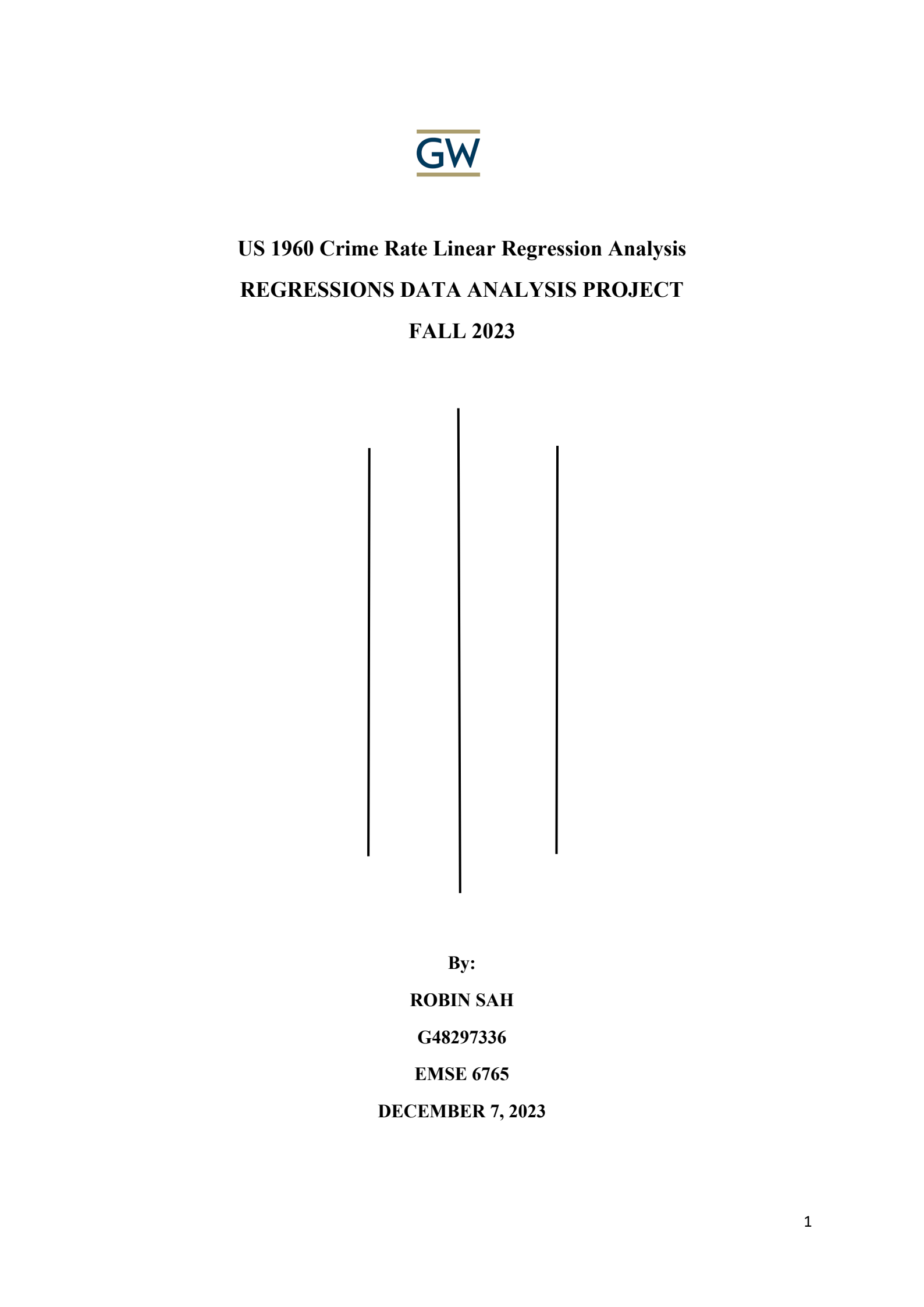


US 1960 Crime Rate Linear Regression Analysis

REGRESSIONS DATA ANALYSIS PROJECT

FALL 2023

By:

ROBIN SAH

G48297336

EMSE 6765

DECEMBER 7, 2023

Table of Contents:

1. Introduction

This project report aims to delve into the US 1960 Crime Dataset to build a linear regression model that investigates the relationship between various socio-economic indicators and the crime rate. By examining factors such as wealth conditions, education, unemployment rates, and population density, we can attempt to discern patterns and correlations that may explain the dynamics of crime during this decade.

2. Dataset

The dataset (Table 1) below presents a detailed portrait of crime and related socio-economic factors across 47 U.S. states in 1960. The primary focus is on the aggregate crime rate, denoted as Y, which quantifies offenses per 100,000 population, and its natural logarithm, Log(Y), to possibly linearize relationships and mitigate skewness in the data.

	Crime (Y)	Log(Crime)	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
2	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
3	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
4	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
5	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
6	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
7	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
8	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
9	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
10	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
11	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
12	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
13	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
14	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
15	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

Fig 1. US 1960 Crime Dataset snippet

Dataset Description

Dependent Variable

Crime (Y) - Crime rate: Number of offenses per 100,000 population in 1960

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: ratio of number of commitments to number of offenses
Pop - state population in 1960 in hundred thousand
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 - labour force participation rate of civilian urban male 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 release
M - percentage of males aged 14-24 in total state population

3. Motivation to use Log(Y)

The histograms and probability plots generated by Minitab provide visual and statistical insights into the suitability of these variables for linear regression analysis.

