Data Analysis Regression Project

Punishment Regimes on Crime Rates Using Multiple Regression

Karen Rius

EMSE 6765

Dr. Johan van Drop

December 10, 2020

Table of Contents

INTRODUCTION

Currently, criminologists are interested in the effect of punishment regimes on crime rates. In the data provided, city crime rate in the US is determined by a number of attributes. For this regression analysis report, we were provided with crime rate data Y along with the candidate attributes $X_{1,\ldots,}X_{13}$. With this data, we will complete the following analysis:

1. Develop a linear regression model of $Log(Y)$ on a relevant set of explanatory variables;
2. Perform a diagnostic analysis of the fitted model; and
3. Forecast the crime rate of a state using the following independent variables: $X_1 = 16$, $X_2 = 15$, $X_3 = 6890$, $X_4 = 0.01$, $X_5 = 168$, $X_6 = 12$, $X_7 = 0.14$, $X_8 = 5$, $X_9 = 0.6$, $X_{10} = 107$, $X_{11} = 27$, $X_{12} = 44$, and $X_{13} = 17$.

Table 1 shows the crime rate data for 47 cities in the U.S for 1960 over 13 explanatory variables.

The dependent variables are as follows:

- Crime Rate Data, Y
- Log of Crime Rate $Log(Y)$

The independent variables are as follows:

- Per capita expenditure in police protection in 1960 X_1
- Per capita expenditure in police protection in 1959 X_2
- Wealth: median value of transferrable assets or family income X_3
- Probability of Imprisonment: ratio of number of commitments to number of offenses X_4
- State population in 1960 in hundred thousand X_5
- Mean years of schooling of the population aged 25 years or over X_6
- Unemployment rate of urban male 14-24 X_7
- Unemployment rate of urban males 35-39-24 X_8
- Labour force participation rate of civilian urban male in the age-group 14-24 X_9
- Number of males per 100 females X_{10}
- Income inequality: percentage of families earning below half the median income X_{11}
- Average time in months served by offenders in state prisons before their first release X_{12}
- Percentage of males aged 14-24 in total state population X_{13}

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
1674	12.1	11.6	6570	0.016201	101	10.5	0.077	3.5	0.58	96.6	17	41.6	12.4
849	7.5	7.1	5800	0.031201	47	10.8	0.083	3.1	0.595	97.2	17.2	34.2984	13.4
511	6.7	6	5070	0.045302	28	11.3	0.077	2.5	0.624	97.2	20.6	36.2993	12.8
664	6.2	6.1	5290	0.0532	22	11.7	0.077	2.7	0.595	98.6	19	21.501	13.5
798	5.7	5.3	4050	0.0691	30	8.7	0.092	4.3	0.53	98.6	26.4	22.7008	15.2
946	8.1	7.7	4270	0.052099	33	8.8	0.116	4.7	0.497	95.6	24.7	26.0991	14.2
539	6.6	6.3	4870	0.076299	10	11	0.114	3.5	0.537	97.7	16.6	19.1002	14.3
929	12.3	11.5	6310	0.119804	31	10.4	0.089	3.4	0.537	97.8	16.5	18.1996	13.5
750	12.8	12.8	6270	0.019099	51	11.6	0.078	3.4	0.536	93.4	13.5	24.9008	13
1225	11.3	10.5	6260	0.034801	78	10.8	0.13	5.8	0.567	98.5	16.6	26.401	12.5
742	7.4	6.7	5570	0.0228	34	10.8	0.102	3.3	0.602	98.4	19.5	37.5998	12.6
439	4.7	4.4	2880	0.089502	22	8.9	0.097	3.4	0.512	96.2	27.6	37.0994	15.7
1216	8.7	8.3	5130	0.0307	43	9.6	0.083	3.2	0.564	95.3	22.7	25.1989	13.2
968	7.8	7.3	5400	0.041598	7	11.6	0.142	4.2	0.574	103.8	17.6	17.6	13.1
523	6.3	5.7	4860	0.069197	14	11.6	0.07	2.1	0.641	98.4	19.6	21.9003	13
1993	16	14.3	6740	0.041698	3	12.1	0.102	4.1	0.631	107.1	15.2	22.1005	13.1
342	6.9	7.1	5640	0.036099	6	10.9	0.08	2.2	0.54	96.5	13.9	28.4999	13.5
1216	8.2	7.6	5370	0.038201	10	11.2	0.103	2.8	0.571	101.8	21.5	25.8006	15.2
1043	16.6	15.7	6370	0.0234	168	10.7	0.092	3.6	0.521	93.8	15.4	36.7009	11.9
696	5.8	5.4	3960	0.075298	46	8.9	0.072	2.6	0.521	97.3	23.7	28.3011	16.6
373	5.5	5.4	4530	0.041999	6	9.3	0.135	4	0.535	104.5	20	21.7998	14
754	9	8.1	6170	0.042698	97	10.9	0.105	4.3	0.586	96.4	16.3	30.9014	12.5
1072	6.3	6.4	4620	0.049499	23	10.4	0.076	2.4	0.56	97.2	23.3	25.5005	14.7
923	9.7	9.7	5890	0.040799	18	11.8	0.102	3.5	0.542	99	16.6	21.6997	12.6
653	9.7	8.7	5720	0.0207	113	10.2	0.124	5	0.526	94.8	15.8	37.4011	12.3
1272	10.9	9.8	5590	0.0069	9	10	0.087	3.8	0.531	96.4	15.3	44.0004	15
831	5.8	5.6	3820	0.045198	24	8.7	0.076	2.8	0.638	97.4	25.4	31.6995	17.7
566	5.1	4.7	4250	0.053998	7	10.4	0.099	2.7	0.599	102.4	22.5	16.6999	13.3
826	6.1	5.4	3950	0.047099	36	8.8	0.086	3.5	0.515	95.3	25.1	27.3004	14.9

Table 1. Original Crime Rate Data (cont. on next page)

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**
1151	8.2	7.4	4880	0.038801	96	10.4	0.088	3.1	0.56	98.1	22.8	29.3004	14.5
880	7.2	6.6	5900	0.0251	9	12.2	0.084	2	0.601	99.8	14.4	30.0001	14.8
542	5.6	5.4	4890	0.088904	4	10.9	0.107	3.7	0.523	96.8	17	12.1996	14.1
823	7.5	7	4960	0.054902	40	9.9	0.073	2.7	0.522	99.6	22.4	31.9989	16.2
1030	9.5	9.6	6220	0.0281	29	12.1	0.111	3.7	0.574	101.2	16.2	30.0001	13.6
455	4.6	4.1	4570	0.056202	19	8.8	0.135	5.3	0.48	96.8	24.9	32.5996	13.9
508	10.6	9.7	5930	0.046598	40	10.4	0.078	2.5	0.599	98.9	17.1	16.6999	12.6
849	9	9.1	5880	0.052802	3	12.1	0.113	4	0.623	104.9	16	16.0997	13

Table 1. Original Crime Rate Data (cont.)

