Influence of Punishment regimes and Socio-Economic factors on Crime Rates

Crime Rate Analysis

Linear Regression

By:

Karthik Sivaraman Iyer

EMSE 6765

December 07, 2023

1.0) Introduction

Criminologists are working towards understanding the effect of punishment regimes on the crime rates. In doing so, the aggregated data of key factors from the 47 states of the USA from 1960 is studied. This report walks through the estimation of the regression model and dives into the variable selection, diagnostic analysis and evaluation of the model, ultimately leading to the model's evolution. The report aims to ultimately define an estimated best Linear Regression model to forecast the Crime Rate for a state outside of the Dataset.

2.0) Data Description

The dataset used for the forecasting of Crime Rates is from 1960. It contains 13 independent variable attributes denoted with 'X1' through 'X13', and one dependent variable Crime denoted with 'Y'. The dataset has 47 datapoints denoting the aggregated values of each variable for 47 states. The 'Table 2.1' below is representative of the dataset available for the development of the model.

Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
Crime	**Po1**	**Po2**	**Wealth**	**Prob**	**Pop**	**Ed**	**U1**	**U2**	**LF**	**M.F**	**Ineq**	**Time**	**M**
791	5.8	5.6	3940	0.084602	33	9.1	0.108	4.1	0.51	95	26.1	26.2011	15.1
1635	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	4.5	4.4	3180	0.083401	18	8.9	0.094	3.3	0.533	96.9	25	24.3006	14.2
1969	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	10.9	10.1	5780	0.041399	18	12.1	0.091	2	0.591	98.5	17.4	21.2998	14.1
682	11.8	11.5	6890	0.034201	25	11	0.084	2.9	0.547	96.4	12.6	20.9995	12.1
963	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	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	6.5	6.2	4210	0.071697	39	9	0.081	2.8	0.553	95.5	23.9	29.4001	15.7
705	7.1	6.8	5260	0.044498	7	11.8	0.1	2.4	0.632	102.9	17.4	19.5994	14

Table 2.1: US 1960 Crime Data

The variable available in the Dataset are as follows:

Dependent Variable:

Crime (Y): 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 transferable 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: Labor 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

3.0 Preliminary Data Analysis

We perform an initial analysis on the dependent variable Crime (Y) to understand the characteristics of the data. In doing so, we plot the histogram and probability plots for Y to understand the distribution and normality.

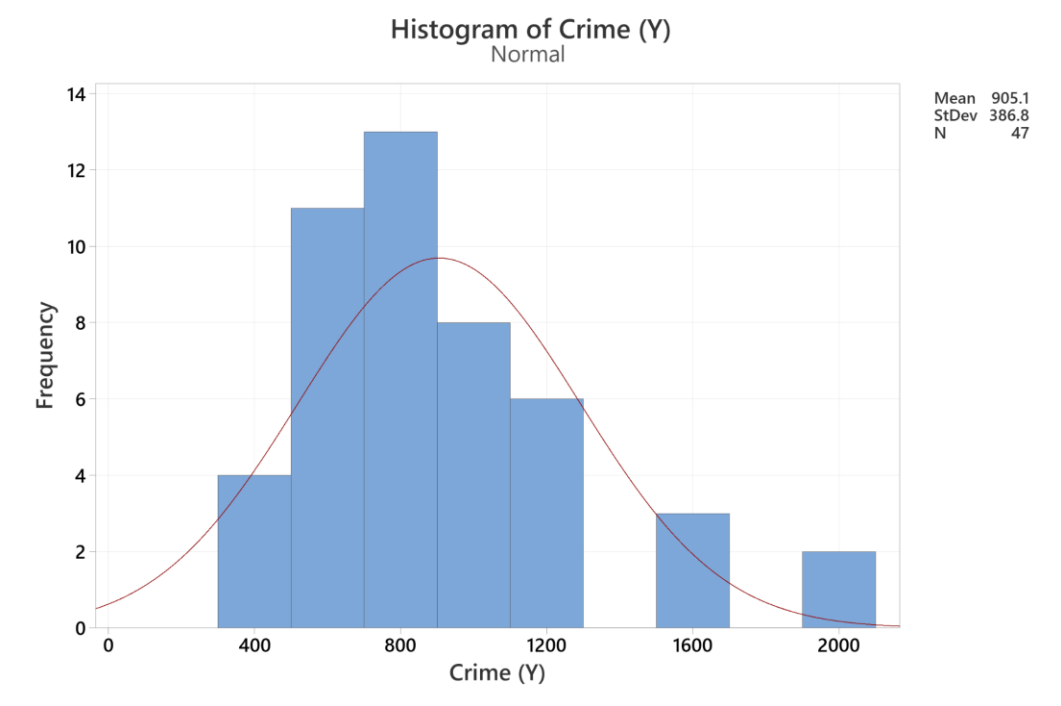

Figure 3.1: Histogram of Crime (Y)

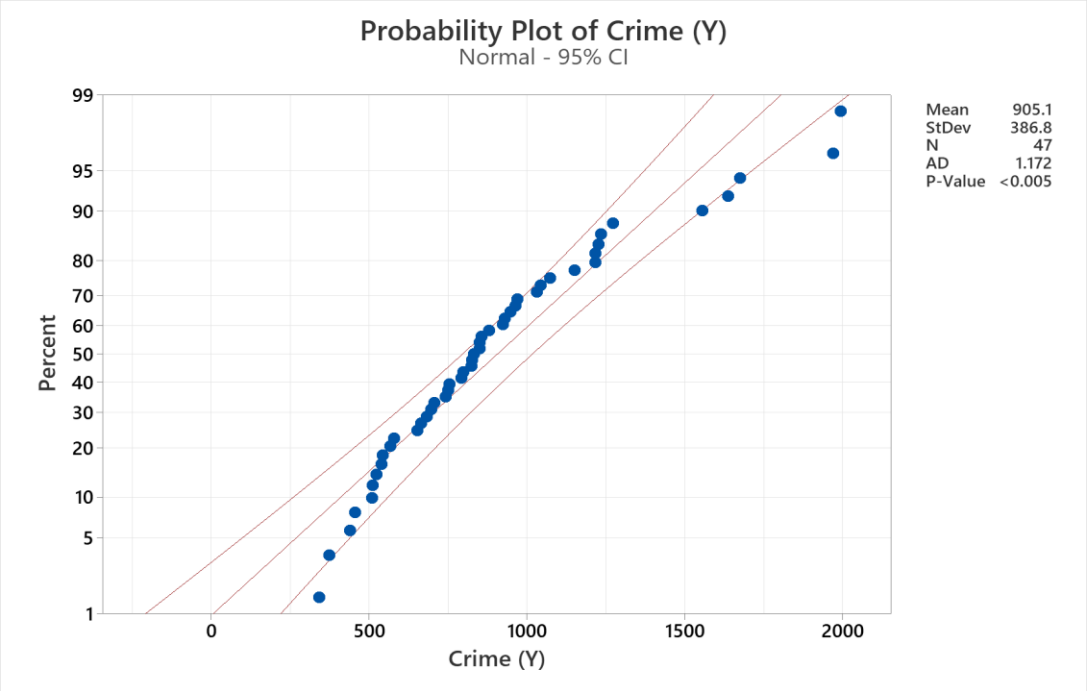

Figure 3.2: Probability Plot of Crime (Y)

In our initial analyses we obtain Figures 3.1 and 3.2 from Minitab which help us derive the following observations:

- **The Histogram** presented in Figure 3.1 for the Crime(Y) variable, presents with some resemblance to a normal distribution. However, a **subtle leftward skewness** can be observed. This skewness implies that, while the general shape of the distribution resembles a normal curve, there is deviation from a perfectly normal distribution.
- The **Probability plot**, in Figure 3.2, **denotes high variability** across data points. Furthermore, the computed **Standard Deviation of 386.8** quantifies the extent of variability within the datapoints, emphasizing the width of the dataset.
- Additionally, the **p-value** associated with the probability plot has a value of **less than 0.005** which is notably less than the conventional significance level of 0.05. These findings from the histogram and probability plot along with the high standard deviation and low p-value are **strongly indicative of non-normality**.
- We can also note the presence of outliers which have to be addressed before developing the model.

From these observations of a high standard deviation and low P-value, *we come to an understanding that the data is not normally distributed*. This information is important when developing a linear regression model as it might impact the reliability, validity and accuracy of the analysis. In general, we prefer to have a larger p-value which would be indicative of a normally distributed variable. So, it is not statistically sound to use Crime(Y) to perform our analysis. These observations *motivate us to perform a transformation of the data* attempting to obtain a more normally distributed dependent variable.

We hence *perform a Log transformation* on the original dependent variable (Y) to obtain Log(Y) and *check for the normality of Log(Y)*. We perform a similar initial analysis on Log(Y) to understand the characteristics. We generate the histograms and probability plots for Log(Y) to understand its distribution and compare it to our original variable (Y).