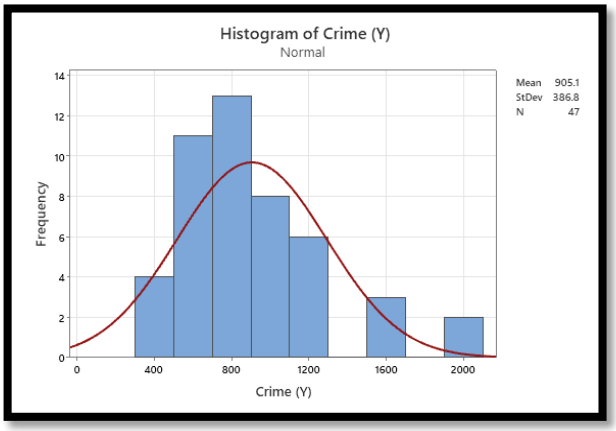


Fig 2. Histogram of Crime Rate (Y)

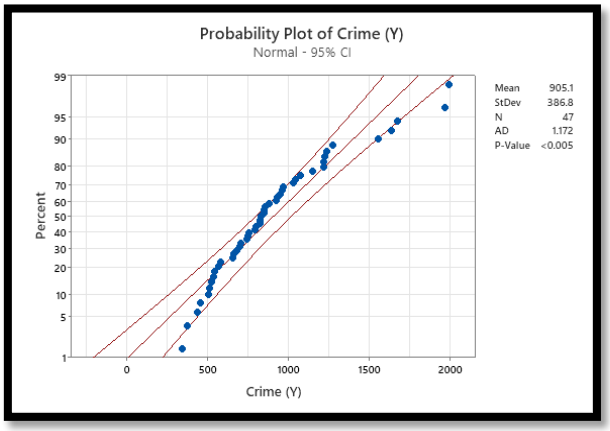


Fig 3. Probability Plot of Crime Rate (Y)

From the original Crime Rate (Y) data:

- The histogram shows a right-skewed distribution, suggesting that the distribution of the dependent variable (Y) is not normally distributed.
- The probability plot, which compares the distribution of Y to a normal distribution, indicates many data points deviating from the line, especially at the tails, highlighting non-normality. The significant standard deviation underlines the spread in the data, and the very low P-value suggests that the distribution of Y is not normal.

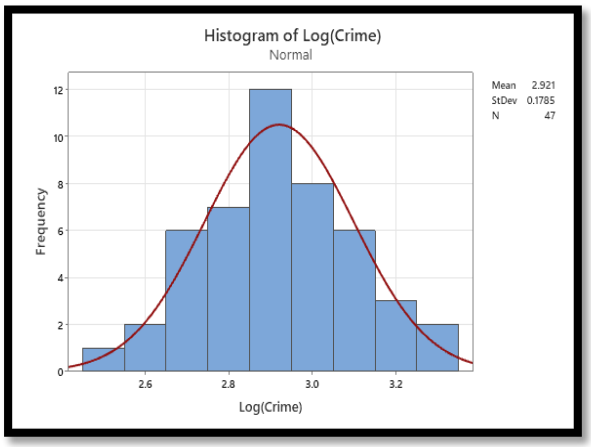


Fig 4. Probability Plot of Log Crime Rate (Log(Y))

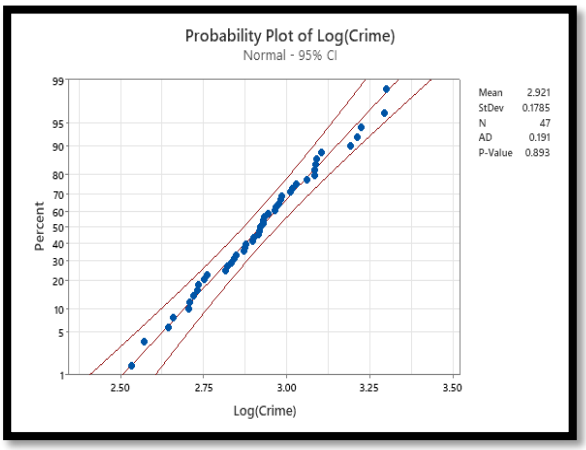


Fig 5. Histogram of Log Crime Rate (Log(Y))

When considering the logarithmic transformation (Log(Y)):

- The histogram of Log(Y) appears more symmetric and bell-shaped, resembling a normal distribution, indicating that the log transformation has helped normalize the data.
- The probability plot for Log(Y) shows data points closer to the trend line, though some deviation still exists. The standard deviation is smaller, indicating less variability in the

log-transformed data compared to the original Y values. The P-value remains low, similar to the original Y, but the improved alignment of data points to the trend line in the probability plot suggests enhanced normality.

Based on the above figures, we can justify the decision to use Log(Y) as the dependent variable in our regression model. The log transformation stabilizes the variance and improves the normality of the distribution, which is preferable for the assumptions underlying linear regression analysis. The use of Log(Y) allows for a clearer interpretation of percentage changes in the crime rate, as the regression coefficients can be understood in terms of percentage changes rather than absolute changes, which is particularly useful when dealing with skewed data.

4. Linear Regression Analysis

Correlation Analysis

The correlation matrix provided highlights the strength of linear relationships between the logarithm of the crime rate (Log (Crime)) and a series of independent variables within the dataset. A threshold of ±0.33 is used to identify substantial correlations, with cells exceeding this threshold highlighted in yellow.

	Log(Crime)	*Po1*	*Po2*	*Wealth*	*Prob*	*Pop*	*Ed*	*U1*	*U2*	*LF*	*M.F*	*Ineq*	*Time*	*M*
Log(Crime	1													
Po1	0.65463	1												
Po2	0.6373	0.99359	1											
Wealth	0.42662	0.78723	0.79426	1										
Prob	-0.41189	-0.47325	-0.47303	-0.55533	1									
Pop	0.33736	0.52628	0.51379	0.30826	-0.34729	1								
Ed	0.30215	0.48295	0.49941	0.736	-0.38992	-0.01723	1							
U1	-0.07487	-0.0437	-0.05171	0.04486	-0.00747	-0.03812	0.0181	1						
U2	0.1674	0.18509	0.16922	0.09207	-0.06159	0.27042	-0.21568	0.74592	1					
LF	0.17273	0.12149	0.10635	0.29463	-0.25009	-0.12367	0.56118	-0.2294	-0.42076	1				
M.F	0.14816	0.03376	0.02284	0.17961	-0.05086	-0.41063	0.43691	0.35189	-0.01869	0.51356	1			
Ineq	-0.15169	-0.6305	-0.64815	-0.884	0.46532	-0.12629	-0.76866	-0.06383	0.01568	-0.26989	-0.16709	1		
Time	0.14258	0.10336	0.07563	0.00065	-0.43625	0.46421	-0.25397	-0.16985	0.10136	-0.12364	-0.4277	0.10182	1	
M	-0.05623	-0.50574	-0.51317	-0.67006	0.36112	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.63921	0.11451	1

