evaluation was performed to verify the improved fit of the adjusted model, as shown in Figure 20. This process involved comparing two models: the Full model and the small model. The small model incorporates all the explanatory variables that are a subset of variables present in the Full model, facilitating the execution of "the increase in R square test." The Full model achieved an R-square value of 76.52%, while the small model attained an R-square value of 73.37%. The difference between these two R-square values is 3.15%. Upon conducting the hypothesis test, the conclusions drawn from Method 1 and Method 2 indicated an improvement in the model. The results demonstrated that Model 4 outperformed the other models. Therefore, Model 4 is the best-fit model.

Figure 20: The Comparison of Full Model and Small Model



Full Model	
R Square	76.52%
Degrees of Freedom (DF)	39

Small Model	
R Square	73.37%
Degrees of Freedom	40

Difference R-Squared	3.15%
Difference Df	1

		Value	
	Numerator	0.0315	
	Denominator	0.0060	
	F-Statistic	5.230	
	α	5%	
Method 1	Critical Value	4.091	Conclusion
	Conclusion	Reject H0	Model Improvement
Method 2	p-value	2.77%	Conclusion
	Conclusion	Reject H0	Model Improvement

H_0: No Model Improvement

H_1: Model Improvement

Explanatory Variables in the full model

Po1	Wealth	Prob	Ed	Ineq	M	M*Ineq

Explanatory Variables in the restricted/small model

Po1	Wealth	Prob	Ed	Ineq	M

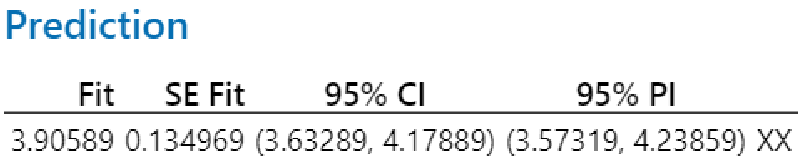
Prediction:

Forecasting for dependent variable for the given independent variables

Po1(X1) = 16, Po1(X2) = 15, Wealth(X3) = 6890, Prob(X4) = 0.01, Pop(X5) = 168, Ed(X6) = 12, U1(X7) = 0.14, U2(X8) = 5, LF(X9) = 0.6, M.F.(X10) = 107, Ineq(X11) = 27, Time(X12) = 44, M(X13) = 17, Ineq*M(X14) = 27 x 17 = 459

Firstly, all variables were substituted into the regression equation of Model 4, and the results are depicted in Figure 21.

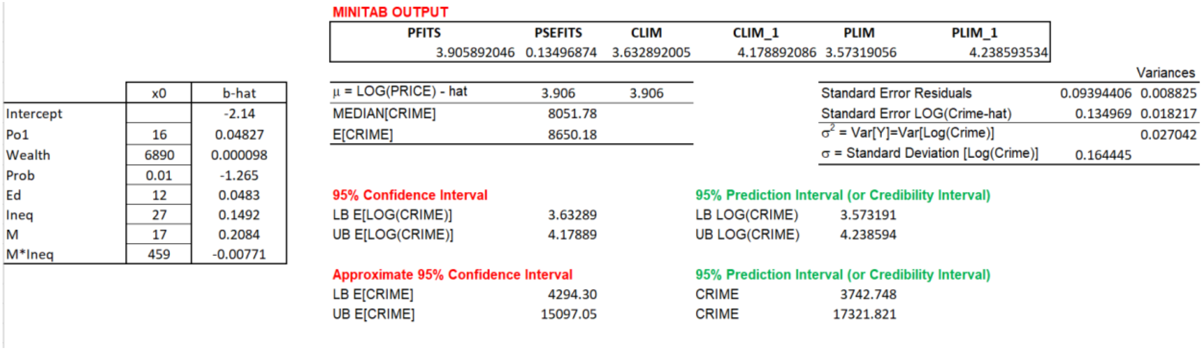
Figure 21: Prediction Minitab Output



Prediction

Fit	SE Fit	95% CI	95% PI	
3.90589	0.134969	(3.63289, 4.17889)	(3.57319, 4.23859)	XX

Figure 22: Prediction Excel output



	x0	b-hat
Intercept		-2.14
Po1	16	0.04827
Wealth	6890	0.000098
Prob	0.01	-1.265
Ed	12	0.0483
Ineq	27	0.1492
M	17	0.2084
M*Ineq	459	-0.00771

MINITAB OUTPUT

PFITS	PSEFITS	CLIM	CLIM_1	PLIM	PLIM_1
3.905892046	0.13496874	3.632892005	4.178892086	3.57319056	4.238593534

μ = LOG(PRICE) - hat	3.906	3.906
MEDIAN[CRIME]	8051.78	
E[CRIME]	8650.18	

		Variances
Standard Error Residuals	0.09394406	0.008825
Standard Error LOG(Crime-hat)	0.134969	0.018217
σ^2 = Var[Y]=Var[Log(Crime)]		0.027042
σ = Standard Deviation [Log(Crime)]	0.164445	

95% Confidence Interval	
LB E[LOG(CRIME)]	3.63289
UB E[LOG(CRIME)]	4.17889

95% Prediction Interval (or Credibility Interval)	
LB LOG(CRIME)	3.573191
UB LOG(CRIME)	4.238594

Approximate 95% Confidence Interval	
LB E[CRIME]	4294.30
UB E[CRIME]	15097.05

95% Prediction Interval (or Credibility Interval)	
CRIME	3742.748
CRIME	17321.821

Observations:

The log of the Crime Rate (3.906) falls within the 95% prediction interval of 3.63289 to 4.17889, as depicted in Figures 21 and 22.

Similarly, the Crime Rate value (8650.18) is within the 95% prediction interval, ranging from 4294.30 to 15097.05, as illustrated in Figure 22.

The prediction interval for the crime rate is substantial, as illustrated in Figure 22.

Conclusion:

The analysis provides a predictive model (Model 4) with a high R-squared value of 76.49%, indicating a high level of explained variance. The inclusion of seven variables, along with an interaction variable, reflects a careful consideration of factors influencing the dependent variable.

The model predicts a Crime Rate value (8650.18) within its corresponding 95% prediction interval (4294.30 to 15097.05), reinforcing the model's accuracy in capturing the variability inherent in the data. The broad range of this interval indicates the presence of diverse factors influencing crime rates, and the model successfully accommodates this complexity by providing a comprehensive prediction interval.

While the model demonstrates robust predictive capabilities, the implications of the limited sample size underscore the need for cautious interpretation and consideration of generalization. Future research may benefit from expanding the dataset to enhance the model's reliability and applicability, ultimately contributing to a more comprehensive understanding of the factors influencing the outcome of interest.

The analysis provides a comprehensive and reliable framework for predicting crime rates, offering statistical accuracy and practical insights.