Figures 3.3 and 3.4 generated using Minitab, are the histogram for Log(Y) and the probability for Log(Y) respectively.

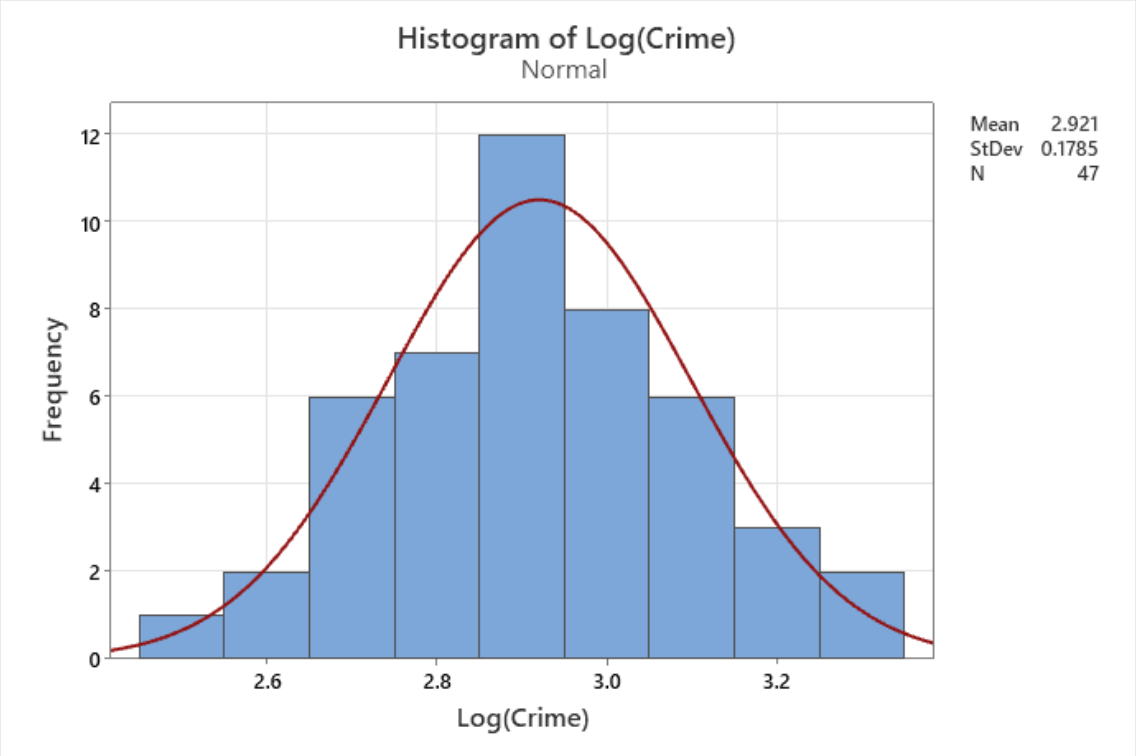

Figure 3.3: Histogram of Log(Y)

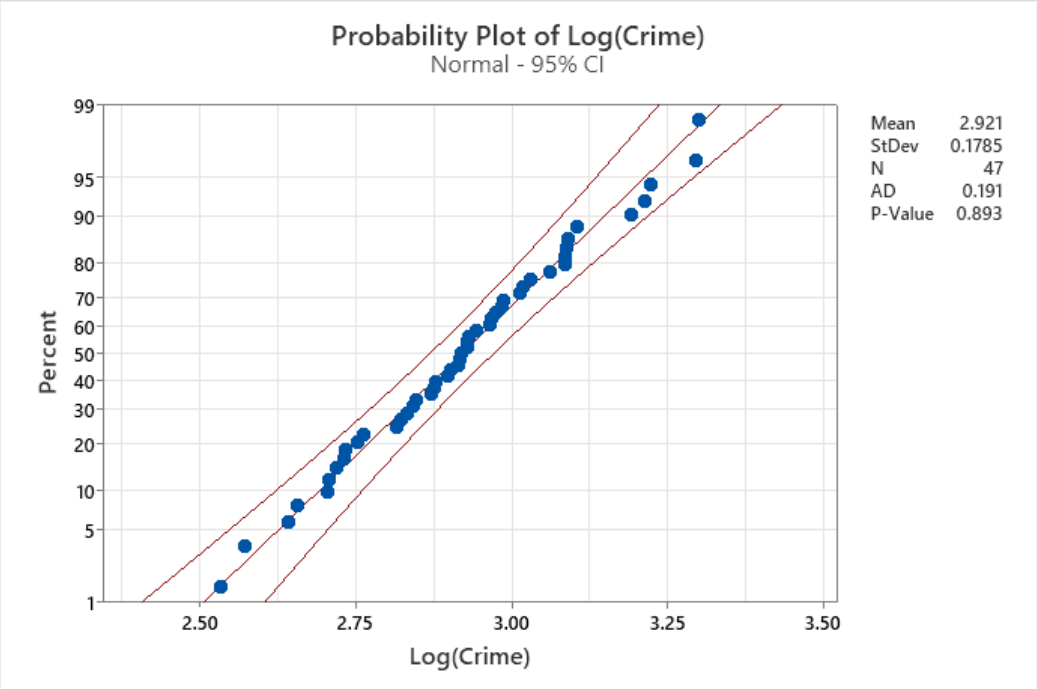

Figure 3.4: Probability plot of Log(Y)

From Figure 3.3 and 3.4 we can observe:

- The **Histogram for Log(Y) is normally distributed** and does not seem to be skewed to either side. Unlike the original Crime variable, the log-transformed data exhibits normality in its distribution.
- The **Standard Deviation for Log(Y) is 0.1785 which is relatively low**. This is indicative of a stabilization of the variance in the original data.
- The **p-Value for the Log(Y)** in the probability plot is **0.893**. This is a very good indication of normality, as the value is higher than the conventional significance level of 0.05.
- In addition to these indicators, we can note that most if not all the points on the probability plot lie along the line and within the Confidence Interval boundaries. We are hence able to note that the outliers previously found in the Crime datapoints have now been mitigated by the Log transformation of the data.
-

With these observations it is evident that Log(Y) shows higher consistency and normality. Furthermore, when compared to our original independent variable (Y), we see a significant rise in the P-value. This escalation in the p-value is indicative of a statistically justified assumption for normality. The outliers present in the Crime(Y) data have been handled with the help of the Log transformation. The Log transformation has hence achieved a more statistically reliable and normal distribution. These observations help us conclude that using Log(Y) as our dependent variable is imperative.

4.0 Model Development

With the incorporation of Log(Y) into our dataset, our aim is now to examine the relationships between independent variables and Log(Y). To identify the most influential attributes for our regression model, correlation analysis is conducted between Log(Y) and each independent variable in the dataset. The outcomes of this correlation exploration will steer the succeeding steps in variable selection, enabling the identification of key contributors for the development of the regression model. The correlation matrix amongst all the independent variables as well as Log(Y) is found in Table 4.1.

	LOG(Y)	PO1	PO2	WEALTH	PROB	POP	ED	U1	U2	LF	M.F	INEQ	TIME
PO1	0.66												
PO2	0.64	0.99											
WEALTH	0.43	0.79	0.79										
PROB	-0.41	-0.47	-0.47	-0.56									
POP	0.34	0.53	0.51	0.31	-0.35								
ED	0.30	0.48	0.50	0.74	-0.39	-0.02							
U1	-0.08	-0.04	-0.05	0.05	-0.01	-0.04	0.02						
U2	0.17	0.19	0.17	0.09	-0.06	0.27	-0.22	0.75					
LF	0.17	0.12	0.11	0.30	-0.25	-0.12	0.56	-0.23	-0.42				
M.F	0.15	0.03	0.02	0.18	-0.05	-0.41	0.44	0.35	-0.02	0.51			
INEQ	-0.15	-0.63	-0.65	-0.88	0.47	-0.13	-0.77	-0.06	0.02	-0.27	-0.17		
TIME	0.14	0.10	0.08	0.00	-0.44	0.46	-0.25	-0.17	0.10	-0.12	-0.43	0.10	
M	-0.06	-0.51	-0.51	-0.67	0.36	-0.28	-0.53	-0.22	-0.24	-0.16	-0.03	0.64	0.12

Table 4.1: Correlation Matrix

For the correlation matrix in Table 4.1, we consider a **threshold value of ±0.35** to highlight a significant correlation. With this threshold value we aim to obtain the independent variables that are most significantly correlated to Log(Y). We can observe that the attributes **Po1, Po2, Wealth and Prob** are most correlated with Log(Y).

In addition to this, we can also observe that **factors Po1 and Po2 are extremely correlated with a factor of 0.99**. Considering *both these variables in our model will lead to multi-collinearity* which will reduce the precision of prediction. When variables in a model are highly colinear amongst themselves, it becomes challenging for the model to deduce the individual contributions toward the dependent variable, leading to high standard errors. To resolve this and to reduce multicollinearity, we made the analytical decision **to remove 'Po2'** from our initial model for the purposes of this report.