Fig 6. Correlation between Log (Crime) and other independent variables

From the matrix, the variables **Po1**, **Wealth**, and **Pop** show a strong and positive correlation with Log (Crime), surpassing the set threshold. This suggests that as these factors increase, there is a corresponding increase in the crime rate when looked at on a logarithmic scale. On the other hand, **Prob** exhibits a strong negative correlation, indicating that an increase in the probability of imprisonment is associated with a decrease in the crime rate.

Notably, **Po2** is also strongly correlated with Log (Crime). However, due to its high correlation with Po1, it is considered redundant for inclusion in a smaller model to avoid multicollinearity, which could distort the analysis. The selection of these variables for further analysis suggests they are key factors to consider in modelling the crime rate in the United States for the year 1960.

5. Model 0

Based on the threshold set on correlation matrix, we can use the independent variables Po1, Po2, Wealth, Prob, and Pop as the small model for our initial regression analysis. Po2 can be discarded since the correlation between Po1 and Po2 is too high which will only result in high VIF values for them suggesting multi-collinearity.

		discarded due to high correlation with Po1			
Log(Cri	Po1	Po2	Wealth	Prob	Pop
2.898	5.8	5.6	3940	0.084602	33
3.214	10.3	9.5	5570	0.029599	13
2.762	4.5	4.4	3180	0.083401	18
3.294	14.9	14.1	6730	0.015801	157
3.091	10.9	10.1	5780	0.041399	18
2.834	11.8	11.5	6890	0.034201	25
2.984	8.2	7.9	6200	0.0421	4
3.192	11.5	10.9	4720	0.040099	50
2.932	6.5	6.2	4210	0.071697	39
2.848	7.1	6.8	5260	0.044498	7

Fig 7. Table showing independent variables Po1, Wealth, Prob and Pop for Model 0

The observations from the Model 0 present the following result:

- Since, the variance inflation factors (VIF) values obtained from Minitab for the variables used below is less than 5, it guarantees that we don't have the issue of collinearity in this model.
- The R-Squared of 48.57% indicates a weaker model and only 48.57% of the variability in the crime rate is explained by the model.
- The Durbin-Watson statistic of 2.38897 suggests that there is no significant autocorrelation in the residuals of the model.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.93	0.171	17.15	0	
Po1	0.0533	0.0123	4.33	0	3.44
Wealth	-6.9E-05	0.000037	-1.88	0.067	3.19
Prob	-1.81	1.07	-1.69	0.098	1.52
Pop	-0.00045	0.000638	-0.7	0.486	1.51

The p-value of Pop is high, so we can remove this variable when building another model.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.133978	48.57%	43.68%	33.66%

Durbin-Watson Statistic

Durbin-Wa 2.38897

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regressior	4	0.7121	0.178024	9.92	0
Po1	1	0.33604	0.336037	18.72	0
Wealth	1	0.06333	0.06333	3.53	0.067
Prob	1	0.05135	0.051353	2.86	0.098
Pop	1	0.00888	0.00888	0.49	0.486
Error	42	0.75391	0.01795		
Total	46	1.46601			

Fig 8. Observations obtained from Minitab for Model 0

- The F-value of the regression is 9.92, with a P-value of 0 indicating that the model is statistically significant.
- The P-value of the first three variables is low which is preferred while the variable Pop has a high P-value. We should consider dropping this variable in our next model.

Regression Equation

Log(Crime) = 2.930 + 0.0533 Po1 - 0.000069 Wealth - 1.81 Prob - 0.000449 Pop

Fig 9. Regression equation (Model 0)

The residual analysis of the small model is more left-skewed than the normal distribution.

- The skewness is reflected in the normal probability plot, where the residuals do not perfectly align with the expected diagonal line. An upward curve on the plot further suggests the left-skewed nature of the residuals.
- The normal probability plot reveals deviations from normality, with the residuals not fitting neatly on the trend line.
- Within the probability plot, there are indications of two influential outliers.
- The plot of residuals versus fitted values suggests the presence of heteroscedasticity, meaning the variance of the residuals may not be constant across all levels of the independent variables.

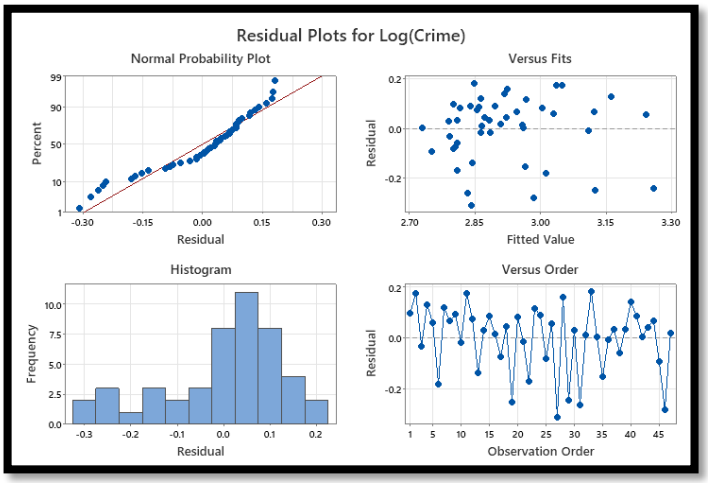


Fig 10. Residual Plots (Model 0)

