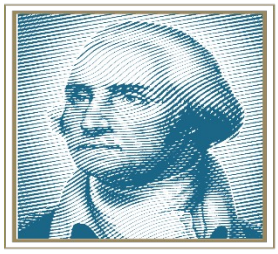


THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

Regression Analysis Report

Name: Shrishail Ravi Terni

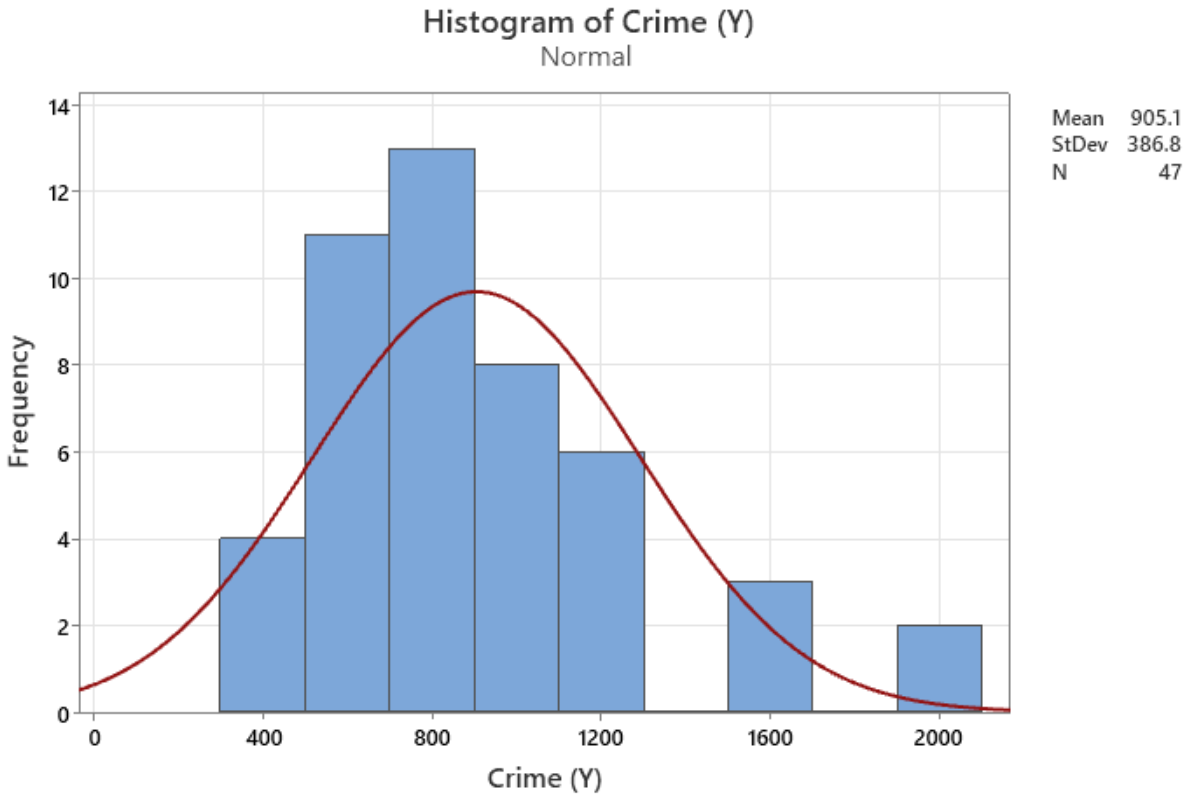
GWID: G28972385

Professor: Dr. Johan Rene Van Dorp

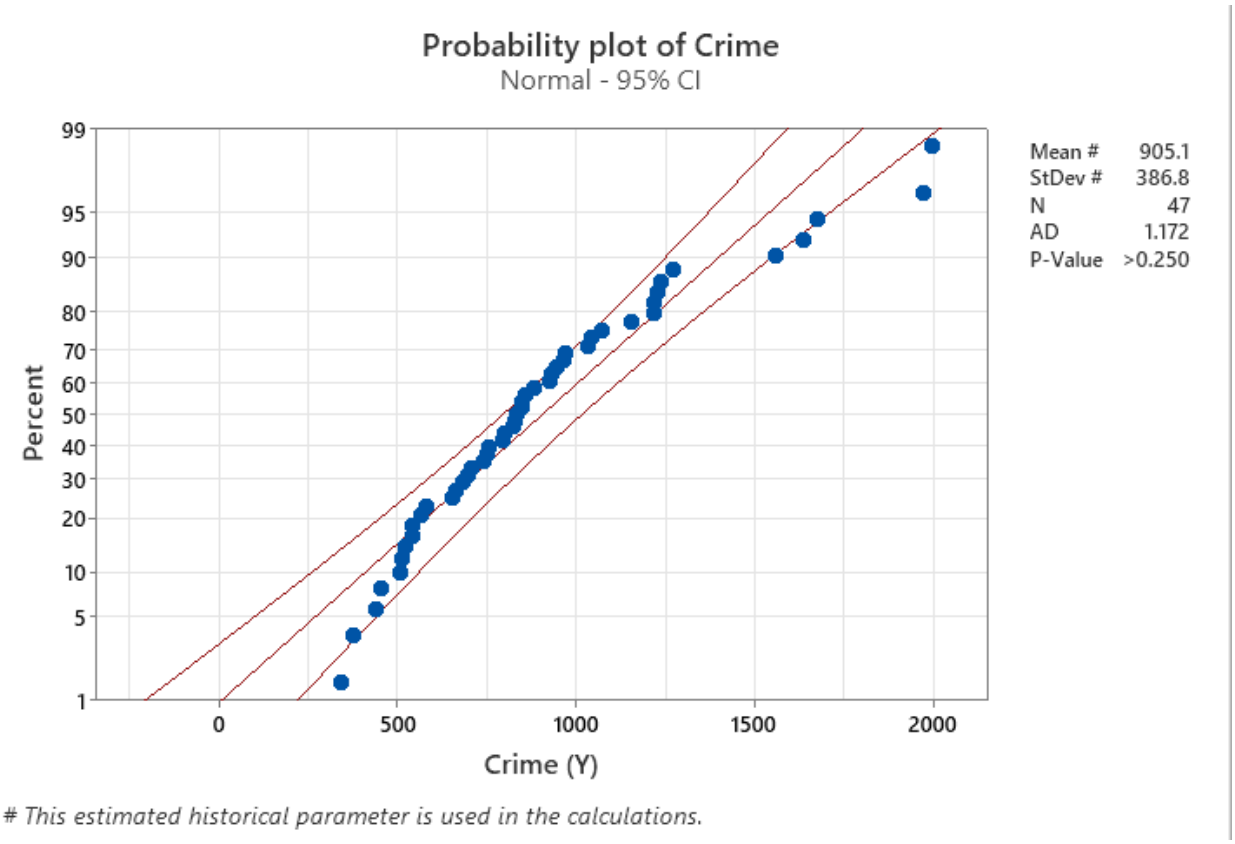
Course code: EMSE 6765

Table of Contents

Analysis of Dependent Variable



The histogram plot of the dependent variable (Crime Data) the plot is not symmetric. It has a mean of 905.1 and standard