Based on these findings and analytical decisions, we keep the following initial explanatory variables as the best predictors for the regression model:

- **Po1**: Per capita expenditure in police protection in 1960,
- **Wealth:** Median value of transferable assets or family income
- **Prob**: Probability of imprisonment

Model A (Po1, Wealth, Prob):

We can develop an initial regression model using the information deduced from the correlation matrix. The results of the initial regression model statistics are displayed in Figure 4.2.

SUMMARY OUTPUT

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	0.234405335	13.21391289	2.99583E-06
Residual	43	0.76278915	0.017739283		
Total	46	1.466005154			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.905567182	0.166303478	17.47147575	3.79136E-21	2.570184256	3.240950109	2.570184256	3.240950109
Po1	0.049157873	0.01074194	4.576256465	4.00127E-05	0.027494686	0.07082106	0.027494686	0.07082106
Wealth	-6.18681E-05	3.50463E-05	-1.765323329	0.08461027	-0.000132546	8.80956E-06	-0.000132546	8.80956E-06
Prob	-1.651295952	1.041130141	-1.586061038	0.120052952	-3.750934986	0.448343081	-3.750934986	0.448343081

Figure 4.2: Base Model Statistical Summary

Additionally, we generate the following information using Minitab for a deeper analysis.

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

Durbin-Watson statistic (DW-stat) = 2.43

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.906	0.166	17.47	0.000	
Po1	0.0492	0.0107	4.58	0.000	2.64
Wealth	-0.000062	0.000035	-1.77	0.085	2.97
Prob	-1.65	1.04	-1.59	0.120	1.45

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.133189	47.97%	44.34%	36.67%

Figure 4.3: Minitab Output for Model A

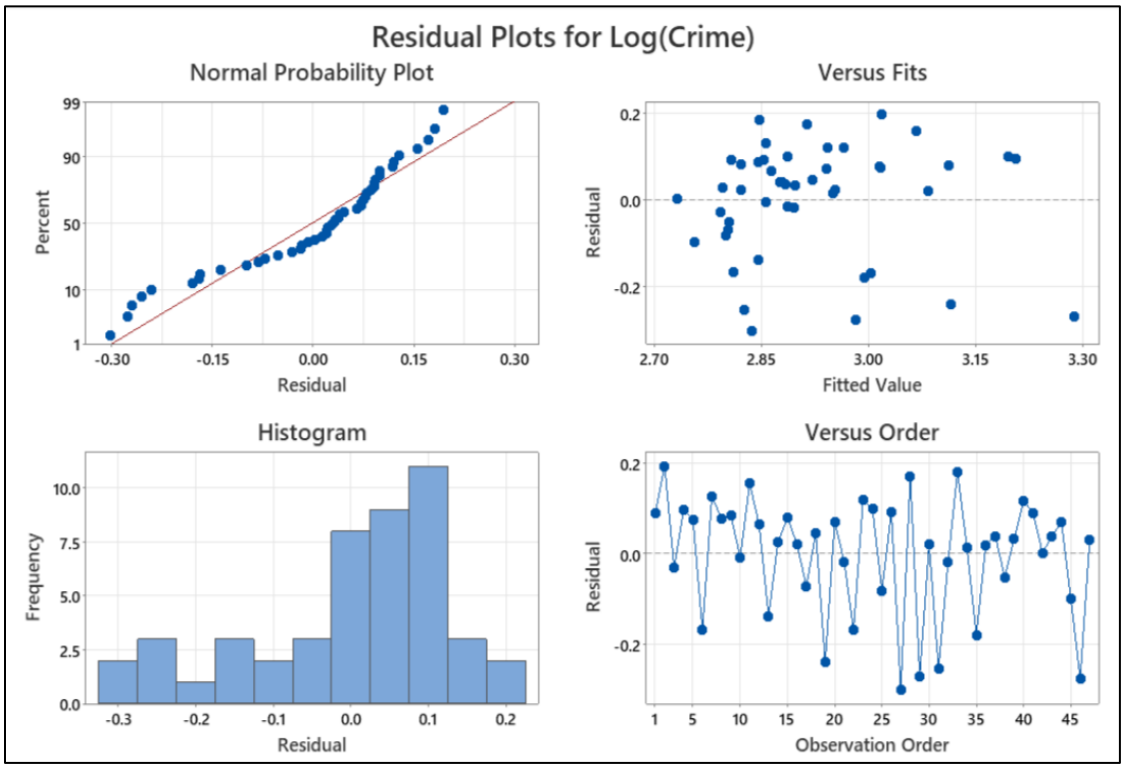

Figure 4.4: Residual Plots for Log(Crime)

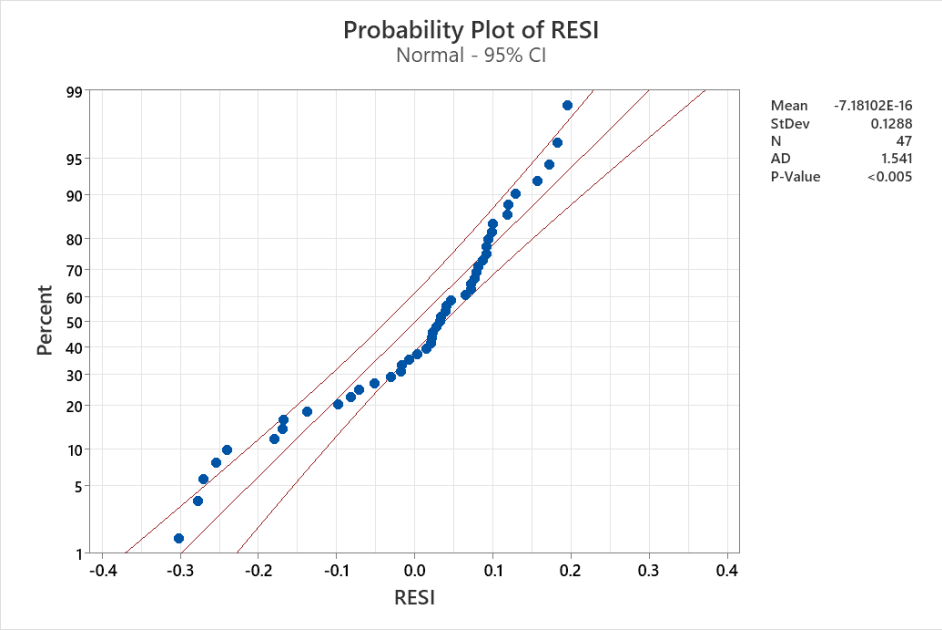

Figure 4.5: Probability plot of Residuals.

Diagnostic Analysis of Model A:

- **R-squared and Adjusted R-squared** – The R-squared and Adjusted R-squared values are statistical measures used to evaluate the goodness of fit of a regression model. The R-squared value does not take into consideration phenomena such as overfitting of the model. The adjusted R-squared value is a modified version, which takes into consideration the number of predictors available in the model. Both these values range from 0 to 1, where a higher value is indicative of a better model. **We observe a R-Squared value of 47.97% and R-squared(adjusted) = 44.34% in Figure 4.2 and 4.3 which is lower than expected**.

- The **p-value** of the independent variables in linear regression analysis is an indicator of the individual contribution of the variable to the model. While developing the model, a hypothesis of all coefficients being 0 is considered. An explanatory variable having a p-value larger than 0.05 would be indicative that the coefficient for that variable is closer to zero and that discarding this variable might help improve the model. **Wealth and Prob have relatively high p-values in Figure 4.3, indicating that dropping these variables can be considered.**

- **Variance Inflation Factor (VIF**) is a statistical measure of multicollinearity amongst the independent variables. If the VIF is in the acceptable range of 0 – 5, it indicates none to minimal multicollinearity amongst the independent variables. If explanatory variables with high VIF values are present in our model, it may lead to deviation in our predictions. **Our VIF as found in Figure 4.3, are in the acceptable range** as of now, **indicating there is no multicollinearity** in our model.

- The **Durbin Watson-statistic** is a measure to detect autocorrelation in the residuals of the regression model. This statistic is important to recognize if there are any patterns in the residuals. The residuals plots, in Figure 4.4 can also be used to visually identify if there are any patterns in the residual values. In the initial model, although moderately high, the **DW-stat is 2.43** is in the

acceptable range of 1.5 to 2.5 which is **indicative of no to minimal autocorrelation** in the residuals.

- The **F-statistic** for the model helps us understand the overall significance of the model. It helps determine whether the model provides a statistically significant improvement over a model with no predictors. The initial model has an **F-statistic of 13.21** which is moderately high, indicating a good model.