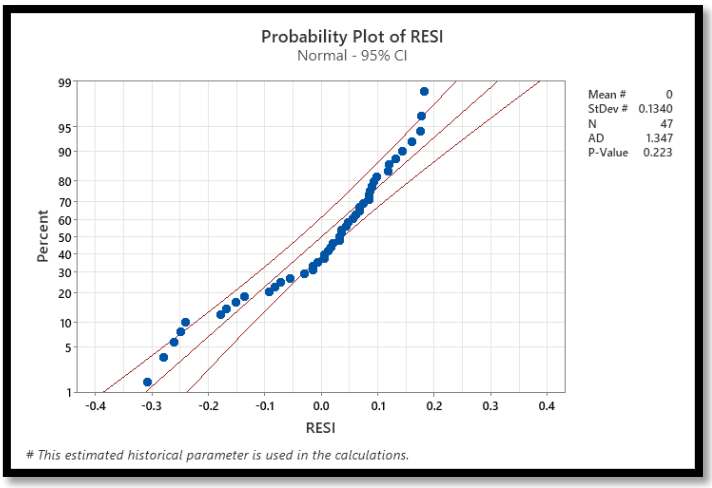


Fig 11. Probability Plot (Model 0)

6. Model 1

The four independent variables that we chose when the correlation threshold was set to 0.33 were Po1, Wealth, Prob, and Pop. We might want to use Ed for our next model since it exhibits a relatively strong correlation with the dependent variable when we set the threshold to 0.3.

The independent variable Pop was dropped due to high P-value and Ed was considered for building the next model. No significant changes were seen in the R-squared and adjusted R-squared values. The Durban-Watson increased to 2.52972. It appeared that the regression analysis equation with 49.12% R-Squared is insufficient to produce a reliable forecast result. Our next objective is to determine which combination of the independent variable works best by adding one or two more new independent variables. It turned out that the improvement comes from adding the variable M.F, Ineq and M. One of the reasons to add three independent variables can be justified from the correlation matrix where we can observe that these variables are highly correlated with most of the other variables.

Log(Cri	Po1	Wealth	Prob	Ed	M.F	Ineq	M
2.898	5.8	3940	0.084602	9.1	95	26.1	15.1
3.214	10.3	5570	0.029599	11.3	101.2	19.4	14.3
2.762	4.5	3180	0.083401	8.9	96.9	25	14.2
3.294	14.9	6730	0.015801	12.1	99.4	16.7	13.6
3.091	10.9	5780	0.041399	12.1	98.5	17.4	14.1
2.834	11.8	6890	0.034201	11	96.4	12.6	12.1
2.984	8.2	6200	0.0421	11.1	98.2	16.8	12.7
3.192	11.5	4720	0.040099	10.9	96.9	20.6	13.1
2.932	6.5	4210	0.071697	9	95.5	23.9	15.7
2.848	7.1	5260	0.044498	11.8	102.9	17.4	14
3.224	12.1	6570	0.016201	10.5	96.6	17	12.4
2.929	7.5	5800	0.031201	10.8	97.2	17.2	13.4

Fig 12. Table showing independent variables Po1, Wealth, Prob, Ed, M.F, Ineq and M for Model 1

The observations from the Model 1 present the following result:

- The VIF for Wealth and Ineq is highlighted, suggesting that these independent variables might be linearly related to other variables in the model. Specifically, the VIF for Wealth is greater than the commonly used threshold of 5, indicating strong multicollinearity concerns.
- The R-squared value of 73.40% shows that the model explains a significant portion of the variance in the dependent variable and the adjusted R-squared value of 68.63% adjusts for the number of predictors and is also relatively high, indicating a good fit.
- The Durbin-Watson statistic of 1.86434 suggests that there is no significant autocorrelation in the residuals.
- Given the high VIF values for Wealth and Ineq, one may consider investigating if combining these variables or excluding one could improve the model to reduce multicollinearity.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF	
Constant	-0.091	0.621	-0.15	0.884		
Po1	0.04488	0.00844	5.32	0	2.9	
Wealth	0.000089	0.000046	1.93	0.061	9.17	
Prob	-1.465	0.784	-1.87	0.069	1.46	
Ed	0.0624	0.0242	2.58	0.014	3.38	
M.F	0.00123	0.00595	0.21	0.838	1.41	The VIF of Wealth and Ineq is greater than 5,
Ineq	0.04151	0.00885	4.69	0	5.73	we may opt to choose their combination to
M	0.0465	0.0162	2.88	0.007	1.9	check if the model improves
				The p-value of **M.F** is significantly high, we may want to disregard this.		

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)			Durbin-Watson Statistic	
0.09999	73.40%	68.63%	58.81%			Durbin-Wa	1.86434

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regressior	7	1.07608	0.153726	15.38	0
Po1	1	0.28269	0.282694	28.27	0
Wealth	1	0.03728	0.037284	3.73	0.061
Prob	1	0.0349	0.034899	3.49	0.069
Ed	1	0.06633	0.066326	6.63	0.014
M.F	1	0.00043	0.000426	0.04	0.838
Ineq	1	0.22018	0.220184	22.02	0
M	1	0.08267	0.082666	8.27	0.007
Error	39	0.38993	0.009998		
Total	46	1.46601			

Fig 13. Observations obtained from Minitab for Model 1

Regression Equation

Log(Crime) = -0.091 + 0.04488 Po1 + 0.000089 Wealth - 1.465 Prob + 0.0624 Ed + 0.00123 M.F + 0.04151 Ineq + 0.0465 M

Fig 14. Regression equation (Model 1)

- The histogram of residual of this model is less left-skewed than model 0.
- The skewness is reflected in the normal probability plot, where the residuals do not perfectly align with the expected diagonal line.
- The normal probability plot reveals deviations from normality, with the residuals not fitting neatly on the trend line.
- There seems to improvement in the number of outliers and the P-value is 0.223, which implies that the residuals could be considered normally distributed.