deviation of 386.8. The plot of the dependent variable is left-skewed, this can be one of the reasons to reject.

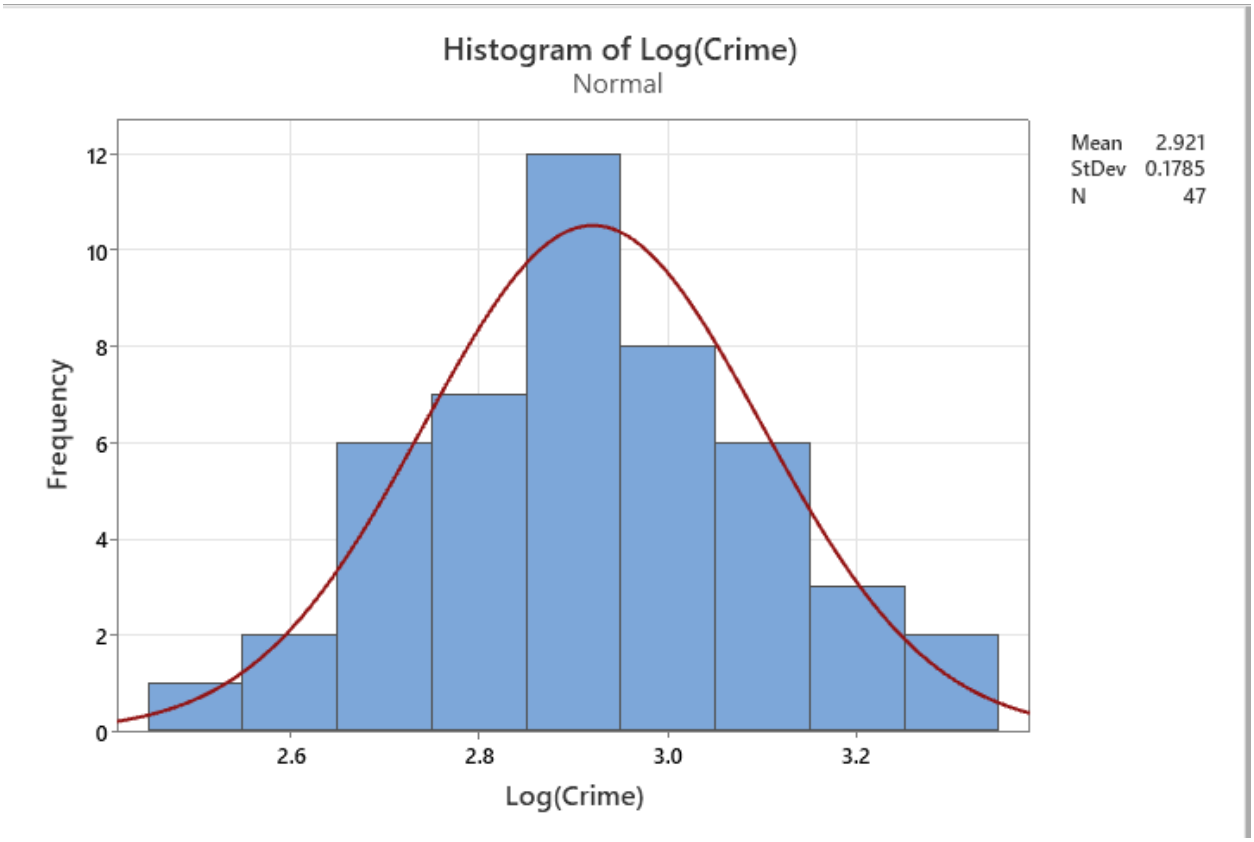


 # This estimated historical parameter is used in the calculations.

The probability plot of the dependent variable has a p-value greater than 0.25 which and Anderson-Darling statistic of 1.172 (which is a high value and deviating from normality). The other supporting factors to reject the choice of our dependent variable are: -