INITIAL OBSERVATIONS

In order to begin our multiple regression analysis, we must first study the distribution of the dependent variable *Y* across various crime rate attributions $X_1, \ldots, X_{13}$. To do this, a histogram plot and normal probability plot will be generated in Minitab. The plots will be generated under the assumption that the data is normally distributed.

Taking a look at the first histogram generated (**Figure 1)**, we observe that the data is not symmetric and skewed to the left toward lower crime rates. We can also see that the standard deviation is quite large at 386.8. These initial observations made can pose issues in our regression analysis.

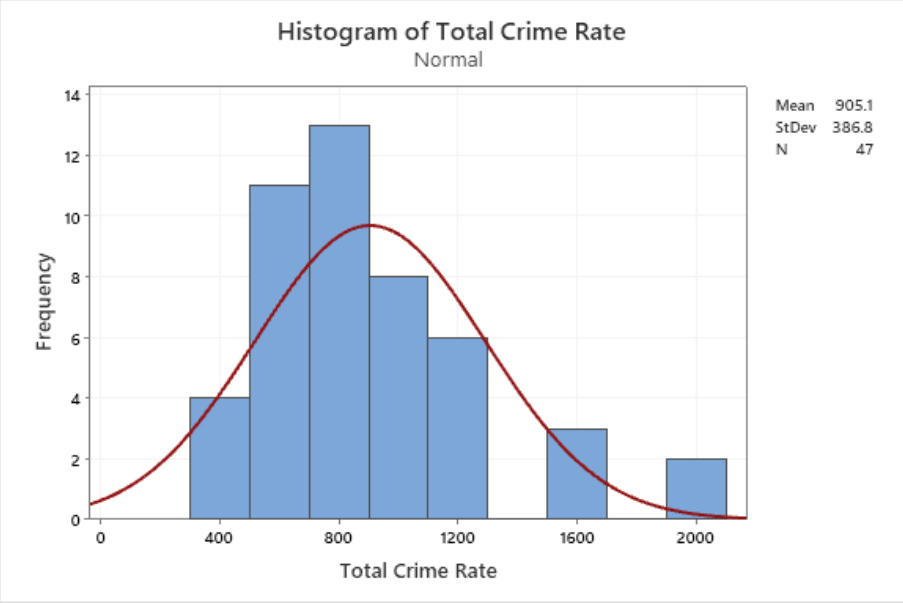


Figure 1. Histogram of Total Crime Rate

Another method we can use to analyze the initial data is to graph the probability plot. By examining the probability plot in **Figure 2,** we observe that the data fails to make a straight line and multiple outliers are seen. This indicates that the data is most likely not normally distributed. We also see that the p-value is <0.005, indicating yet again that the data does not follow a normal distribution.

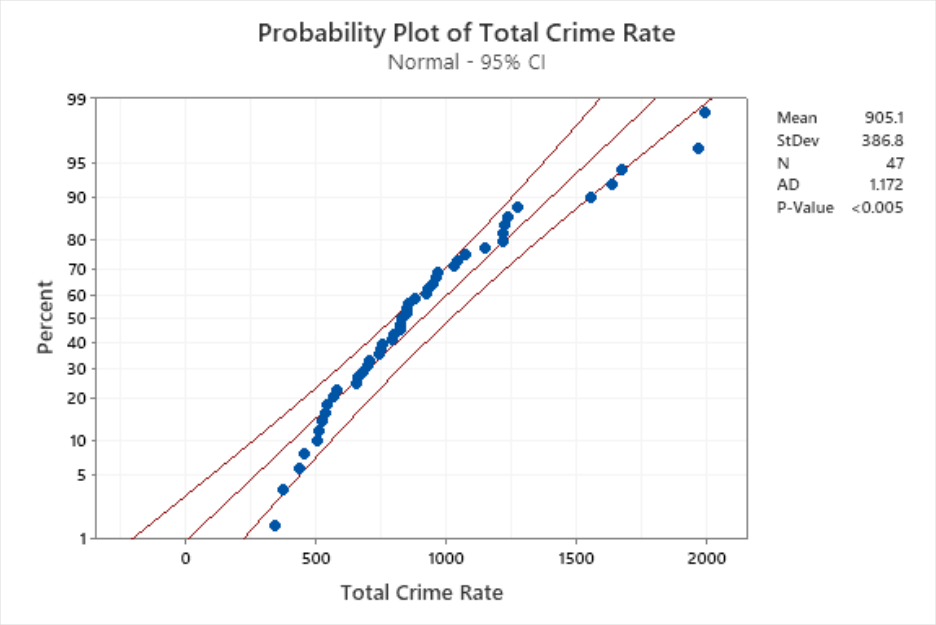


Figure 2. Normal Probability Plot of Total Crime Rate

Due to the fact our initial observations of dependent variable *Y* did not show a symmetric, bell shaped curve, we will now replace *Y* for the *Log(Y).*

As shown in **Figure 3,** replacing the dependent variable to *Log(Y)* generated a histogram plot with much better symmetry. We can also see that in addition to improved symmetry, the standard deviation has greatly decreased from the original 386.8 to 0.1785.

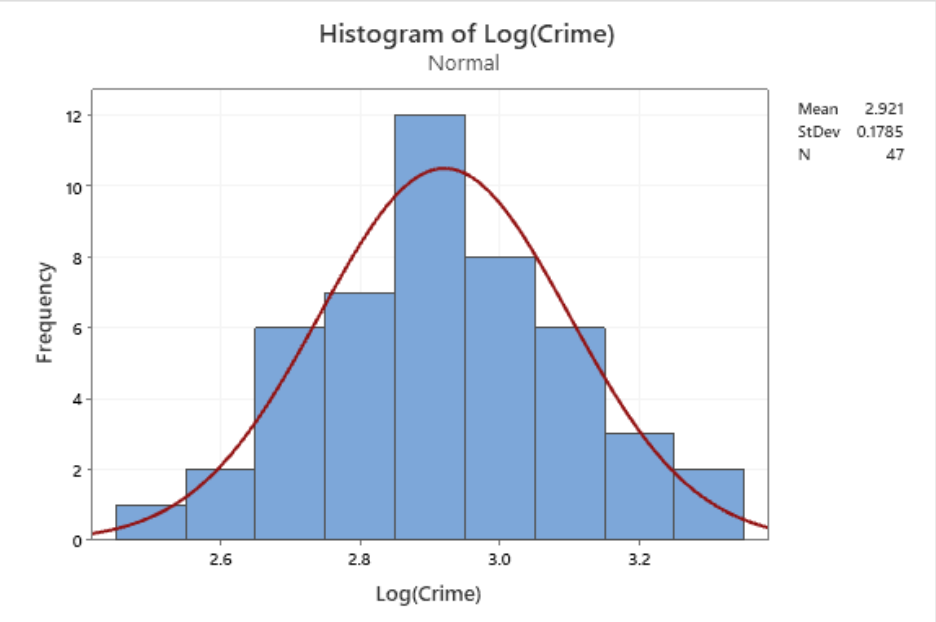


Figure 3. Histogram Plot of Log(Total Crime)

The normal probability plot also shows great improvement (**Figure 4).** In the normal probability plot below, we can see that the data now follows a much more linear straight line with no outliers. We also see that the Anderson-Darling (AD) statistic is reduced from the original 1.172 value to 0.191 while the p-value has increased drastically. An increased p-value indicates greater normality in the data. With this new information, we can now fail to reject normality in the *Log(Y)* data.