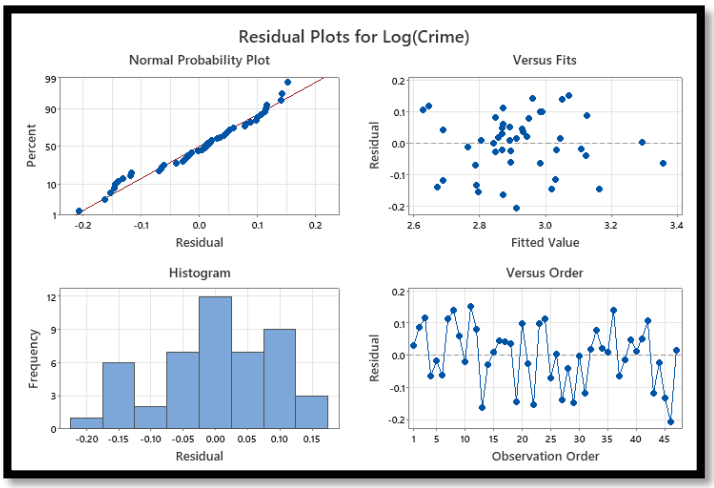


Fig 15. Residual Plots (Model 1)

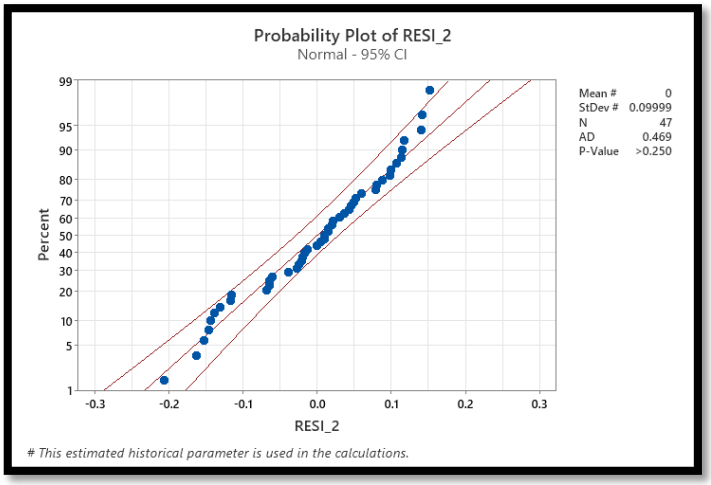


Fig 16. Probability Plot (Model 1)

7. Model 2

The combination of Wealth and Ineq was used to observe any significant improvement in the model but there was no significant increase in the R-square value which was 74.57%. We might consider dropping some independent variables since the use of many variables did not seem to be an ideal approach. We needed to be parsimonious in this case. Through repeated testing, we opted to drop variables M.F, Wealth*Ineq and observed a smaller model which gave a better adjusted R-squared values than previous model.

Log(Cri	Po1	Wealth	Prob	Ed	Ineq	M
2.898	5.8	3940	0.084602	9.1	26.1	15.1
3.214	10.3	5570	0.029599	11.3	19.4	14.3
2.762	4.5	3180	0.083401	8.9	25	14.2
3.294	14.9	6730	0.015801	12.1	16.7	13.6
3.091	10.9	5780	0.041399	12.1	17.4	14.1
2.834	11.8	6890	0.034201	11	12.6	12.1
2.984	8.2	6200	0.0421	11.1	16.8	12.7
3.192	11.5	4720	0.040099	10.9	20.6	13.1
2.932	6.5	4210	0.071697	9	23.9	15.7
2.848	7.1	5260	0.044498	11.8	17.4	14
3.224	12.1	6570	0.016201	10.5	17	12.4
2.929	7.5	5800	0.031201	10.8	17.2	13.4

Fig 17. Table showing independent variables Po1, Wealth, Prob, Ed, Ineq and M for Model 2

The observations from the Model 2 present the following result:

- The R-squared value is 73.37%, which shows a strong explanatory power of the model.
- The adjusted R-squared of 69.38% suggests that after adjusting for the number of predictors, the model still accounts for a substantial portion of the variance in the dependent variable.

The overall model is parsimonious compared to previous iterations, with a slight drop in R-squared value, but an improved adjusted R-squared value which is more representative of the model's predictive power when penalizing for the number of predictors.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.014	0.49	-0.03	0.977	
Po1	0.04468	0.00828	5.39	0	2.86
Wealth	0.00009	0.000045	1.99	0.054	9.08
Prob	-1.455	0.773	-1.88	0.067	1.46
Ed	0.0648	0.021	3.08	0.004	2.6
Ineq	0.04186	0.00859	4.88	0	5.53
M	0.0471	0.0158	2.99	0.005	1.85

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	Durbin-Watson Statistic
0.098787	73.37%	69.38%	61.17%	Durbin-Wa 1.88882

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regressior	6	1.07565	0.179276	18.37	0
Po1	1	0.28396	0.283961	29.1	0
Wealth	1	0.0385	0.038501	3.95	0.054
Prob	1	0.03457	0.034569	3.54	0.067
Ed	1	0.0928	0.092801	9.51	0.004
Ineq	1	0.23197	0.231968	23.77	0
M	1	0.08702	0.087021	8.92	0.005
Error	40	0.39035	0.009759		
Total	46	1.46601			

Fig 18. Observations obtained from Minitab for Model 2

Regression Equation

Log(Crime) = -0.014 + 0.04468 Po1 + 0.000090 Wealth - 1.455 Prob + 0.0648 Ed + 0.04186 Ineq + 0.0471 M

Fig 19. Regression equation (Model 2)

- From residuals' probability plot, we can find only 1 influential outlier.
- The data point is more fit to the trend line and distributes a smaller curve.
- Residuals versus order plot in this model is less chaotic than previous models. Since the distribution of residuals does not show severe deviations from normality. This

indicates that the model may be well-specified and the residuals behave as expected under the assumptions of linear regression.