With these observations, we can arrive at a conclusion that **although the model is reasonable**, the **low R-squared values suggest** that **adding** more explanatory **variable** will improve the model. When looking at the probability plots for residuals in figure 4.5 we notice non-normality, but this is not enough to invalidate the model. To improve the model, we consider introducing another explanatory variable and perform similar analysis on the next model.

Model B

In the following model we introduce the variable 'M', and perform similar statistical analysis, to determine whether there is improvement in the model.

Log(Crime)	Po1	Wealth	Prob	M
2.898	5.8	3940	0.0846	15.1
3.214	10.3	5570	0.0296	14.3
2.762	4.5	3180	0.0834	14.2
3.294	14.9	6730	0.0158	13.6
3.091	10.9	5780	0.0414	14.1
2.834	11.8	6890	0.0342	12.1
2.984	8.2	6200	0.0421	12.7
3.192	11.5	4720	0.0401	13.1
2.932	6.5	4210	0.0717	15.7
2.848	7.1	5260	0.0445	14

Table 4.6: Adjusted explanatory variables for Model B

With the addition of the new explanatory variable, we now analyze the regression model with the new variable to determine whether the added variable results in any improvement of the model.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.897	0.397	4.78	0.000	
Po1	0.0479	0.0100	4.78	0.000	2.65
Wealth	-0.000011	0.000037	-0.29	0.771	3.91
Prob	-1.612	0.969	-1.66	0.104	1.45
M	0.0541	0.0196	2.76	0.009	1.82

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.123992	55.95%	51.76%	45.13%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.82029	0.205073	13.34	0.000
Po1	1	0.35163	0.351627	22.87	0.000
Prob	1	0.04254	0.042543	2.77	0.104
M	1	0.11708	0.117078	7.62	0.009
Wealth	1	0.00132	0.001325	0.09	0.771
Error	42	0.64571	0.015374		
Total	46	1.46601			

Fits and Diagnostics for Unusual Observations

Obs	Log(Crime)	Fit	Resid	Std Resid	
18	2.9680	2.9543	0.0137	0.16	X
19	2.8751	3.1140	-0.2389	-2.01	R
27	2.5340	2.8381	-0.3041	-2.54	R

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.08821

Figure 4.7 & 4.8: Minitab Output for Model B

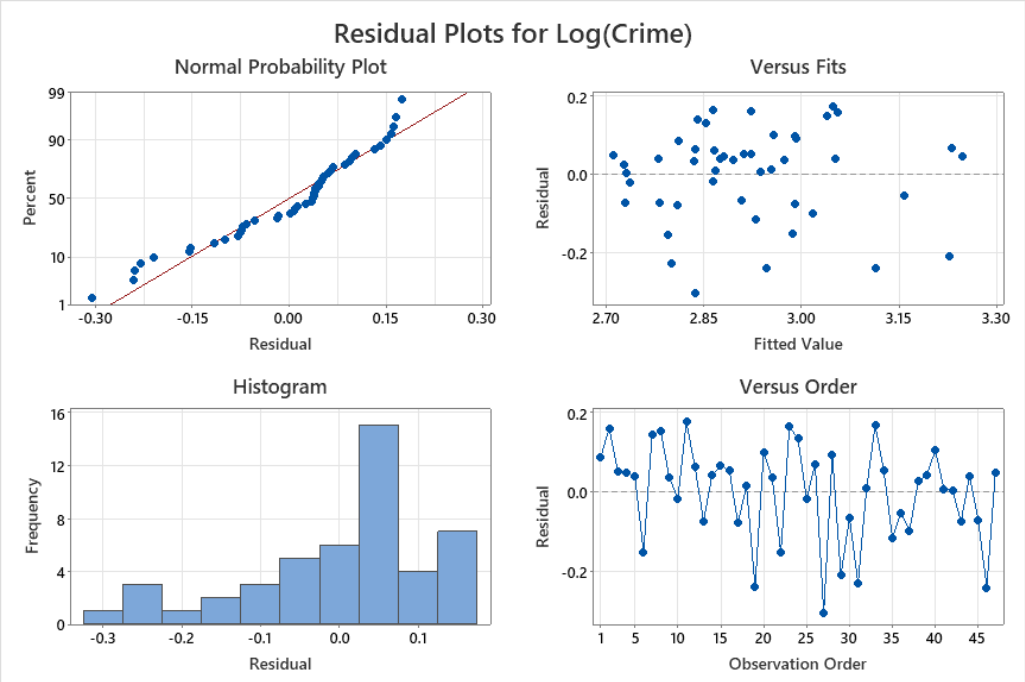

Figure 4.9: Residual plots

Log(Crime) = 0.723 + 0.04056 Po1 + 0.000122 Wealth – 1.455 Prob + 0.0460 M + 0.03304 Ineq

Diagnostic Analysis of Model B:

- The **R-sq(adj)** values has increased from 44.34% to **51.76% indicating model improvement.**
- The **Durbin-Watson statistic** has gone from 2.43 to **2.08**, indicating model improvement.
- **The p-valu**e is again **at 0** which is an indication of a good model.
- The **f-value** has risen from 13.21 to **13.34**. Although there is **not a significant rise** in value it doesn't disclose any reasons to consider the model as worsened. Ultimately, we expect to have a higher f-value.
- The VIF values obtained from Minitab for all the variables are predominantly low, and within acceptable range, hence indicating little to no multicollinearity.
- There **is no apparent heteroscedasticity** in the residual plot. The residuals **seem to be skewed to the right, but this is not enough to disregard the model.**
- The individual **p-value for wealth** is **0.771** indicates that we **may consider dropping** the variable because of the high p-value. **We hence look at our data** and notice that the **values for wealth in our dataset are in the higher thousands**, while the remaining datapoints are on a lower scale.
- The null-hypothesis here, is that all the coefficients are equal to '0'. When encountering a **high p-value**, it is **indicative** that the coefficient for Wealth is closer to if not exactly **'0'**. Considering that the values for the Wealth are extremely high**, it is expected that coefficient for Wealth be extremely low**.
- We hence use our knowledge of the data to better understand the p-value before making the analytical **decision to not remove Wealth** from our model at this stage.

Overall, the improved R-sq(adj), f-value, DW-statistic denote that addition of the variable results in a better fit model.

Although, an improvement on the previous model, the R-sq value achieved in this model is moderately low and we expect a higher value for the R-sq(adj). To achieve this, we test with the addition of another explanatory variable.

Model C

In the following model, we will introduce the explanatory variable Ineq and perform regression analysis to determine whether there is improvement in the model.

Log(Crime)	Po1	Wealth	Prob	Ineq	M
2.898	5.8	3940	0.084602	26.1	15.1
3.214	10.3	5570	0.029599	19.4	14.3
2.762	4.5	3180	0.083401	25	14.2
3.294	14.9	6730	0.015801	16.7	13.6
3.091	10.9	5780	0.041399	17.4	14.1

Table 4.10: Adjusted explanatory variables for Model C

With the addition of the new explanatory variable, we now analyze the regression model with the new variable to determine whether the added variable results in any improvement of the model.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.723	0.470	1.54	0.131	
Po1	0.04056	0.00898	4.52	0.000	2.78
Wealth	0.000122	0.000049	2.52	0.016	8.60
Prob	-1.455	0.850	-1.71	0.094	1.46
M	0.0460	0.0173	2.65	0.011	1.85
Ineq	0.03304	0.00890	3.71	0.001	4.92

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.108555	67.04%	63.02%	55.19%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	0.98285	0.19657	16.68	0.000
Po1	1	0.24024	0.24024	20.39	0.000
Wealth	1	0.07461	0.07461	6.33	0.016
Prob	1	0.03455	0.03455	2.93	0.094
M	1	0.08299	0.08299	7.04	0.011
Ineq	1	0.16256	0.16256	13.79	0.001
Error	41	0.48315	0.01178		
Total	46	1.46601			