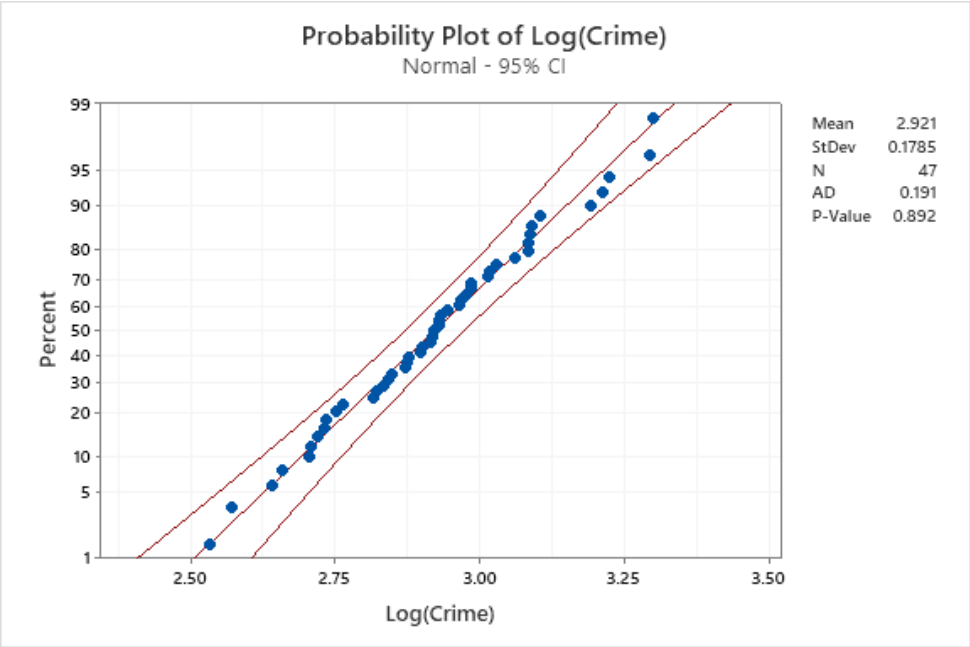


Figure 4. Normal Probability Plot of Log(Total Crime)

CORRELATION ANAYLSIS

Before beginning our first regression model, a correlation matrix must be constructed to compare the dependent variable *Log(Total Crime Rate)* to the independent variables X_1-X_{13}. The correlation matrix will display correlation measures of the linear dependence between the variables to give us a basis of what explanatory variables will be used for the first regression model. This matrix is displayed in **Figure 5** below.

	Log(Crime)	*Po1*	*Po2*	*Wealth*	*Prob*	*Pop*	*Ed*	*U1*	*U2*	*LF*	*M.F*	*Ineq*	*Time*	*M*
Log(Crime)	1.000													
Po1	0.655	1.000												
Po2	0.637	0.994	1.000											
Wealth	0.427	0.787	0.794	1.000										
Prob	-0.412	-0.473	-0.473	-0.555	1.000									
Pop	0.337	0.526	0.514	0.308	-0.347	1.000								
Ed	0.302	0.483	0.499	0.736	-0.390	-0.017	1.000							
U1	-0.075	-0.044	-0.052	0.045	-0.007	-0.038	0.018	1.000						
U2	0.167	0.185	0.169	0.092	-0.062	0.270	-0.216	0.746	1.000					
LF	0.173	0.121	0.106	0.295	-0.250	-0.124	0.561	-0.229	-0.421	1.000				
M.F	0.148	0.034	0.023	0.180	-0.051	-0.411	0.437	0.352	-0.019	0.514	1.000			
Ineq	-0.152	-0.631	-0.648	-0.884	0.465	-0.126	-0.769	-0.064	0.016	-0.270	-0.167	1.000		
Time	0.143	0.103	0.076	0.001	-0.436	0.464	-0.254	-0.170	0.101	-0.124	-0.428	0.102	1.000	
M	-0.056	-0.506	-0.513	-0.670	0.361	-0.281	-0.530	-0.224	-0.245	-0.161	-0.029	0.639	0.115	1.000

Figure 5. Correlation Matrix of Total Crime Rate with a Threshold of 0.4

The above correlation matrix contains a threshold of 0.4 and any explanatory variables above the defined threshold were considered significant. Based on this matrix, it was determined that a strong correlation exists between *Log(Total Crime Rate)* and X_1, X_2, X_3, and X_4.The variables are defined as follows: Per capita expenditure in police protection in 1960 (X_1), per capita expenditure in police protection in 1960 (X_2), wealth: median value of transferrable assets or family income (X_3), and probability of imprisonment: ratio of number of commitments to number of offenses (X_4).

Based on these findings, we will use the defined explanatory variables to begin our first regression model. It is also important to note however, that variables X_1-X_4 also seem to be highly correlated with one another which may cause multicollinearity.

FIRST REGRESSION MODEL

Based on the initial explanatory variables found in the correlation matrix, the first regression model was constructed below in **Figure 6.**

WORKSHEET 1

Regression Analysis: Log(Crime) versus Po1, Po2, Wealth, Prob

Regression Equation

Log(Crime) = 2.892 + 0.0946 Po1 - 0.0498 Po2 - 0.000057 Wealth - 1.62 Prob

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.892	0.168	17.22	0.000	
Po1	0.0946	0.0588	1.61	0.115	78.43
Po2	-0.0498	0.0634	-0.79	0.436	80.72
Wealth	-0.000057	0.000036	-1.59	0.119	3.06
Prob	-1.62	1.05	-1.55	0.128	1.45

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.133802	48.72%	43.83%	34.09%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.71429	0.17857	9.97	0.000
Po1	1	0.04635	0.04635	2.59	0.115
Po2	1	0.01106	0.01106	0.62	0.436
Wealth	1	0.04546	0.04546	2.54	0.119
Prob	1	0.04307	0.04307	2.41	0.128
Error	42	0.75192	0.01790		
Total	46	1.46622			

Fits and Diagnostics for Unusual Observations

Obs	Log(Crime)	Fit	Resid	Std Resid	
18	2.9680	2.9286	0.0394	0.43	X
27	2.5340	2.8111	-0.2771	-2.22	R
29	3.0180	3.2791	-0.2611	-2.25	R
46	2.7060	2.9980	-0.2920	-2.25	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.34470

Figure 6. First Regression Model using X_1, X_2, X_3, and X_4

Looking at the initial model, we can see that R-sq and R-sq(adj) values are moderately high, but not as high as we would like to see in regression analysis. Also, in order to test for the predicted multicollinearity between variables, the VIF values were added to the model. Right away, one can see that X_1 - Per capita expenditure in police protection in 1960, and X_2 - Per capita expenditure in police protection in 1959 are both greater than 5, indicating that regression coefficients were inadequately estimated.