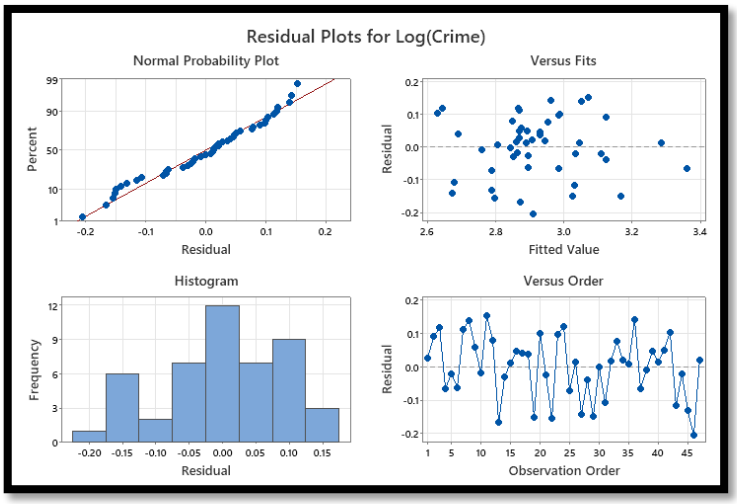


Fig 20. Residual Plots (Model 2)

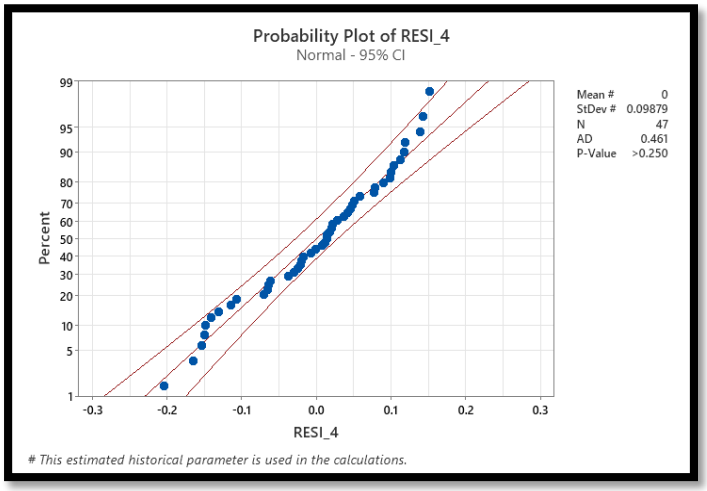


Fig 21. Probability Plot (Model 2)

8. Model 3

In the process of model refinement, an iterative approach was taken to enhance the explanatory power of the linear regression model. This involved both the addition of new variables and the exploration of interaction effects between existing variables. The incorporation of 'Ineq', representing income inequality, and 'M', the percentage of young males in the population, individually contributed to the model, yet only marginally increased the adjusted R-squared value. This modest improvement suggested that while each variable individually holds some explanatory power, there might be an amplified effect when considering their interaction.

Given the observed high correlation between different variables, the variables 'Po1' and 'Ineq', indicative of potential interplay between these factors, an interaction term was created by multiplying them. Our new table observed is displayed in figure.

Log(Cri	Po1	Wealth	Prob	Ed	Ineq	M	Po1*Ineq
2.898	5.8	3940	0.084602	9.1	26.1	15.1	151.38
3.214	10.3	5570	0.029599	11.3	19.4	14.3	199.82
2.762	4.5	3180	0.083401	8.9	25	14.2	112.5
3.294	14.9	6730	0.015801	12.1	16.7	13.6	248.83
3.091	10.9	5780	0.041399	12.1	17.4	14.1	189.66
2.834	11.8	6890	0.034201	11	12.6	12.1	148.68
2.984	8.2	6200	0.0421	11.1	16.8	12.7	137.76
3.192	11.5	4720	0.040099	10.9	20.6	13.1	236.9
2.932	6.5	4210	0.071697	9	23.9	15.7	155.35
2.848	7.1	5260	0.044498	11.8	17.4	14	123.54
3.224	12.1	6570	0.016201	10.5	17	12.4	205.7
2.929	7.5	5800	0.031201	10.8	17.2	13.4	129

*Fig 22. Table showing independent variables Po1, Wealth, Prob, Ed, Ineq, M and Po1*Ineq for Model 3*

The observations from the Model 3 present the following result:

- The interaction term 'Po1*Ineq' has a positive coefficient and is statistically significant with a P-value of 0.01, indicating that the combined effect of police expenditure and income inequality has a notable impact on the dependent variable.
- The R-squared value has increased to 77.54%, suggesting that a larger proportion of the variability in the crime rate is explained by this model compared to previous ones.
- The adjusted R-squared value is 73.51%, showing improvement and a good fit of the model after accounting for the number of predictors used.
- The Durbin-Watson statistic of 2.0732 indicates that there is no serious autocorrelation issue among the residuals.