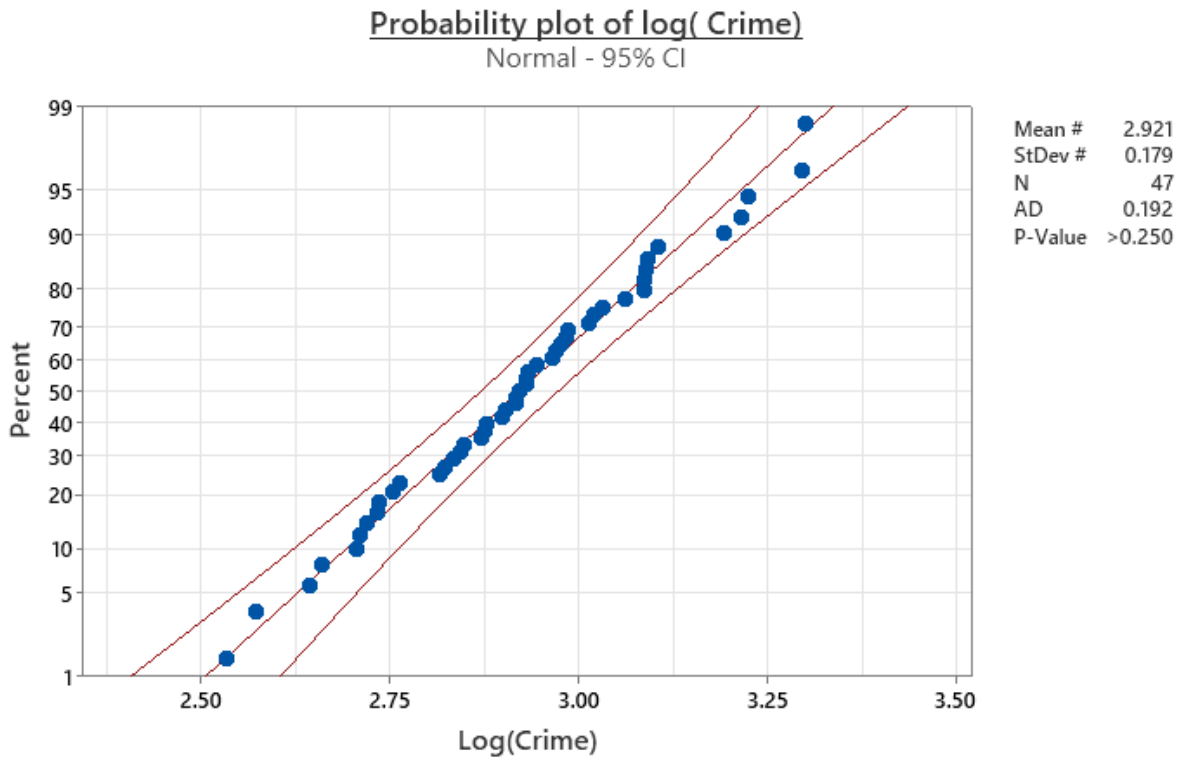
- The plot is not symmetric with the mean.
- There are far too many outliers.
- Many data points on the boundaries.

The reasons listed above are sufficient to reject the use of the price as the dependent variable. Consider taking the log(price) as the dependent variable.



The histogram of the log of crime variable is comparatively normal than the price variable.

Check for the probability plot.



 # This estimated historical parameter is used in the calculations.

The inferences drawn from the probability plot of the log(crime) variable are listed below.

Key Takeaways: -

- P-value greater than 0.250.
- Anderson Darling statistic of 0.192 which is a low value showing signs of normality.
- It has a mean of 2.921 and a standard deviation 0.179.
- The data points are symmetric across the mean.
- There are no visible outliers.

Hence the log(price) is opted satisfactorily as the dependent variable.

The next step is to do the correlation analysis of the dependent variable with the other independent variable to find out which independent variables contribute the most to dependent variable.

Correlation Analysis

The threshold value of 0.4 is used.

	Log(Crime)	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
Log(Crime	1													
Po1	0.654631	1												
Po2	0.637305	0.993586	1											
Wealth	0.42662	0.787225	0.794262	1										
Prob	-0.41189	-0.47325	-0.47303	-0.55533	1									
Pop	0.337359	0.526284	0.513789	0.308263	-0.34729	1								
Ed	0.302145	0.482952	0.49941	0.735997	-0.38992	-0.01723	1							
U1	-0.07487	-0.0437	-0.05171	0.044857	-0.00747	-0.03812	0.018103	1						
U2	0.167404	0.185093	0.169224	0.092072	-0.06159	0.270422	-0.21568	0.745925	1					
LF	0.172732	0.121493	0.10635	0.294632	-0.25009	-0.12367	0.561178	-0.2294	-0.42076	1				
M.F	0.148161	0.03376	0.022843	0.179609	-0.05086	-0.41063	0.436915	0.351892	-0.01869	0.513559	1			
Ineq	-0.15169	-0.6305	-0.64815	-0.884	0.465322	-0.12629	-0.76866	-0.06383	0.015678	-0.26989	-0.16709	1		
Time	0.142578	0.103358	0.075627	0.000649	-0.43625	0.46421	-0.25397	-0.16985	0.101358	-0.12364	-0.4277	0.101823	1	
M	-0.05623	-0.50574	-0.51317	-0.67006	0.361116	-0.28064	-0.53024	-0.22438	-0.24484	-0.16095	-0.02868	0.639211	0.114511	1
	Threshold		0.4											

Based on the correlation matrix above we can see that the dependent variables (PO1,PO2,Wealth,Prob) are most highly correlated with log(price).

Let us do a regression analysis in Excel.