Although observations can also be made regarding p-values and the Durbin Watson coefficient, the multicollinearity of the X_1 and X_2 indicates that one or more variables need to be eliminated in order to conduct a better regression analysis.

We can also see in the probability plot of the residuals (**Figure 7**), that a linear pattern is not displayed, and we do have an outlier. The Anderson-Darling statistic also remains high in the probability plot of the residuals.

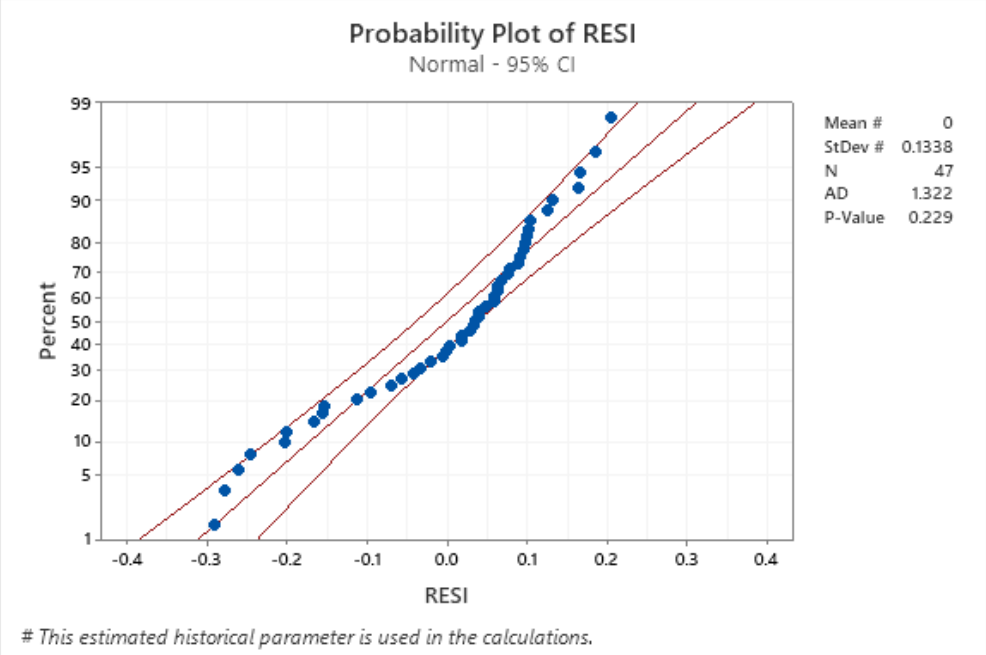


Figure 7. First Regression Model Probability Plot

SECOND REGRESSION MODEL

As discussed previously, in order to optimize our regression analysis, the two variables with the highest VIF values (indicating high collinearity) were removed. When removing both variables however, the R-values were not preserved and significantly decreased from 48.7% and 43.83% to 22.62% and 19.10%. Therefore, in order to ensure the R-values did not drastically change, variables X_2 and X_4 were eliminated. The preserved variables are defined as follows: Per capita expenditure in police protection in 1960 (X_1) and wealth: median value of transferrable assets or family income (X_3). The results are summarized in **Figure 8** below.

WORKSHEET 1

Regression Analysis: Log(Crime) versus Po1, Wealth

Regression Equation

Log(Crime) = 2.719 + 0.0504 Po1 - 0.000043 Wealth

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.719	0.120	22.70	0.000	
Po1	0.0504	0.0109	4.62	0.000	2.63
Wealth	-0.000043	0.000034	-1.29	0.205	2.63

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.135473	44.92%	42.42%	34.85%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	0.65869	0.32934	17.95	0.000
Po1	1	0.39185	0.39185	21.35	0.000
Wealth	1	0.03036	0.03036	1.65	0.205
Error	44	0.80753	0.01835		
Total	46	1.46622			

Fits and Diagnostics for Unusual Observations

Obs	Log(Crime)	Fit	Resid	Std Resid	
27	2.5340	2.8233	-0.2893	-2.21	R
29	3.0180	3.2803	-0.2623	-2.22	R X
46	2.7060	2.9972	-0.2912	-2.19	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.29841

Figure 8. Second Regression Model using Variables X_1 and X_3

Based on the initial observations of the second regression analysis model, we can see that the VIF values are less than 5, indicating that multicollinearity likely does not exist between the two variables. Unfortunately, when the two variables, X_2 and X_4, were removed, the R-sq value and the Adj. R-sq values were not entirely preserved. In this model, the R-values decreased from the original 48.7% and 43.83% to 44.92% and 42.42%, which means that again, this model may not be the best fit for this particular dataset.

Although the R-values were not maintained, we can still see that the p-values coefficients remain low. The F-value has also increased indicating we are on the right track to optimizing our model, but we are not quite there yet.

There is little concern however over the Durbin-Watson statistic as it remains within a range of 1.5 and 2.5 coming in at 2.29.

In **Figure 9**, we see the residuals plots for the second regression model. We can see from the plots below that although the R-values were not preserved, the bandwidth of the versus fits plot remains consistent representing no alarming heteroscedasticity. We also see that the versus order plots remains chaotic, supporting the independence of residuals assumption.

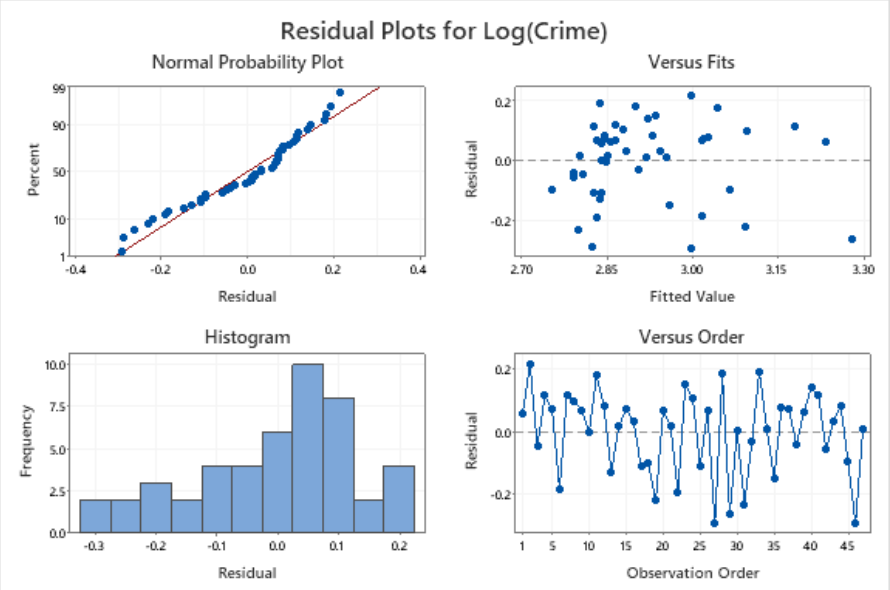


Figure 9. Residual Plots for Second Regression Model

Figure 10 displays the probability plot of the residuals. As you can see compared to the first regression probability plot, the Anderson-Darling statistic has decreased from 1.322 to 0.934. We also see that the p-value is greater than 0.250 and the data is starting to follow a more linear pattern. An outlier however, still remains.