Model 3, with the inclusion of the interaction term 'Po1*Ineq', has demonstrated an improved fit with a higher R-squared and adjusted R-squared. The interaction term's significance suggests that it captures a combined effect that is not represented when 'Po1' and 'Ineq' are considered solely as main effects. However, the high VIF values for some predictors indicate multicollinearity, which may be distorting the estimated relationships and should be addressed.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.867	0.561	1.54	0.131	
Po1	-0.0311	0.0292	-1.06	0.294	41.04
Wealth	0.000074	0.000043	1.74	0.089	9.25
Prob	-1.152	0.728	-1.58	0.122	1.49
Ed	0.0496	0.0203	2.44	0.019	2.82
Ineq	0.0047	0.016	0.3	0.769	22.07
M	0.0458	0.0147	3.12	0.003	1.85
Po1*Ineq	0.00464	0.00173	2.69	0.01	26.14

Model Summary **Durbin-Watson Statistic**

S	R-sq	R-sq(adj)	R-sq(pred)	Durbin-Wa	2.0732
0.09189	77.54%	73.51%	66.92%		

Increase in adjusted R-squared than previous model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regressior	7	1.1367	0.162385	19.23	0
Po1	1	0.00955	0.009551	1.13	0.294
Wealth	1	0.02561	0.025614	3.03	0.089
Prob	1	0.02114	0.021138	2.5	0.122
Ed	1	0.05024	0.050241	5.95	0.019
Ineq	1	0.00074	0.000741	0.09	0.769
M	1	0.08224	0.082241	9.74	0.003
Po1*Ineq	1	0.06104	0.061044	7.23	0.01

Fig 23. Observations obtained from Minitab for Model 3

Regression Equation	
Log(Crime) =	0.867 - 0.0311 Po1 + 0.000074 Wealth - 1.152 Prob + 0.0496 Ed + 0.0047 Ineq
+ 0.0458 M + 0.00464 Po1*Ineq	

Fig 24. Regression equation (Model 3)

- The residuals mostly fall around the line in this graph, it indicates that the residuals are approximately normally distributed.
- No influential observation (outlier) is identified from the probability plot of residuals.
- The absence of patterns in the residuals-versus-fits and residuals-versus-order plots further indicate that the model's assumptions of homoscedasticity and independence of errors are being met.

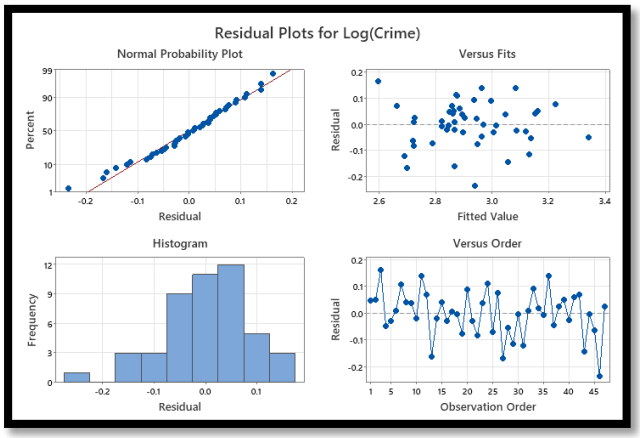


Fig 25. Residual Plots (Model 3)

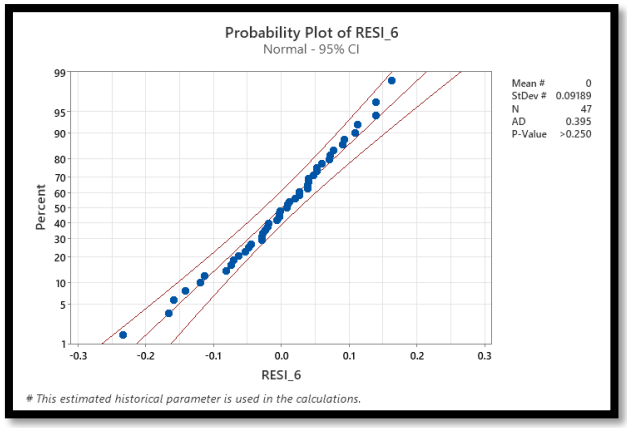
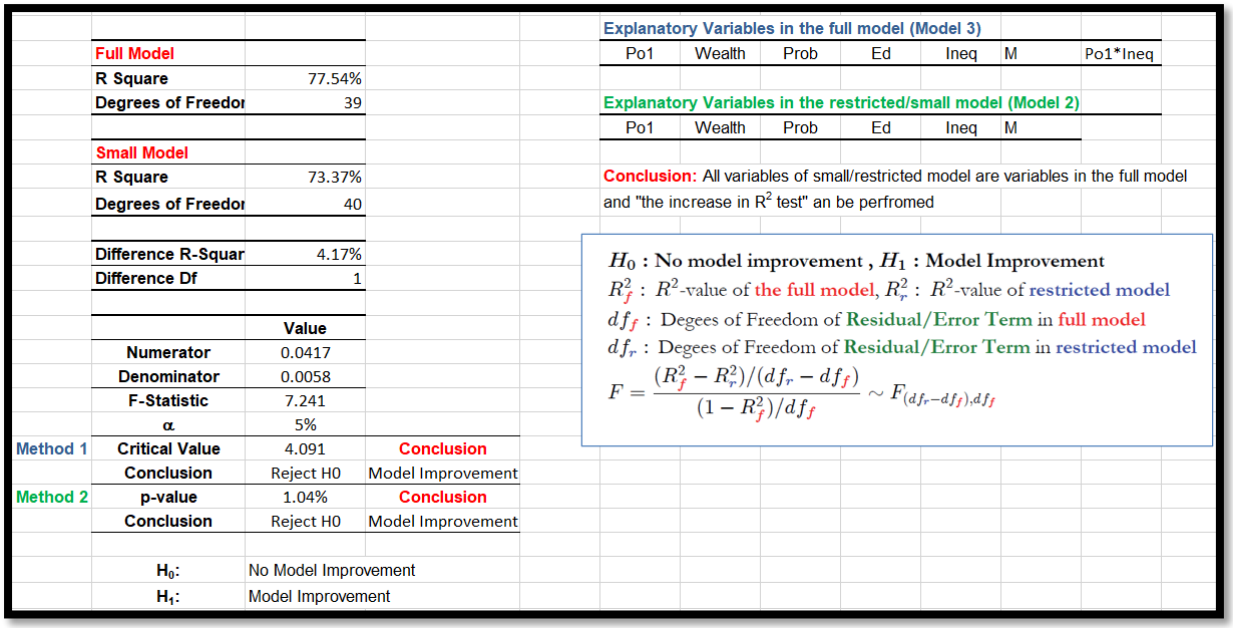


Fig 26. Probability Plot (Model 3)

9. Comparing the models

Based on various criteria, the best model would be Model 3 with a high adjusted R-squared value, statistically significant predictors, residuals that closely follow a normal distribution with no apparent patterns or outliers, and a Durbin-Watson statistic close to 2.