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.88136

Figure 4.11: Minitab Output for Model C

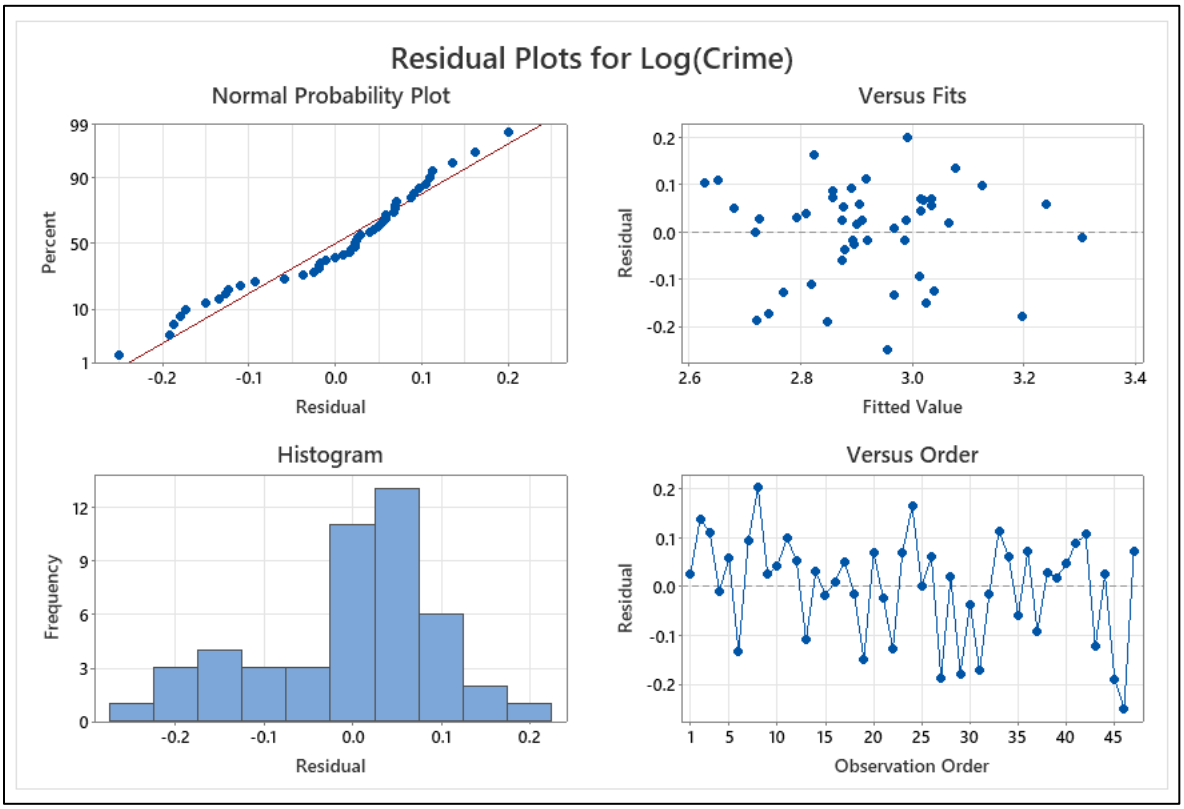

Figure 4.12: Residual Plots Model C

Log(Crime) = 0.723 + 0.04056 Po1 – 1.455 Prob + 0.0460 M + 0.000122 Wealth + 0.03304 Ineq

Diagnostic Analysis of Model C:

- **The R-sq(adj)** has increased from 51.76% to **63.02**% indicating a much better fit model.
- **The p-value** for the model is at 0.000 indicating a good model.
- **The f-value** for the model has increased to **16.68,** which is a significant rise, indicating model improvement.
- The **VIF** values are still in the acceptable range of 5-10 for model C.
- The **DW-statistic is 1.88** which is close to 2.0, indicating no autocorrelation in the variables.
- Additionally, the residual plots for Model C in Figure 4.12 when compared to the initial residuals plot are starting to represent more normality.
- In Figure 4.11 we can also observe that the **p-value for wealth**, which was previously high, is now at **0.016** indicating that it is a crucial contributor.

Overall, when taking into consideration the R-sq(adj), f-value, p-value and residual plots for Model C we can observe clear indication of improvement in the Linear Regression model.

The R-sq(adj) value of 63.02% is moderately high. We attempt to improve our R-squared value more by adding another explanatory variable.

Model D

We add an additional explanatory variable to Model C to improve the R-sq(adj) of our model. With the addition of the variable ‘Ed’ we get a new dataset.

Log(Crime)	Po1	Wealth	Prob	Ineq	M	Ed
2.898	5.8	3940	0.084602	26.1	15.1	9.1
3.214	10.3	5570	0.029599	19.4	14.3	11.3

2.762	4.5	3180	0.083401	25	14.2	8.9
3.294	14.9	6730	0.015801	16.7	13.6	12.1
3.091	10.9	5780	0.041399	17.4	14.1	12.1

Table 4.13: Adjusted explanatory variables for Model D

With the addition of the new explanatory variable, we perform statistical analysis to determine whether there is any improvement in our model. The observations from Minitab are found below.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.014	0.490	-0.03	0.977	
Po1	0.04468	0.00828	5.39	0.000	2.86
Prob	-1.455	0.773	-1.88	0.067	1.46
M	0.0471	0.0158	2.99	0.005	1.85
Wealth	0.000090	0.000045	1.99	0.054	9.08
Ineq	0.04186	0.00859	4.88	0.000	5.53
Ed	0.0648	0.0210	3.08	0.004	2.60

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	1.07565	0.179276	18.37	0.000
Po1	1	0.28396	0.283961	29.10	0.000
Prob	1	0.03457	0.034569	3.54	0.067
M	1	0.08702	0.087021	8.92	0.005
Wealth	1	0.03850	0.038501	3.95	0.054
Ineq	1	0.23197	0.231968	23.77	0.000
Ed	1	0.09280	0.092801	9.51	0.004
Error	40	0.39035	0.009759		
Total	46	1.46601			

Figure 4.14: Minitab Output for Model D

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0987866	73.37%	69.38%	61.17%

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.88882

Figure 4.15: Minitab Output for

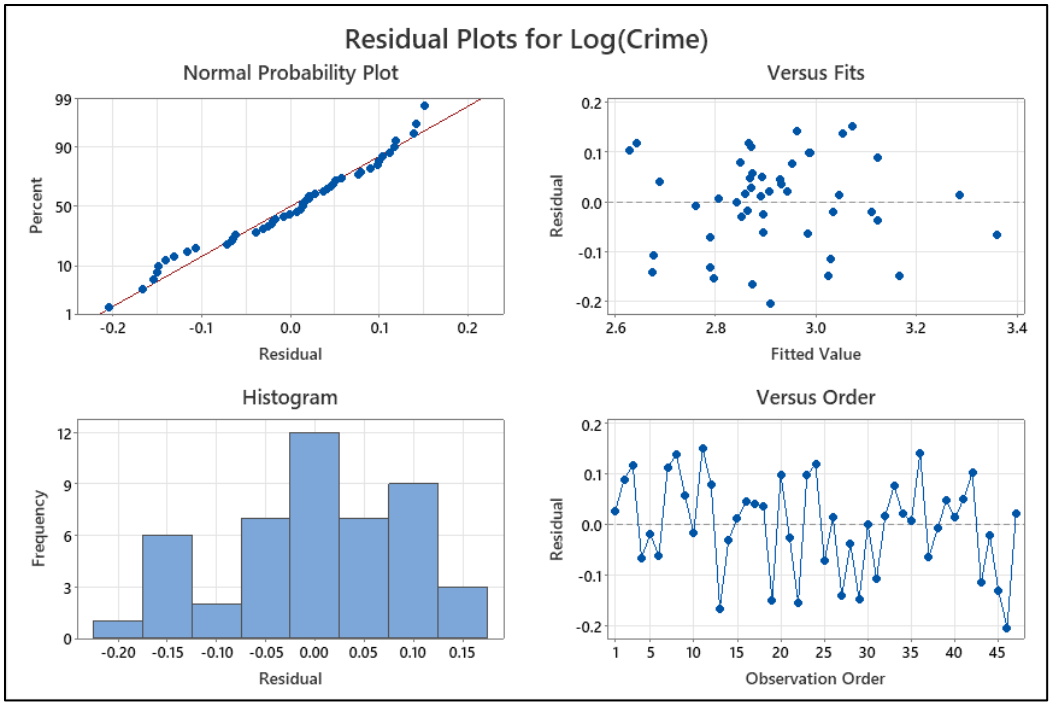

Figure 4.16: Residual plots for Model D

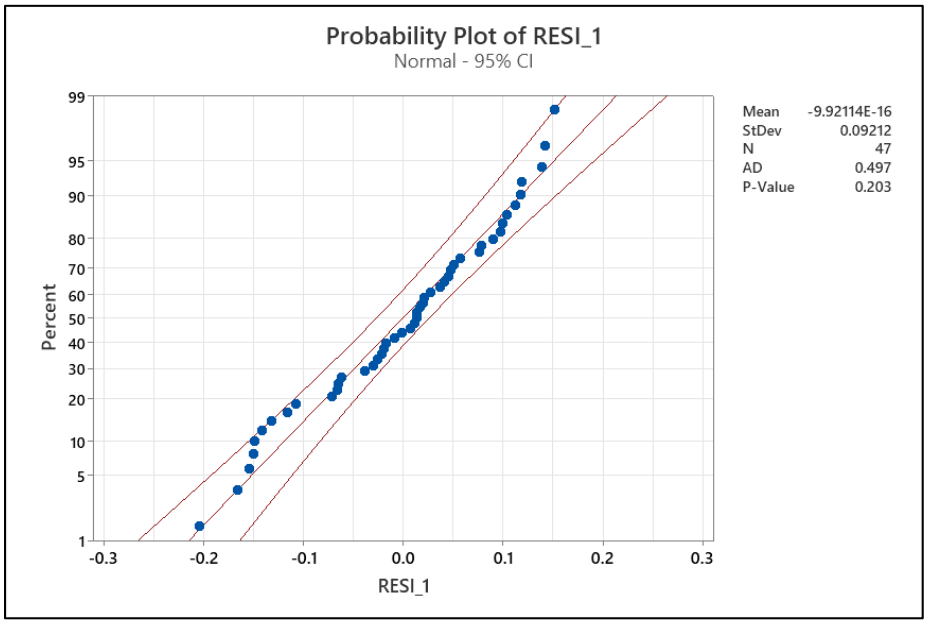

Figure 4.17: Probability plot for Residuals Model D

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

Diagnostic Analysis of Model D:

- The R-sq(adj) value has increased from 63.02% to 69.38% indicating a better fit model.
- The p-value for the model continues to stay at 0.0 indicating a good model.
- The f-value for the model has increased from 16.68 to 18.37. As the f-value has increased to 18.37, it indicated that the model is improved.
- The VIF values are within the acceptable range of 5-10 although the VIF for wealth is slightly high.
- The DW-stat remains at 1.88 indicating that there is no autocorrelation in the residuals.
- Additionally, when we observe the residual plot of in Figures 4.16 and 4.17, we can notice that there is no heteroscedasticity in the residual plot. The probability plot in Figure 4.17, has a p-value of 0.203. This p-value is a strong indication that the residuals are now following normality and hence can be used to perform statistical analysis for the model.

Taking into consideration the increase in the R-sq(adj) value and the f-value, we can say that there is a significant improvement in the model when compared to the previous model. The model currently has a R-sq(adj) value of approximately 69%, which is a moderately high and substantial value.

With the aim of defining the ‘best’ Linear Regression model, a higher r-squared value is preferred. We aim to introduce another variable into our model. We take into consideration that adding too many variables from our data may lead to overfitting and multicollinearity and auto correlation in out model.

Re-interpreting our original data, we understand that the time duration of imprisonment and probability of imprisonment should be impactful in understanding the effect of punishment regimes. Based on subject knowledge, there is an underlying relation between the variables ‘Time’ and ‘Prob’. We can study this relation through the following scatterplot.

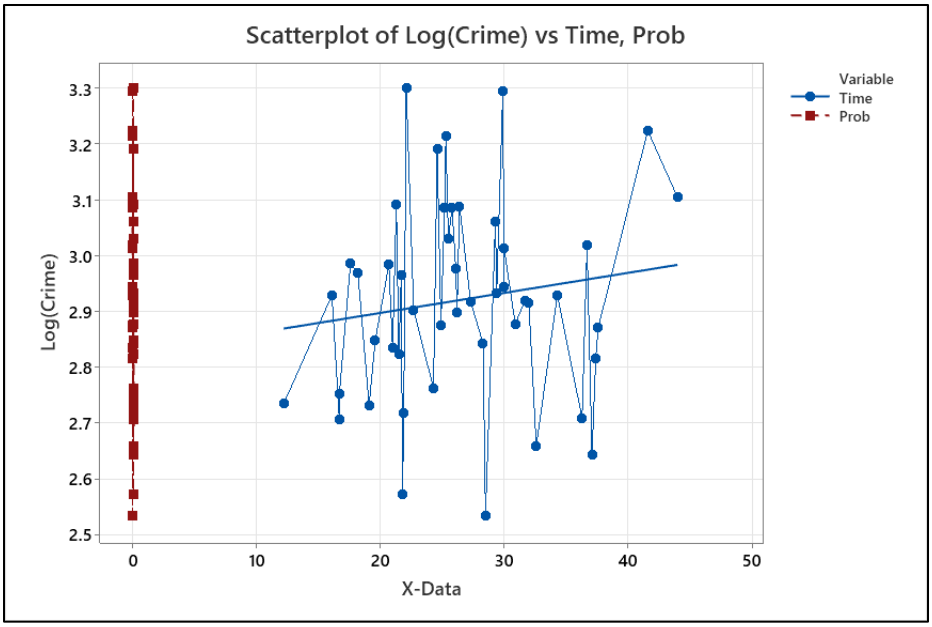

Figure 4.18: Scatterplot of Log(Crime) vs Prob, Time

We observe that individual regressions cross over each other, providing minimal effect in our model individually. This motivates us to explore the effect of their interaction. Upon plotting the interaction term, Prob*Time we observe the following linear interaction (Figure 4.19) which motivates the inclusion of this term in our model.

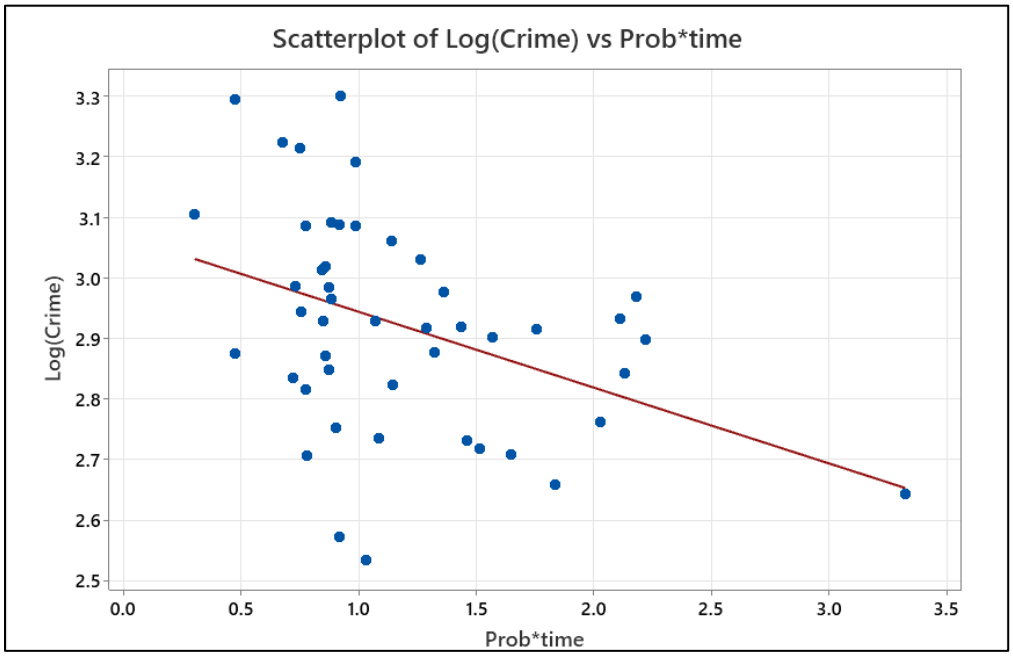

Figure 4.19: Scatterplot of Log(Crime) vs Prob*Time

Model E

We add the additional explanatory variable of 'Time' to Model E while also introducing the interaction term Prob*Time. We hence get the following dataset for our Model E.

Log(Crime)	Po1	Wealth	Prob	Ineq	M	Ed	Time	Prob*Time
2.898	5.8	3940	0.084602	26.1	15.1	9.1	26.2011	2.216665462
3.214	10.3	5570	0.029599	19.4	14.3	11.3	25.2999	0.74885174
2.762	4.5	3180	0.083401	25	14.2	8.9	24.3006	2.026694341
3.294	14.9	6730	0.015801	16.7	13.6	12.1	29.9012	0.472468861
3.091	10.9	5780	0.041399	17.4	14.1	12.1	21.2998	0.88179042

Table 4.20: Adjusted explanatory variables for Model E

With the addition of these two variables, we generate our statistical indicators for Model E and perform the diagnosis of the regression model.