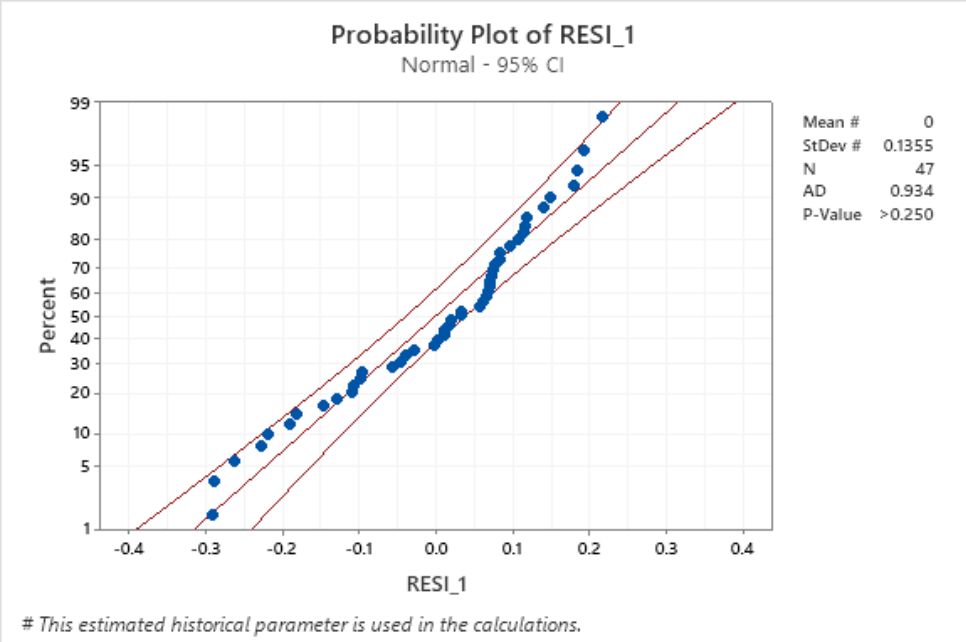


Figure 10. Second Regression Model Probability Plot

Generally, the second regression model is slightly better than the first, but there is still room for this model to improve. Although eliminating variables removed the issue of multicollinearity, we must now work on increasing the R-values. In order to do this, we will add additional independent variables to the model.

THIRD REGRESSION MODEL

From our second regression model, we added two explanatory variables in order to increase the R-values along with preventing multicollinearity. The variables added were as follows: Labor force participation rate of civilian urban male in the age-group 14-24 (X_9) and percentage of males aged 14-24 in total state population (X_{13}). A summary of this model is shown below in **Figure 11.**

WORKSHEET 1

Regression Analysis: Log(Crime) versus Po1, Wealth, LF, M

Regression Equation

Log(Crime) = 1.438 + 0.0515 Po1 - 0.000007 Wealth + 0.621 LF + 0.0530 M

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.438	0.440	3.27	0.002	
Po1	0.0515	0.0103	5.00	0.000	2.73
Wealth	-0.000007	0.000038	-0.19	0.847	3.94
LF	0.621	0.489	1.27	0.212	1.14
M	0.0530	0.0199	2.66	0.011	1.83

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.125649	54.78%	50.47%	42.60%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.80313	0.200784	12.72	0.000
Po1	1	0.39450	0.394501	24.99	0.000
Wealth	1	0.00059	0.000593	0.04	0.847
LF	1	0.02541	0.025413	1.61	0.212
M	1	0.11193	0.111929	7.09	0.011
Error	42	0.66308	0.015788		
Total	46	1.46622			

Fits and Diagnostics for Unusual Observations

Obs	Log(Crime)	Fit	Resid	Std Resid	
27	2.5340	2.8032	-0.2692	-2.24	R
37	2.9200	3.0435	-0.1235	-1.23	X
46	2.7060	2.9806	-0.2746	-2.26	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.98185

Figure 11. Third Regression Model using Variables X_1, X_3, X_9, and X_{13}.

The main objective of creating this third model was to increase the R-values and that was successfully completed. Compared to the first and second models, the R-square value increased to 54.78% and the adjusted R-square value increased to 50.47%. Although the R-values are not as high, the increase in value indicates this model is superior to the second regression model.

We also see that even though two explanatory variables were added, all VIF values still remain less than 5, showing that multicollinearity is not an issue here. We also observe that the p-values still remain relatively low with the Durbin-Watson statistic also decreasing from 2.30 to 1.98.

Figure 12 shows a four-in-one plot of the residuals for this model. Compared to the second model, the plots below support normality much more efficiently as the histogram gains a more bell-shaped curve (although slightly skewed to the right), the versus order plot remains chaotic, and the normal probability plot (**Figure 13)** follows a linear pattern.