Excel Regression Analysis(basic model)

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

ANOVA

	df	SS	MS	F	Significance F
Regression	4	0.714302014	0.178575504	9.977570596	9.09055E-06
Residual	42	0.75170314	0.017897694		
Total	46	1.466005154			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2.892264251	0.167897367	17.22638245	1.19098E-20	2.553433647	3.231094854	2.553433647	3.231094854
Po1	0.094633166	0.058779966	1.609956115	0.114898907	-0.023989608	0.21325594	-0.023989608	0.21325594
Po2	-0.049881187	0.063379336	-0.787026035	0.435685392	-0.177785866	0.078023491	-0.177785866	0.078023491
Wealth	-5.6972E-05	3.57479E-05	-1.593714881	0.118498791	-0.000129114	1.51702E-05	-0.000129114	1.51702E-05
Prob	-1.624611633	1.046317922	-1.552694069	0.127999942	-3.736166687	0.486943422	-3.736166687	0.486943422

From the regression statistics above it can be inferred that: -

- The model has a very low R-squared and a very low adjusted R-squared value (49% and 43.8% respectively).
- The p-value of the F-statistic is significantly very low which signifies that at least one of the coefficients of the regression model is non-zero.
- Now we examine the p-values of the Intercept and other dependent variables.
- The p-value of the Intercept is very low so we can conclude that there will be an intercept in the equation.
- The p-values of the Po1, Prob, wealth are considerably low also we can expect coefficients of the Po1 and wealth explanatory variables.
- Whereas the coefficients of Po2 are very high so we may discard these explanatory variables.

Minitab Regression Analysis(basic model)

Regression Equation:

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

Coefficients Analysis:

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.892	0.168	17.23	0.000	
Po1	0.0946	0.0588	1.61	0.115	78.43
Po2	-0.0499	0.0634	-0.79	0.436	80.72
Wealth	-0.000057	0.000036	-1.59	0.118	3.06
Prob	-1.62	1.05	-1.55	0.128	1.45

Key Takeaways: -

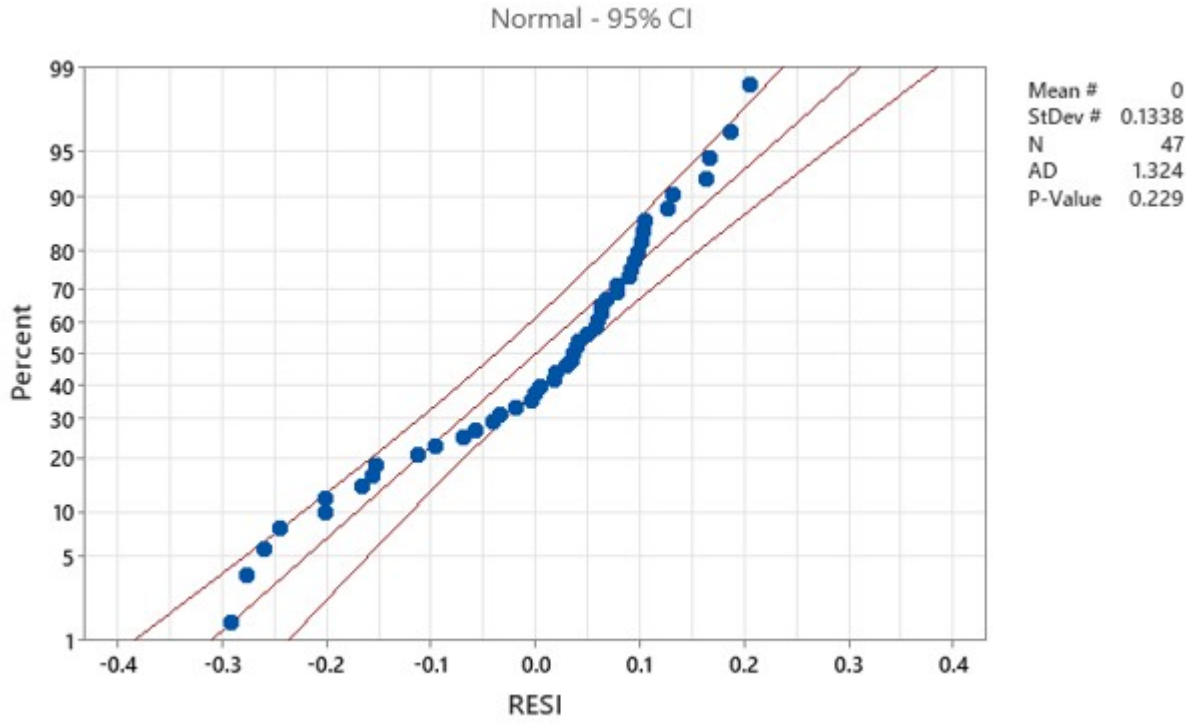
- High Variable Inflation factors which indicate Multicollinearity among the dependent variables.

Durbin-Watson Statistic:

Durbin-Watson Statistic = 2.34368

The Durbin-Watson statistic is 2.3 which is not that ideal but not a sole factor to reject the model.