	Full Model		
	R Square	77.54%	
	Degrees of Freedor	39	
	Small Model		
	R Square	73.37%	
	Degrees of Freedor	40	
	Difference R-Squar	4.17%	
	Difference Df	1	
		Value	
	Numerator	0.0417	
	Denominator	0.0058	
	F-Statistic	7.241	
	α	5%	
Method 1	Critical Value	4.091	Conclusion
	Conclusion	Reject H0	Model Improvement
Method 2	p-value	1.04%	Conclusion
	Conclusion	Reject H0	Model Improvement
	H_0:	No Model Improvement	
	H_1:	Model Improvement	

Explanatory Variables in the full model (Model 3)

Po1	Wealth	Prob	Ed	Ineq	M	Po1*Ineq

Explanatory Variables in the restricted/small model (Model 2)

Po1	Wealth	Prob	Ed	Ineq	M

Conclusion: All variables of small/restricted model are variables in the full model and "the increase in R^2 test" an be perfromed

H_0 : No model improvement , H_1 : Model Improvement

R_f^2 : R^2-value of the full model, R_r^2 : R^2-value of restricted model

df_f : Degees of Freedom of Residual/Error Term in full model

df_r : Degees of Freedom of Residual/Error Term in restricted model

$$F = \frac{(R_f^2 - R_r^2)/(df_r - df_f)}{(1 - R_f^2)/df_f} \sim F_{(df_r - df_f), df_f}$$

Fig 27. Comparison between model 3 and model 2

- The null hypothesis H0 posits that the full model does not provide a significant improvement over the small model, while the alternative hypothesis H1 suggests there is an improvement.
- In this case, the F-statistic value is 7.241, indicating the test statistic value from the F-distribution.
- The critical value at the 5% significance level is 4.091, and since the calculated F-statistic value is greater than the critical value, we would reject H0 at this level.
- The p-value for the test is 1.04%, which is below the conventional alpha level of 5%, further leading to the rejection of H1.

The statistical evidence suggests that the full model with the interaction term (Model 3) provides a significant improvement in the explanatory power compared to the small model (Model 2). This is demonstrated by the higher R-squared value, the significant F-statistic, and the low p-value. The conclusion is that including the interaction between Po1 and Ineq leads to a better-fitted model for predicting the logarithm of the crime.

10. Prediction

Forecasting Dependent Variable Value

Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
16	15	6890	0.01	168	12	0.14	5	0.6	107	27	44	17

Regression Equation

Log(Crime) = 0.867 - 0.0311 Po1 + 0.000074 Wealth - 1.152 Prob + 0.0496 Ed + 0.0047 Ineq + 0.0458 M + 0.00464 Po1*Ineq

Settings

Variable	Setting
Po1	16
Wealth	6890
Prob	0.01
Ed	12
Ineq	27
M	17

- The 95% confidence interval for the prediction is (4.00845, 4.74388). This interval estimates the range within which the true value of the dependent variable is expected to fall, with 95% confidence, given the model and the specific values of the independent variables.
- The 95% prediction interval is (3.96414, 4.78818). This interval provides a range for where a new observation's dependent variable value would likely fall, with 95% confidence, given the model and the set values of the independent variables.

Prediction

Fit	SE Fit	95% CI	95% PI	
4.37616	0.181795	(4.00845, 4.74388)	(3.96414, 4.78818)	XX

XX denotes an extremely unusual point relative to predictor levels used to fit the model.

Fig 28. Results of Prediction

MINITAB OUTPUT							
PFITS	PSEFITS	CLIM	CLIM_1	PLIM	PLIM_1		
4.376162285	0.181795385	4.008446409	4.74387816	3.964141883	4.788183		
							Variances
μ = LOG(crime) -	4.376	4.373		Standard Error Residuals		0.146820	0.021556
MEDIAN[crime]	2377.728617			Standard Error LOG(CRI		0.181795	0.03305
E[crime]	2748.0806			s2 = Var[Y]=Var[Log(CRIME)]			0.054606
				s = Standard Deviation [L		0.233679	
95% Confidence Interval			95% Prediction Interval (or Credibility Interval)				
LB E[LOG(CRIMI	4.00845		LB LOG(CRIME)	3.964142			
UB E[LOG(CRIM	4.74388		UB LOG(CRIME)	4.788183			
Approximate 95% Confidence Interval			95% Prediction Interval (or Credibility Interval)				
LB E[CRIME]	1019.638932		CRIME	920.7503295			
UB E[CRIME]	5544.701362		CRIME	6140.202392			
Approximate because Log is not a linear function							

Log Crime Rate is within the 95% prediction interval.

Crime Rate is within the 95% prediction interval.

11. Conclusion

The final regression model developed in this project offers a nuanced estimate of the crime rate in the United States for the year 1960, utilizing a carefully selected set of independent variables. This model accounts for the complexity of social factors influencing crime rates, such as police expenditure, median family income, education levels, income inequality, and the proportion of young males in the population.

This project demonstrates a balance between model complexity and predictive accuracy, leading to a more parsimonious and interpretable model. This model is preferred not only for its statistical robustness, as evidenced by a better fit reflected in the adjusted R-squared value, but also for providing a more cost-effective approach to understanding and potentially addressing the factors that contribute to crime.

Model 3 serves as a strong analytical tool, offering valuable insights into the relationship between socio-economic indicators and crime rates. While recognizing the model's limitations, its simplicity and accuracy make it a useful resource for policymakers and social scientists aiming to interpret historical crime data and inform future decisions.