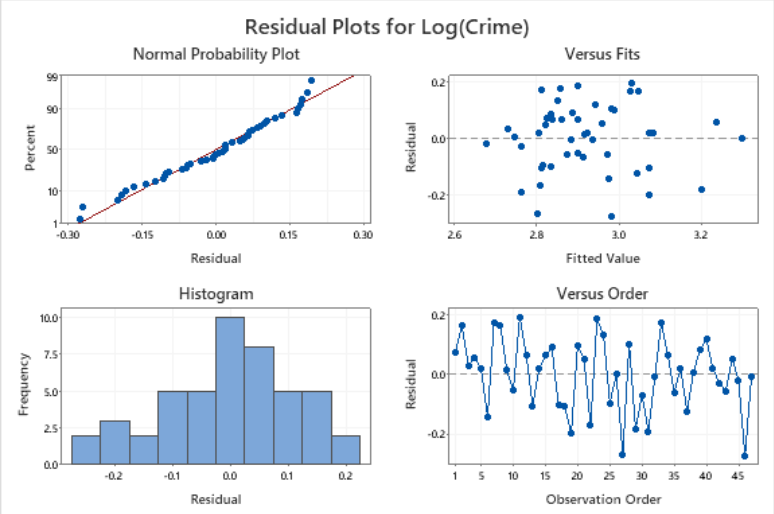


Figure 12. Four-in-One Plot of Residuals for Third Regression Model

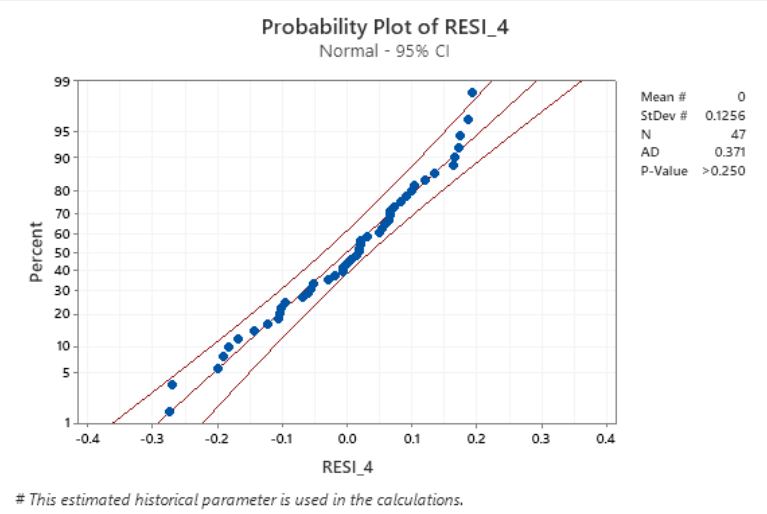


Figure 13. Probability Plot for Third Regression Model

It is also important to note that the probability plot also shows a decreased Anderson-Darling statistic, going from the original 0.934 to 0.371. P-values also remain greater than 0.250.

This regression model is described by the following equation:

Regression Equation

Log(Crime) = 1.438 + 0.0515 Po1 - 0.000007 Wealth + 0.621 LF + 0.0530 M

Equation 1. Third Regression Model Equation

INTERACTION TERM TEST

To further test our regression model, several interaction terms were tested in Minitab. The interaction term that produced the best improvement on our regression model was the product of X_7, unemployment rate of urban males 14-24 and X_{12}, income inequality: percentage of families earning below half the median income. The results are summarized in **Figure 14.**

WORKSHEET 1

Regression Analysis: Log(Crime) versus Po1, Wealth, LF, M, ineq*u2

Regression Equation

Log(Crime) = 0.560 + 0.04762 Po1 + 0.000037 Wealth + 1.241 LF + 0.0628 M + 0.00298 ineq*u2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.560	0.511	1.10	0.280	
Po1	0.04762	0.00964	4.94	0.000	2.79
Wealth	0.000037	0.000039	0.95	0.348	4.70
LF	1.241	0.502	2.47	0.018	1.40
M	0.0628	0.0187	3.35	0.002	1.89
ineq*u2	0.00298	0.00105	2.84	0.007	1.77

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.116237	62.22%	57.61%	47.35%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	0.91226	0.18245	13.50	0.000
Po1	1	0.33007	0.33007	24.43	0.000
Wealth	1	0.01219	0.01219	0.90	0.348
LF	1	0.08238	0.08238	6.10	0.018
M	1	0.15185	0.15185	11.24	0.002
ineq*u2	1	0.10912	0.10912	8.08	0.007
Error	41	0.55396	0.01351		
Total	46	1.46622			

Fits and Diagnostics for Unusual Observations

Obs	Log(Crime)	Fit	Resid	Std Resid	
37	2.9200	3.0917	-0.1717	-1.89	X
46	2.7060	2.9440	-0.2380	-2.13	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.71609

Figure 14. Interaction Term Model using Variables X_1, X_3, X_9, X_{13}, and X_7*X_{12}

The above interaction term was selected as it showed the highest increase in the R-values while also maintaining low VIF values and coefficient p-values for the explanatory variables remained low. When testing other interaction terms, R-values did increase, but VIF values also drastically increased between explanatory variables.

Along with the observations mentioned previously, other observations seen were:

- The R-square and Adj R-Square values increased from 54.4% and 50.47% to 62.2% and 57.61% respectfully. Values clearly increased for both values from the third regression model.
- VIF values all remain less than 5 indicating no concern over multicollinearity
- The Durbin-Watson statistic decreased slightly, but still remains close to a target value of 2
- The coefficient p-values of independent variables carried over from the previous model remain small. The coefficient p-value for wealth also decreased using the interaction term.

Plots of the residuals (**Figure 15 and Figure 16)** were also produced in order to analyze the interaction term against the third regression model. Based on the residual plots below, the following observations were made:

- Normality becomes a better assumption with the interaction term added as the histogram becomes more evenly distributed (not as skewed) and creates the bell-shaped curve
- Plots of the residuals over fitted values show constant bandwidth indicating no apparent heteroscedasticity. Constant variance of residuals may be assumed.
- The versus order plot remains chaotic
- The probability plot of residuals remains in bounds of the confidence interval with no outliers, a low Anderson-Darling value of 0.161 (decreasing from the third regression model), and p-values remain greater than 0.250