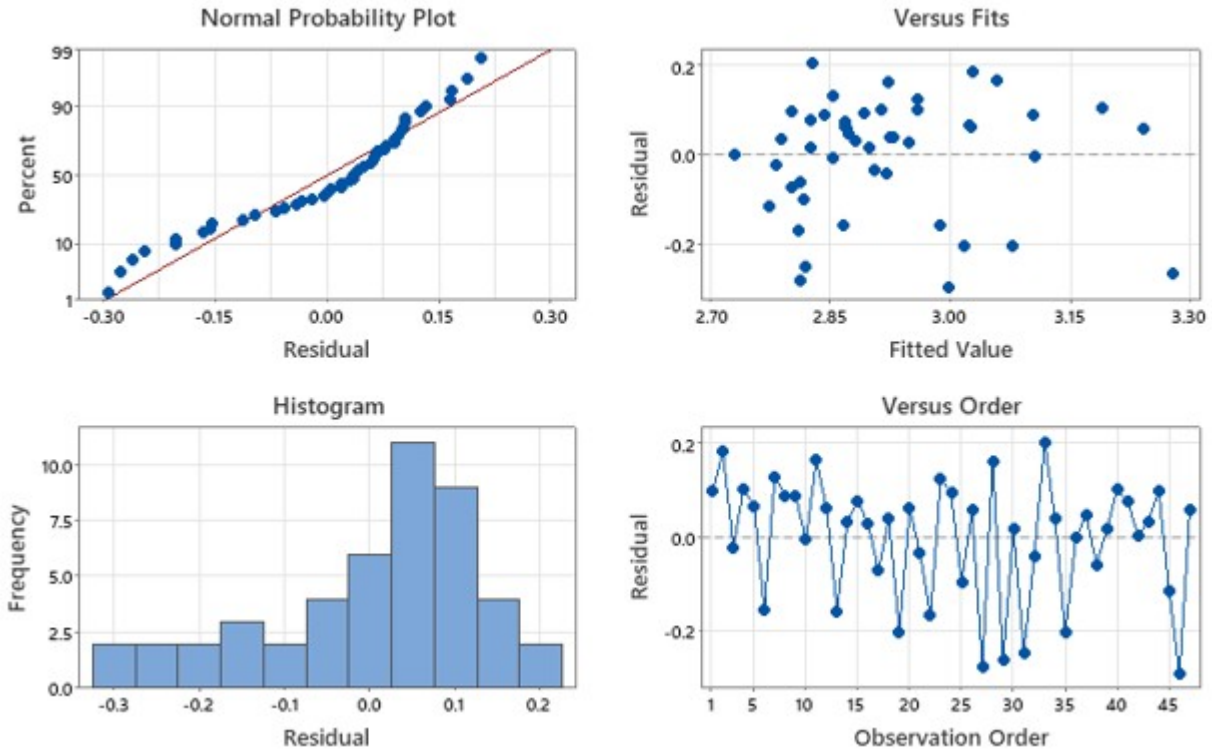
Probability plot of Residuals



 # This estimated historical parameter is used in the calculations.

The probability plot of the residuals is normal but has 1 outlier and is not symmetric with the mean.

Residual Analysis



Here are a few takeaways from the Residual plot: -

- In the Residual plot vs Fits plot that there is Heteroscedasticity i.e. (the residuals are not a constant function of the variance)
- In the residual vs Frequency plot, we can see that it is rightly skewed.
- In the Residual plot vs Observation order plot, we can see a chaotic function which is ideal.

Since the model used above has very high variable inflation factors for all the independent variables, we omit all of them and use a different set of explanatory variables.

Model-2

Regression Equation

Log(Crime) = 0.101 + 0.1056 Ed + 0.0870 U2 + 0.0530 M + 0.03043 Ineq - 1.736 Prob
+ 0.04572 Po1 - 2.33 U1

Coefficient Analysis

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.101	0.434	0.23	0.818	
Ed	0.1056	0.0234	4.52	0.000	3.37
U2	0.0870	0.0348	2.50	0.017	4.25
M	0.0530	0.0160	3.30	0.002	2.00
Ineq	0.03043	0.00673	4.52	0.000	3.55
Prob	-1.736	0.736	-2.36	0.023	1.38
Po1	0.04572	0.00760	6.01	0.000	2.52
U1	-2.33	1.46	-1.59	0.119	3.42

Key Takeaways: -

- The U1 explanatory variable has a comparatively higher P-value.
- The Variable Inflation factor of U2 explanatory variable is high.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0966021	75.17%	70.72%	61.88%

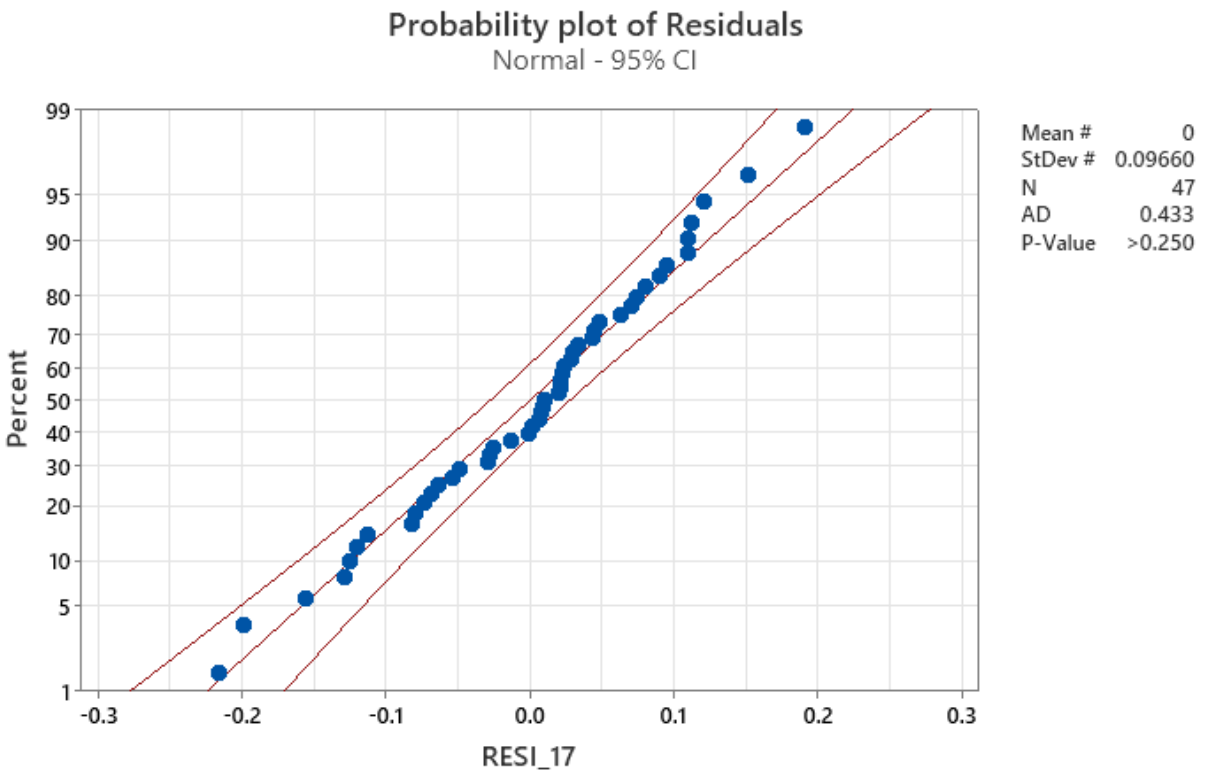
The R-squared value is 75.17% and the adjusted R-squared value is 70.72%.