Regression Equation

Log(Crime) = -0.477 + 0.04507 Po1 + 4.28 Prob + 0.0476 M + 0.000086 Wealth + 0.05162 Ineq + 0.0746 Ed + 0.00834 Time - 0.2604 Prob*time

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.477	0.486	-0.98	0.332	
Po1	0.04507	0.00747	6.03	0.000	2.86
Prob	4.28	2.05	2.09	0.044	12.64
M	0.0476	0.0144	3.31	0.002	1.89
Wealth	0.000086	0.000041	2.08	0.044	9.13
Ineq	0.05162	0.00827	6.24	0.000	6.32
Ed	0.0746	0.0209	3.57	0.001	3.18
Time	0.00834	0.00399	2.09	0.043	4.65
Prob*time	-0.2604	0.0798	-3.26	0.002	12.14

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0889791	79.48%	75.16%	68.45%

Figure 4.21: Minitab output for Model E

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	1.16515	0.145644	18.40	0.000
Po1	1	0.28815	0.288153	36.40	0.000
Prob	1	0.03448	0.034481	4.36	0.044
M	1	0.08695	0.086946	10.98	0.002
Wealth	1	0.03439	0.034395	4.34	0.044
Ineq	1	0.30863	0.308631	38.98	0.000
Ed	1	0.10093	0.100931	12.75	0.001
Time	1	0.03455	0.034550	4.36	0.043
Prob*time	1	0.08424	0.084240	10.64	0.002
Error	38	0.30086	0.007917		
Total	46	1.46601			

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.82805

Figures 4.22: Minitab output for Model E

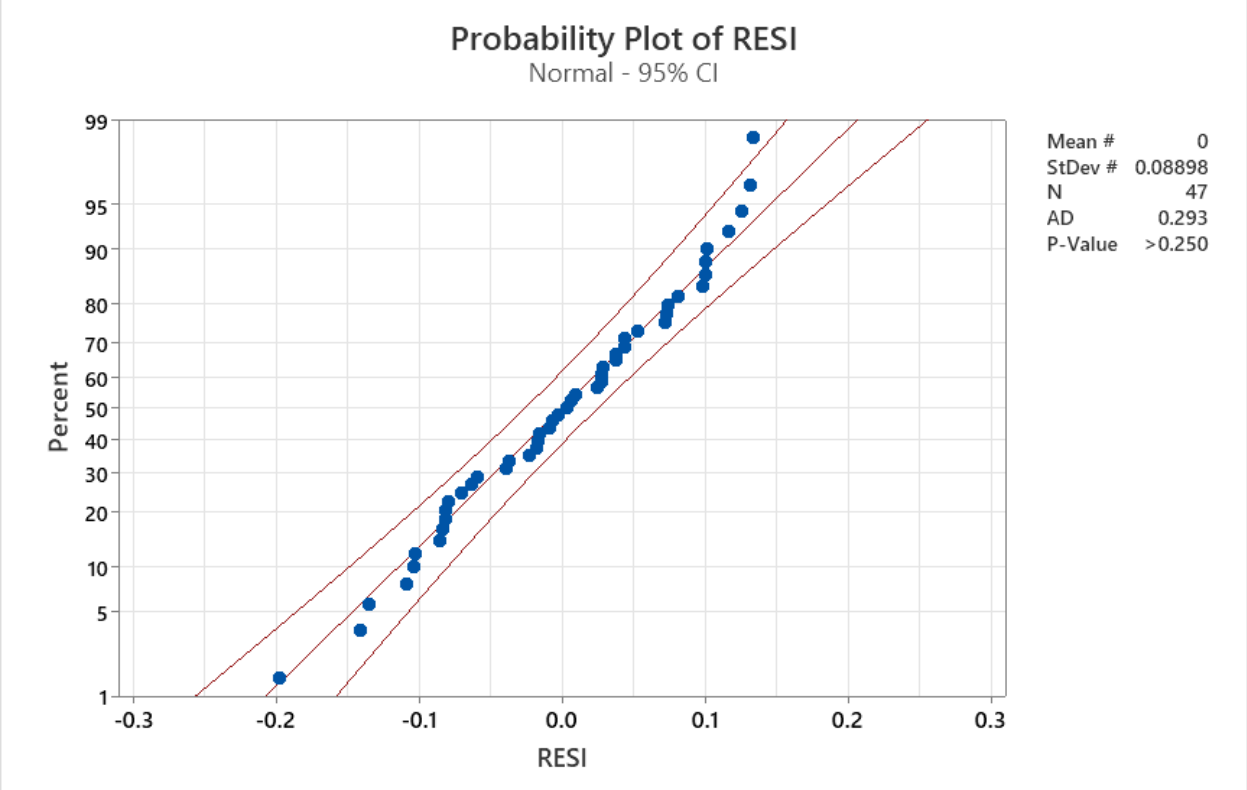

Figure 4.23: Probability plot for Model E

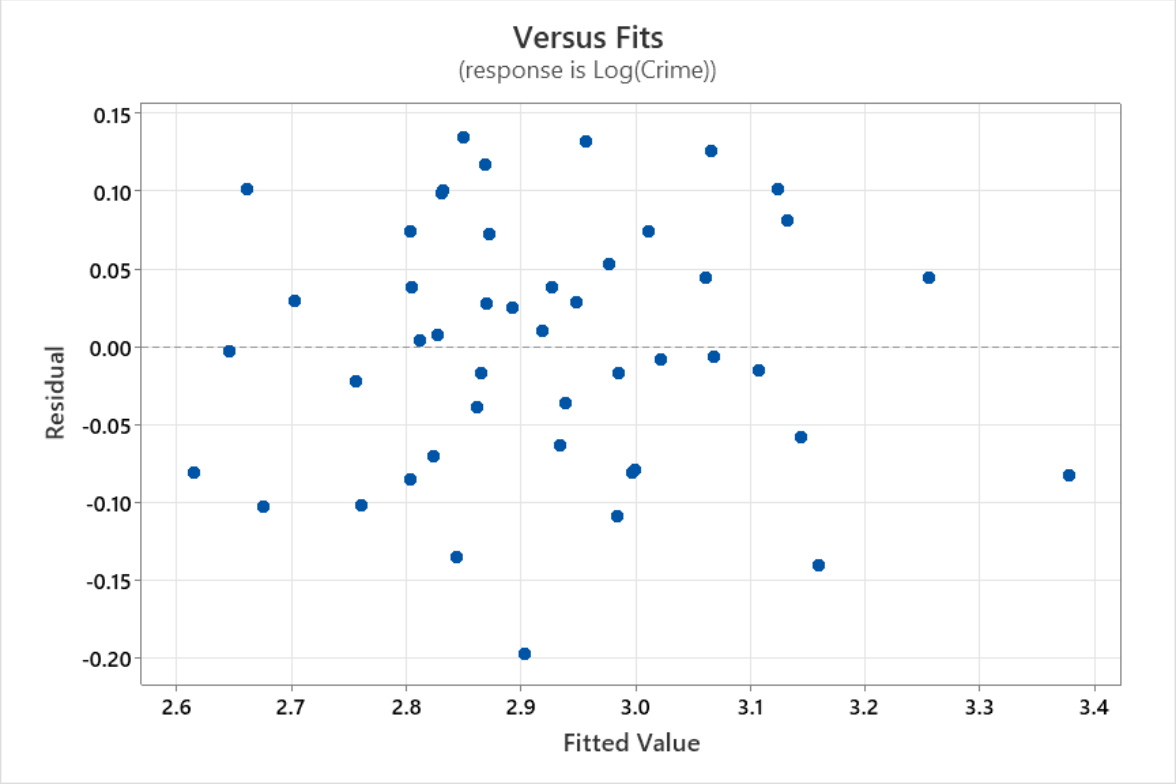

Figure 4.24: Residuals vs Fitted values

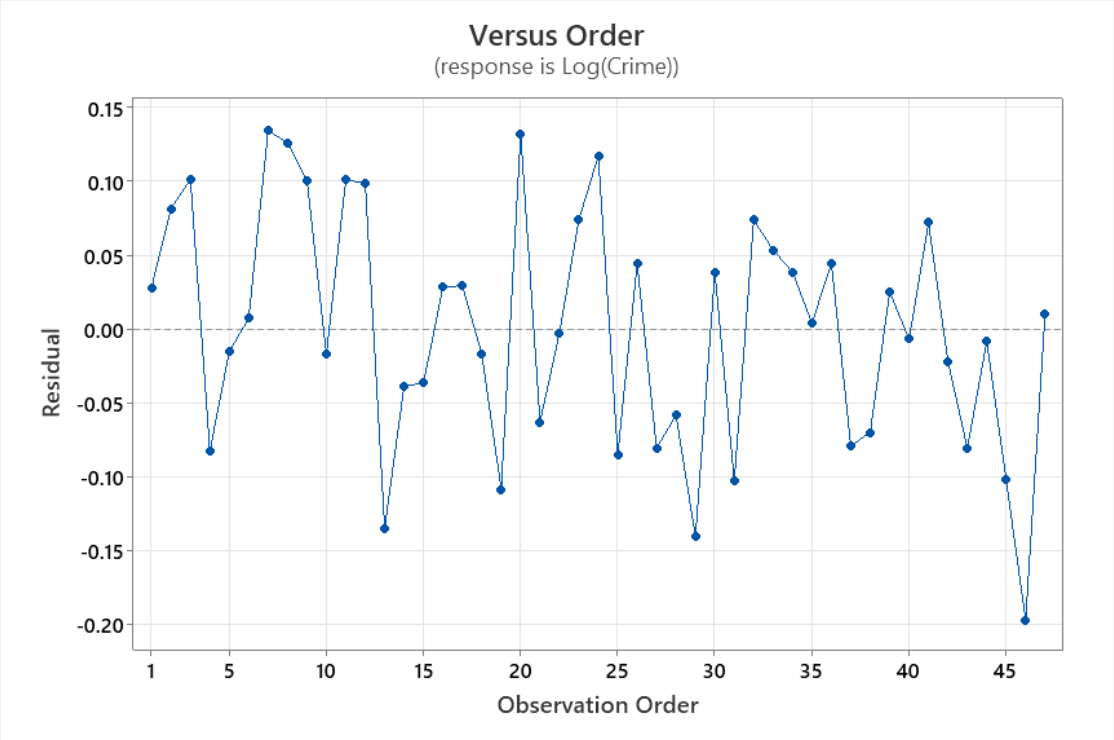

Figure 4.25: Observation of Residuals

Diagnostic Analysis of Model E:

- The R-sq value of the latest regression model is 79.48% and the **R-sq(adj)** value is **75.16%** This value is relatively high and indicates a good regression model.
- The p-value for the regression model is also indicative of a good model.
- The **p-values** for all the explanatory variables are also **below 0.05**.
- The **F-value** for the model is at **18.4** as observed in Figure 4.22. The higher the f-value, the better the model. A value of 18.4 is a good indication that this model can be considered ideally suitable for our analysis.
- The **DW-statistic** is **1.83** which is in the acceptable range of 1.5-2.5. Additionally, since the value is closer to 2.0, it is indicative of minimal to no autocorrelation within the model, which is crucial for our model.
- The **VIF values for the variable Prob and Prob*Time are 12.64 and 12.14** respectively. This is **slightly above** the acceptable range **but is expected**. The VIF factor is **indicative of multicollinearity** within the variables. **Since** we have introduced an **interaction term** using the explanatory variable 'Prob', it is **expected that there will be correlation between these two variables**. Hence the VIF factor, although high, can be disregarded for this case.
- The **residual analysis** is **evident of normality** with a p-value higher than the initial model. We can note **certain outliers** when observing the probability plot in Figure 4.23 but are not enough to invalidate the model.
- There **is no apparent heteroscedasticity** in the residual vs fitted plot in Figure 4.23. The Observation of Residuals is also not evident of any patterns, hence denoting that the variance of the residuals remains relatively constant across different levels of the fitted values. This evidence supports the reliability of the regression model in terms of consistent and unbiased prediction errors.

The model E shows a strong improvement from our initial model. The statistical indicators suggest that this model can be considered as a good linear regression model for our use and analysis.

At this stage we take into consideration that adding too many variables from our data may lead to overfitting, multicollinearity and auto correlation in the model. We

hence consider the effect that removing an independent variable would have on our model.

Based on trend in p-value we understand that ‘Wealth’ and ‘Prob’ have high p-value from our initial to our latest model. We consider dropping the ‘Wealth’ variable since it has a high p-value as well as VIF value.

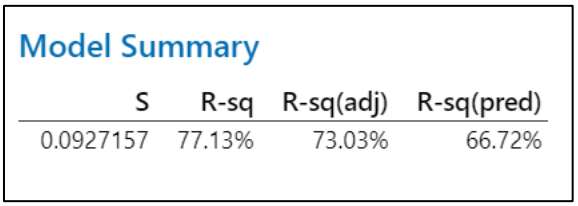

Figure 4.26: Model summary after dropping Wealth

Immediately, we can observe that dropping the explanatory variable ‘Wealth’ leads to a drop in our R-sq and R-sq(adj) values. With similar observation made, when considering the experimental dropping of variables, the idea was quickly disregarded, and the model was finalized.

5.0) Best Regression Model Fit

We can test for improved fit in the adjusted model by thoroughly comparing it to the previous model.

Data Information:
Total datapoints (N) : 47

Small Model Summary (Model D):

Independent Variables	Po1, Wealth, Prob, Ineq, M, Ed
No of Variables (p)	6
R-squared value(R-sq)	73.37%

Large Model Summary(Model E)

Independent Variables	Po1, Wealth, Prob, Ineq, M, Ed, Time, Prob*Time

No of Variables (p)	8
R-squared value(R-sq)	79.48%

Let us consider an initial hypothesis that the Large Model (Model E) has no improvement over the smaller model (Model D)

H_0 = No model improvement
H_1 = Model improvement
Degrees of Freedom (DoF) = N-p-1
We conduct the hypothesis test using the above information, the results of which are available in Figure 5.1

Explanatory Variables in the full model							
Time	Po1	Wealth	Prob	Ed	Ineq	M	Prob*Time
Explanatory Variables in the restricted/small model							
	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							

Figure 5.1: Variables in models

	N	47	**P**
	Full Model		
	R Square	79.48%	8
	Degrees of Freedom (DF)	38	
	Small Model		
	R Square	73.37%	6
	Degrees of Freedom	40	
	Difference R-Squared	6.11%	
	Difference Df	2	
		Value	
	Numerator	0.0306	
	Denominator	0.0054	
	F-Statistic	5.657	
	α	5%	
Method 1	**Critical Value**	3.245	**Conclusion**
	Conclusion	Reject H0	Model Improvement
Method 2	**p-value**	0.71%	**Conclusion**
	Conclusion	Reject H0	Model Improvement
	H_0:	No Model Improvement	
	H_1:	Model Improvement	

Figure 5.2: Hypothesis test for model comparison

Our hypothesis test from Figure 5.1 and 5.2 suggests that Null Hypothesis may be rejected. This suggests that there is significant improvement in the model with the inclusion of the additional explanatory and interaction variables.

Based on this test, we conclude that the best fit regression model for the purposes of this report can be represented by the following information.

Regression Equation:
Log(Crime) = -0.477 + 0.04507 Po1 + 0.000086 Wealth + 4.28 Prob + 0.0476 M + 0.05162 Ineq + 0.0746 Ed + 0.00834 Time - 0.2604 Prob*time
Dependant Variable:
Y = Log(Crime)
Independent Variables:
Po1: Per capita expenditure in police protection in 1960
Wealth: Median value of transferable assets or family income
Prob: Probability of imprisonment: ratio of number of commitments to number of offenses
Ed: Mean years of schooling of the population aged 25 years or over
Ineq: Income inequality
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
Prob*Time: Probability of serving more/less time in prison

As we have finalized our model, we now aim to identify the outliers in our model if any and use the Standardized Residuals and the Deleted Fit values to do so.

p	8	
n	47	
DFIT THRESHOLD	0.87519	
α	0.1	
TRES1 Threshold	1.304854	
INFLUENCIAL OBS.		
Data	TRES	DFIT
8	1.679668	0.98786
13	-1.78638	-0.99271
29	-1.98438	-1.34071
46	-2.59286	-1.11752

Figure 5.3: Influential Observation identification

From the analysis for influential observations, we can get the data points as seen in Figure 5.3. From the probability plot for residuals in Figure 4.23 we can identify that these points are on the extremes ends of the plot, hence verifying our statistical analysis. The model is used to perform a prediction on the given values as found in figure 5.4. The predictions are then interpreted in Figure 5.5 as part.

	x0	b-hat
Intercept	1	-0.476995537
Time	44	0.008336001
Po1	16	0.045073295
Wealth	6890	8.56488E-05
Prob	0.01	4.281727465
Ed	12	0.07464985
Ineq	27	0.051617312
M	17	0.047590009
Prob*Time	0.44	-0.260448722

Figure 5.4: Intercepts and Prediction values

MINITAB OUTPUT

PFITS	PSEFITS	CLIM	CLIM_1	PLIM	PLIM_1
4.227797203	0.139431777	3.945532328	4.510062078	3.89295417	4.562640232

			Variances
μ = LOG(Crime) - hat	4.228	Standard Error Residuals	0.146820	0.021556
MEDIAN[Crime]	16,896.52	Standard Error LOG(Price-hat)	0.139432	0.019441
E[Crime]	18,836.37	σ^2 = Var[Y]=Var[Log(Price)]		0.040997
		σ = Standard Deviation [Log(Price)]	0.202478	

95% Confidence Interval		95% Prediction Interval (or Credibility Interval)	
LB E[LOG(Crim)]	3.94553	LB LOG(Crime)	3.892954
UB E[LOG(Crime)]	4.51006	UB LOG(Crime)	4.562640
UB - LB	0.56453	UB - LB	0.66969
Approximate 95% Confidence Interval		**95% Prediction Interval (or Credibility Interval)**	
LB E[Crime]	8,821.29	Crime	7,815.45
UB E[Crime]	32,363.99	Crime	36,529.21
UB - LB	23,542.70	UB - LB	28,713.75

Figure 5.5: Model E Prediction

95% Confidence Interval for E[Log(Crime)] = (3.94,4.51)

95% Prediction Interval for E[Log(Crime)] = (3.89,4.56)

95% Confidence Interval for E[Crime] = (8,821.29, 32,363.99)

95% Prediction Interval for E[Crime] = (7,815.45, 36,529.2)

Conclusion: The E[Crime] values maybe be exaggerated because Log is not a linear function, but we can see that **the E[Crime] value is 18,836.37** for the given datapoints which indicates a relatively median high rate for the given data. Based on this we can comment **that there will be a rise in crime rate if the respective factors are as indicated.**