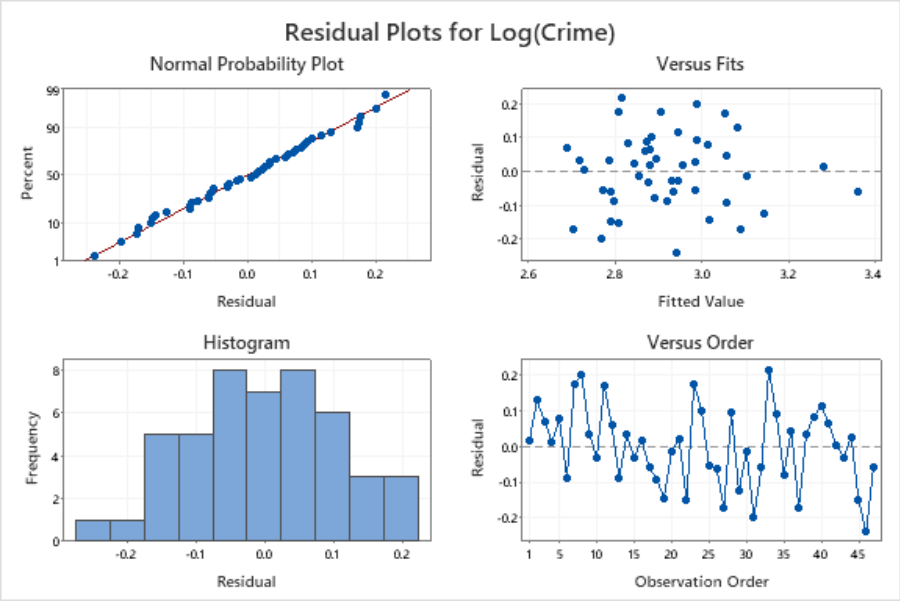


Figure 15. Four-in-One Plot of Residuals for Interaction Term Model

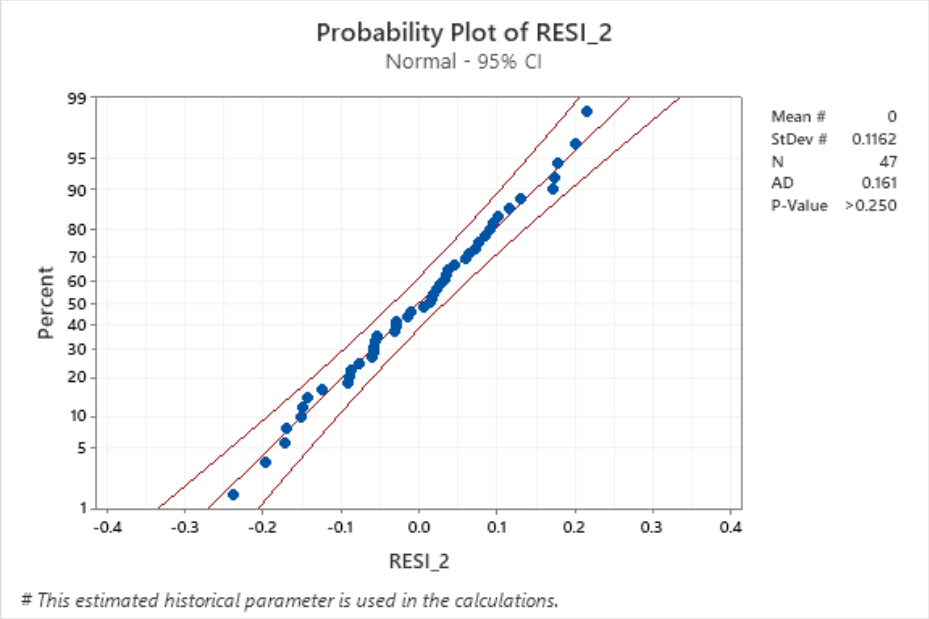


Figure 16. Probability Plot for Interaction Term Model

Overall, adding an interaction term ultimately improved our regression model by increasing the R-value terms by 7.8% (R-square) and 4.6% (Adj. R-Square). The equation that describes this regression model is below.

Regression Equation

Log(Crime) = 0.560 + 0.04762 Po1 + 0.000037 Wealth + 1.241 LF + 0.0628 M + 0.00298 ineq*u2

Equation 2. Interaction Term Test Model Equation

Although initial observations show that this regression model may be the best model, this will be further explored under the "Best Regression Model" section of the report.

DIAGNOSTIC DATA ANALYSIS

In order to determine any outliers or influential data points in the given data set, an analysis was conducted in Minitab and Excel to flag such observations.

This analysis was conducted with the third regression model and the interaction term model as they are the models being considered for our final regression model. To do this, the studentized (deleted) residuals and DFIT coefficients were found. We then exported this data into Excel in order to find the influential datapoints through conditional formatting and calculated thresholds.

Through this method, the following observations were made (**Figure 17)**:

- The studentized deleted residual threshold with a significance level of 5% was calculated to be ~2.02 for both the third regression model and the interaction term model. With this threshold, data points 27 and 46 were found to be outliers. Data point 46 was an outlier in both models. These datapoints should be reconsidered in the data set.
- The DFIT threshold was calculated to be 0.65 and 0.71 for each model. With this threshold data points 8, 24, 27, 29, 37, and 45 were found to be outliers. Data points 8,27,29, and 37 were found in both models while data point 45 was an outlier only in the interaction term model.

Unfortunately, with no other information of how the data was acquired, we can identify these data points, but it would not be enough information to remove them from each model.

Third Regression Model							Interaction Term Model					
Data	TRES	DFIT					Data	TRES_1	DFIT_1			
1	0.59875	0.17665		p	4		1	0.15351	0.05354		p	5
2	1.36838	0.30989		n	47		2	1.15618	0.29426		n	47
3	0.27497	0.13087		DFIT Threshold	0.6523281		3	0.68678	0.34297		DFIT Threshold	0.7145896
4	0.49843	0.2073					4	0.12756	0.05685			
5	0.15596	0.03856					5	0.69501	0.217		α	0.05
6	-1.20846	-0.42274		α	0.05		6	-0.80898	-0.32171		n-p-2	40
7	1.51896	0.63453		n-p-2	41		7	1.6619	0.69427		TRES1 Threshold	2.0210754
8	1.48139	0.72279		TRES1 Threshold	2.019541		8	2.01118	1.01914			
9	0.11516	0.03159					9	0.3215	0.09122			
10	-0.43102	-0.14249					10	-0.27613	-0.09338			
11	1.62164	0.43011					11	1.55529	0.42701			
12	0.54973	0.14861					12	0.53535	0.14518			
13	-0.88536	-0.34028					13	-0.81298	-0.31558			
14	0.14806	0.04037					14	0.29917	0.08299			
15	0.52907	0.1354					15	-0.26371	-0.10724			
16	0.75696	0.25946					16	0.17597	0.07445			
17	-0.82503	-0.16771					17	-0.51275	-0.12616			
18	-0.88327	-0.26328					18	-0.80799	-0.24477			
19	-1.68435	-0.50095					19	-1.31114	-0.46103			
20	0.8053	0.18852					20	-0.14026	-0.0637			
21	0.40971	0.12369					21	0.20906	0.0657			
22	-1.49495	-0.70401					22	-1.43512	-0.68318			
23	1.53725	0.31144					23	1.58588	0.3244			
24	1.09693	0.22149					24	0.89249	0.20342			
25	-0.83009	-0.37957					25	-0.50691	-0.24457			
26	0.00663	0.00374					26	-0.60416	-0.37275			
27	-2.35982	-0.72466					27	-1.64642	-0.76074			
28	0.85409	0.24684					28	0.84907	0.24649			
29	-1.76282	-1.10819					29	-1.30836	-0.88562			
30	-0.58224	-0.23156					30	-0.10451	-0.04674			
31	-1.5894	-0.38537					31	-1.79617	-0.43732			
32	-0.04889	-0.01317					32	-0.52526	-0.1683			
33	1.42901	0.27615					33	1.97336	0.46189			
34	0.53447	0.12809					34	0.79767	0.20115			
35	-0.48991	-0.14571					35	-0.68867	-0.20868			
36	0.16922	0.05771					36	0.41363	0.14552			
37	-1.24143	-0.94077					37	-1.94894	-1.54556			
38	0.05327	0.02135					38	0.3197	0.13209			
39	0.68972	0.19773					39	0.75594	0.21673			
40	0.96684	0.16882					40	1.0082	0.17657			
41	0.17466	0.07353					41	0.59803	0.26858			
42	-0.23751	-0.0686					42	0.0544	0.01693			
43	-0.48311	-0.21099					43	-0.27445	-0.12276			
44	0.43777	0.11528					44	0.2446	0.06764			
45	-0.15988	-0.07208					45	-1.61252	-1.12875			
46	-2.37737	-0.60776					46	-2.22735	-0.62804			
47	-0.04613	-0.01345					47	-0.49641	-0.16624			

Figure 17. Diagnostic Analysis for Outliers on Third Regression Model and Interaction Term Model

BEST REGRESSION MODEL

In order to appropriately select the best model for this regression report, increases in the R-value for the third regression model to the interaction term test model were tested for statistically significance.