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.10404

The Durbin-Watson statistic is 2.104 which is ideal

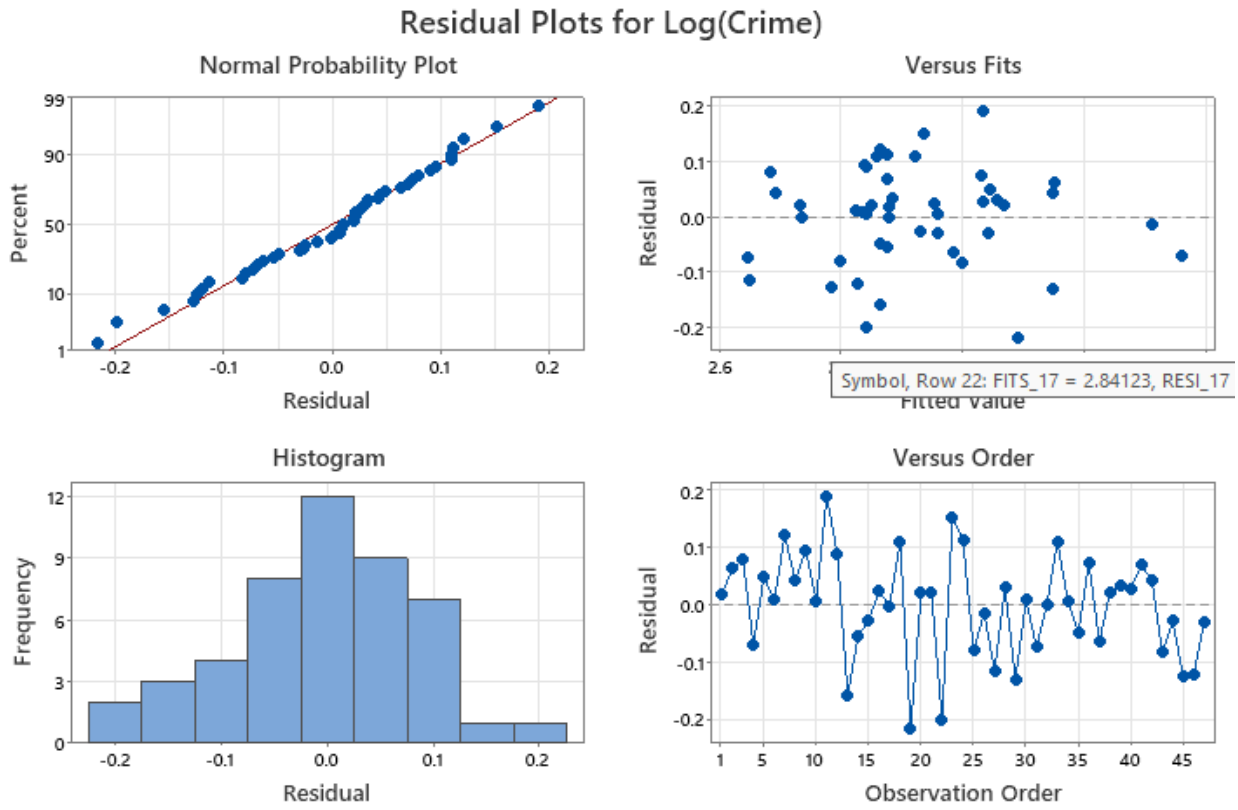
Probability plot of Residuals



The probability plot of Residuals is perfectly normal due to the following reasons:

- The Anderson-Darling statistic is very low and ideal.
- The p-value is greater than 25%.
- No outliers and symmetric with the mean.

Residual Analysis



Key Takeaways: -

- In the residual vs the Fitted value plot there is no heteroscedasticity.
- The residual vs Frequency plot is normal.
- The residual vs the observation order plot is chaotic.

The next model to be considered will include an interaction term to increase the adjusted R-squared value. The interaction term will be Ineq and Po1 as they have a lowest negative correlation (-0.68).

Model-3

Regression Equation

Regression Equation

Log(Crime) = 1.029 - 0.000582 Pop + 0.0819 Ed + 0.0911 U2 + 0.0496 M - 0.0078 Ineq - 2.43 U1 - 1.531 Prob - 0.0371 Po1 + 0.00522 Ineq*Po1

Coefficient Analysis

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.029	0.474	2.17	0.037	
Pop	-0.000582	0.000453	-1.29	0.207	1.83
Ed	0.0819	0.0221	3.71	0.001	3.76
U2	0.0911	0.0313	2.91	0.006	4.30
M	0.0496	0.0146	3.39	0.002	2.09
Ineq	-0.0078	0.0136	-0.57	0.570	18.25
U1	-2.43	1.32	-1.84	0.073	3.48
Prob	-1.531	0.688	-2.23	0.032	1.51
Po1	-0.0371	0.0277	-1.34	0.189	41.80
Ineq*Po1	0.00522	0.00161	3.24	0.003	25.76

Key Takeaways: -

- The p-value of the pop, ineq and po1 variable is high.
- The addition of an Interaction term(Ineq*po1) leads to high variable Inflation factors among the explanatory variables

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0864224	81.15%	76.56%	69.89%

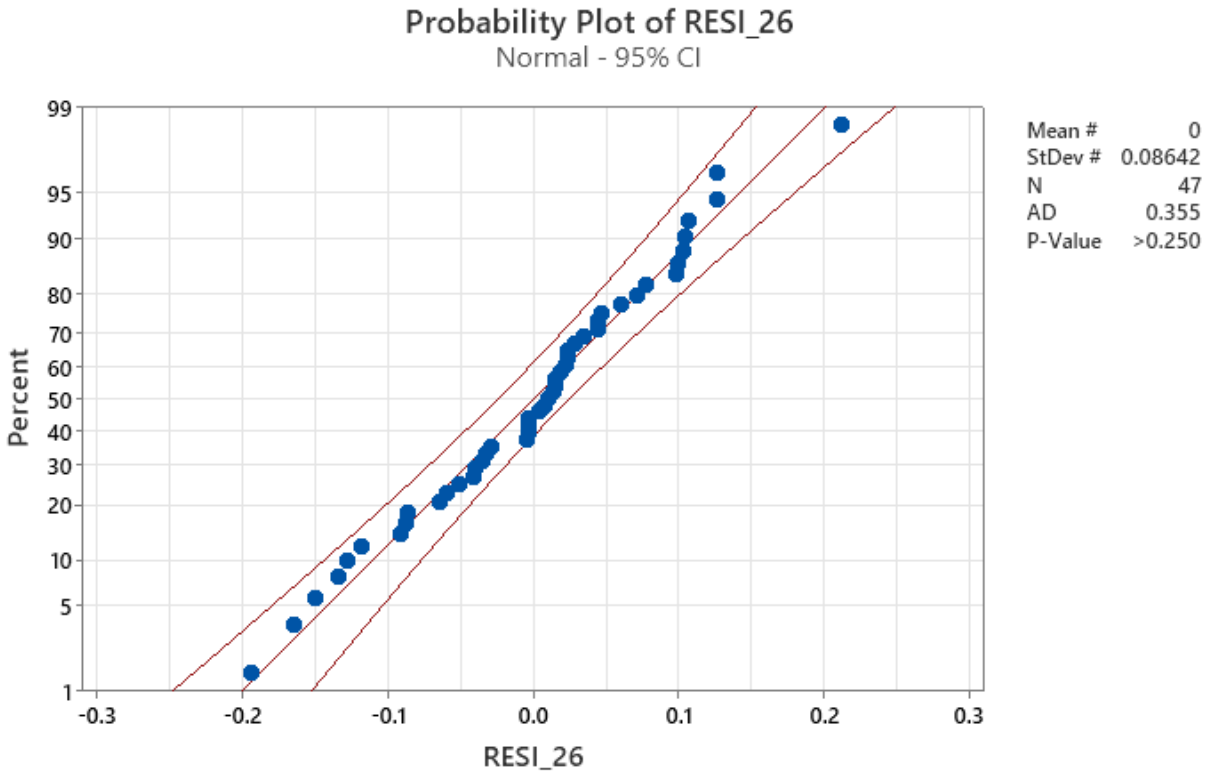
The R-squared value has increased up to 81.15% and increase in the adjusted R-squared value up to 76.56% which can be clearly attributed to the addition of the Interaction term.

Durbin-Watson Statistic

Durbin-Watson Statistic = 1.96244

The Durbin-Watson statistic is 1.96244 which is close to 2 and is very ideal.