Figure 18 summarizes the analysis for the best regression model using data acquired from Minitab and Excel.

With this analysis one can say that the interaction term test model is indeed the better choice for this data set. Here, we have observed that the F-statistic value is greater than the F-critical value at a significance value of 5%. We can also see that the p-value (1%) is less than the significance level of 5% also supporting that the interaction term test model is the better regression model.

Overall, adding an interaction term defined as the product of X_7, unemployment rate of urban males 14-24 and X_{12}, income inequality: percentage of families earning below half the median income can lead one to say that we can accept this model over the third regression model.

Interaction Term Test Model	
R Square	62.20%
Deg Freedom	41

Third Regression Model	
R Square	54.78%
Deg Freedom	42

	Value	Df
Numerator	0.074	1
Denominator	0.009	41
F-statistic	8.048	
a	5%	

Critical Value	4.079
Conclusion	Model Improvement
p-value	1%
Conclusion	Model Improvement

Figure 18. Best Regression Model Analysis

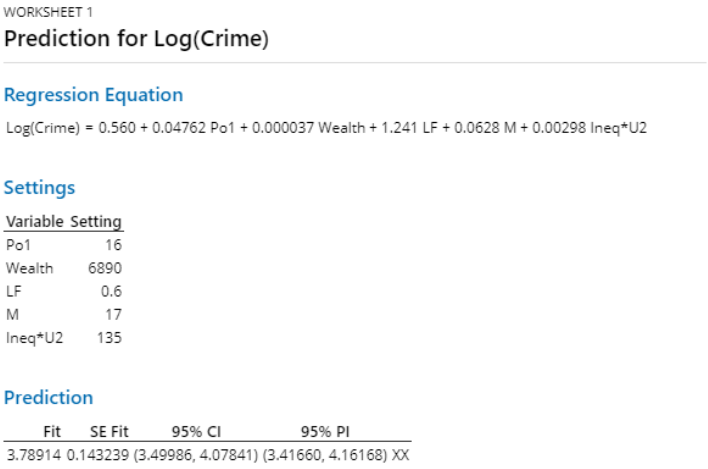
FORECASTING

Using the interaction term regression model, we were also tasked to forecast the total crime rate using the given variables below **(Table 2)**. It is important to note that the only the variables used within the interaction term regression model were used while forecasting.

Po1	Wealth	LF	M	Ineq*U2
16	6890	0.6	17	153

Table 2. Given Variables for Prediction

Minitab was used in order to predict the coefficients of each independent variable using a 95% confidence interval **(Figure 19)**. **Table 3** displays these coefficients along with b-hat values needed to perform forecasting.



WORKSHEET 1

Prediction for Log(Crime)

Regression Equation

Log(Crime) = 0.560 + 0.04762 Po1 + 0.000037 Wealth + 1.241 LF + 0.0628 M + 0.00298 Ineq*U2

Settings

Variable	Setting
Po1	16
Wealth	6890
LF	0.6
M	17
Ineq*U2	135

Prediction

Fit	SE Fit	95% CI	95% PI
3.78914	0.143239	(3.49986, 4.07841)	(3.41660, 4.16168) XX

Figure 19. Predication Analysis in Minitab

	x0	b-hat
Intercept	1	5.60E-01
Po1	16	4.76E-02
Wealth	6890	3.70E-05
LF	0.6	1.24E+00
M	17	6.28E-02
Ineq*U2	135	2.98E-03

Table 3. Variable Coefficients & B-hat Values

This information was then exported from Minitab to Excel in order to calculate bounds for the 95% confidence intervals and the 95% predication intervals. When calculated, the forecasted dependent variable (Crime Rate) was revealed. Summary of this data is listed below in **Figure 20.**

	x0	b-hat
Intercept	1	5.60E-01
Po1	16	4.76E-02
Wealth	6890	3.70E-05
LF	0.6	1.24E+00
M	17	6.28E-02
Ineq*U2	135	2.98E-03

PFITS	PSEFITS	CLIM	CLIM_1	PLIM	PLIM_1
3.78914	0.143239	3.49986	4.07841	3.4166	4.16168

Log(Crime Rate) - hat	3.78914
Median[Crime Rate]	6153.752
E[Crime Rate]	6734.678

Standard Error Residuals	0.116237
Var[Log(Crime Rate)]	0.034029
Standard Deviation [Log(Crime I	0.184468

95% Confidence Interval	
LB E[Log(Crime Rate)]	3.49986
UB E[Log(Crime Rate)]	4.07841

95% Prediction Interval	
LB Log(Crime Rate)	3.4166
UB Log(Crime Rate)	4.16168

Approximate 95% Confidence Interval	
LB E[Crime Rate]	3161.258
UB E[Crime Rate]	11978.71

95% Prediction Interval	
LB Crime Rate	2609.756572
UB Crime Rate	14510.42056

Figure 20. Forecasted Data Summary using Interaction Term Model

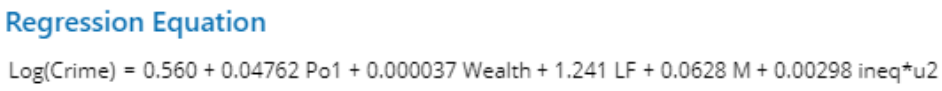
From the parameters given in **Table 2,** the best regression model produces a crime rate of ~6735. The approximate 95% confidence interval of the expected value is between 3161.258 for the lower bound and 11,978.71 for the upper bound. This confidence interval, however, only specifies the range at which crime rate could fall and has no probability interpretation.

The 95% predication interval for crime rate was found to be 2609.756 for the lower bound and 14,510.42 for the upper bound. This means that crime rate has a 95% chance of following within this range given the explanatory variables for prediction.

CONCLUSIONS

Based on the forecasted values for crime rate, one can see that intervals for prediction are rather large. Using any other model, however, would have resulted in much larger intervals. Overall, it was shown that the interaction term model was indeed the best regression model after studying a model of highly correlated explanatory variables, removed variables, and interaction term variables.

Based on all procedural analysis performed, the final model yielded the following equation:



Equation 2. Interaction Term Test Model Equation

With this model, the highest R-values were produced and statistically proven to be significant. It was also determined by observation of the residual plots that normality and independence assumptions could be made.

One can also say that based on the explanatory values used within this model, these predictors can be used to best determine total crime rate in a given state for a specific year. The explanatory variables used were per capita expenditure in police protection in 1960 (X_1), wealth: median value of transferrable assets or family income (X_3), labor force participation rate of civilian urban male in the age-group 14-24 (X_9) and percentage of males aged 14-24 in total state population (X_{13}), and finally the product of unemployment rate of urban males 14-24 (X_7) and income inequality: percentage of families earning below half the median income (X_{12}).

Overall, these explanatory variables helped to build the best regression model in order to determine punishment regimes on city crime rate.