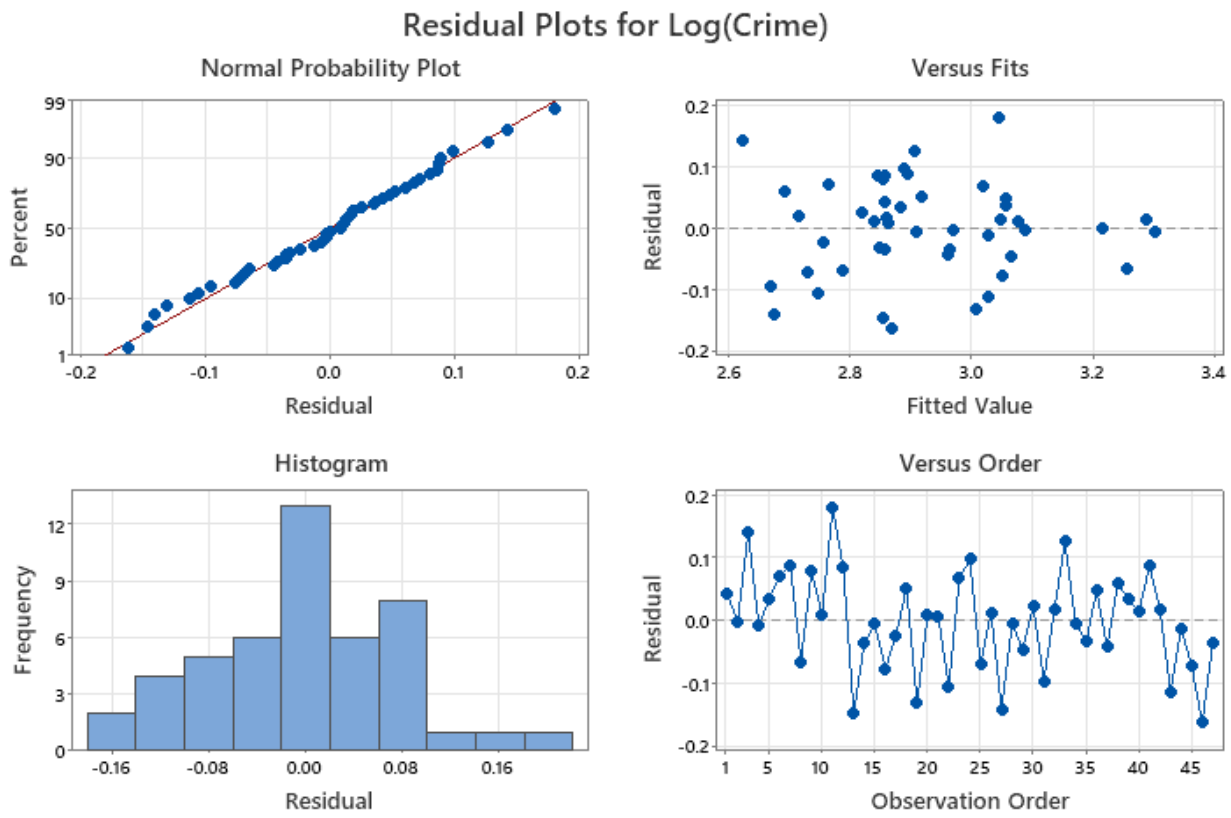
Probability plot of Residuals



Key Takeaways: -

- The plot has a p-value greater than 25% and an Anderson-Darling statistic of 0.355 which is very ideal.
- The plot is symmetrical with the mean and there are no visible outliers

Residual Analysis



Key Takeaways: -

- The Residual vs the Fitted value is Homoscedastic.
- The Histogram plot of the residuals deviates from normality.
- The Residual vs the Observation order plot is chaotic.

It is important to check whether the increase in the R-squared from Model-2 to Model-3 is statistically significant or not.

Hence a F-test is conducted to check for the improvement in the model.

Diagnostic Analysis of the model

	Full Model		
	R Square	81.15%	
	Degrees of Freedom (DF)	37	
	Small Model		
	R Square	75.17%	
	Degrees of Freedom	39	
	Difference R-Squared	5.98%	
	Difference Df	2	
		Value	
	Numerator	0.0299	
	Denominator	0.0051	
	F-Statistic	5.869	
	α	5%	
Method 1	**Critical Value**	3.252	**Conclusion**
	Conclusion	Reject H0	Model Improvement
Method 2	**p-value**	0.61%	**Conclusion**
	Conclusion	Reject H0	Model Improvement
	H_0:	No Model Improvement	
	H_1:	Model Improvement	

Explanatory Variables in the full model

Pop	Ed	U2	M	Ineq	U1	Prob	Po1	Ineq*Po1

Explanatory Variables in the restricted/small model

	Ed	U2	M	Ineq	Prob	Po1	U1

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

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

Details of the Test: -

- An F-test is conducted based on the R-squared values and the degrees of freedom of the larger model and the smaller model.
- Null-Hypothesis: - No model improvement
  Alternate Hypothesis: - Model Improvement
- The value of the F-statistic is 5.859 and the critical value is 3.252.
- Since the F-statistic is greater than the critical value we reject the null hypothesis.
- The p- value (0.61%) is lesser than the alpha value hence we reject the null hypothesis.

To assert that there is model improvement we need to predict with both the models.

Prediction by Models with the given input

Model-2

Settings

Variable	Setting
Ed	12
U2	5
M	17
Ineq	27
Prob	0.01
Po1	16
U1	0.14

Prediction

Fit	SE Fit	95% CI	95% PI	
3.91292	0.123677	(3.66276, 4.16309)	(3.59550, 4.23035)	XX

The width of the Prediction Interval is 0.63485

Model-3

Variable	Setting
Pop	168
Ed	12
U2	5
M	17
Ineq	27
U1	0.14
Prob	0.01
Po1	16

Prediction

Fit	SE Fit	95% CI	95% PI	
4.30689	0.177583	(3.94707, 4.66671)	(3.90672, 4.70705)	XX

Here the width of the prediction interval is 0.80033

Since model-2 has the lowest width of the prediction interval it is preferred.

Prediction of Model-2:

MINITAB OUTPUT

PFITS	PSEFITS	CLIM	CLIM_1	PLIM	PLIM_1
3.91292	0.123677	3.66276	4.16309	3.5955	4.23035

						Variances
μ = LOG(CRIME) - hat	3.913	3.913		Standard Error Residuals	0.096602	0.009332
MEDIAN[CRIME]	8183.14	Times $1000	8183.140352	Standard Error LOG(CRIME-hat)	0.123677	0.015296
E[CRIME]	8735.22	Times $1000		s2 = Var[Y]=Var[Log(CRIME)]		0.024628
				s = Standard Deviation [Log(CRIME)]		

95% Confidence Interval		95% Prediction Interval (or Credibility Interval)	
LB E[LOG(CRIME)]	3.66276	LB LOG(CRIME)	3.595500
UB E[LOG(CRIME)]	4.16309	UB LOG(CRIME)	4.230350

Approximate 95% Confidence Interval		95% Prediction Interval (or Credibility Interval)	
LB E[CRIME]	4600.02	CRIME	3940.03
UB E[CRIME]	14557.61	CRIME	16996.13

Approximate because Log is not a linear function

Key Takeaways: -

- The Median of the crime value obtained is **8183.14**.
- The Expected value of the crime value is **8735.2**.
- Prediction Interval (Crime)
  -> Lower Bound - **3940.03**
  -> Upper Bound – **16996.13**
- Confidence Interval for the Expected value (Crime)
  ->Lower Bound – **4600.02**
  ->Upper bound – **14557.61**

Conclusion

The width of the prediction interval obtained by the prediction of the best model is very large. The mean and the median of the crime value tend to converge with each other, which indicates normality. Both the confidence interval and the prediction interval include the expected value of the mean. The model without the interaction term has a smaller prediction interval even though it has a lower R-squared and a lower adjusted R-squared value.

The explanatory variables which contribute the most to the dependent variable are Ed, U2, M, Ineq, Prob, Po1 and U1.