Dec 11, 2017

Effects of Punishment Regimes on Criminal Rate: A Regression Approach

Mingyuhui Liu (Jane), Master of Science Candidate in Data Analytics

School of Engineering and Applied Sciences

The George Washington University

Table of Contents

Table of Figures and Tables

Acronyms

adjR-sq	Adjusted R-square
CI	Confidence Interval
DW	Durbin–Watson statistic
MSE	Mean Square Error
PI	Probability Interval
R-sq	R-Square
RVF	Residual Vs Fitted
RVO	Residual Vs Observation Order
SE	Standard Error
Std	Standard Deviation

Introduction

An integrated dataset consisting of 14 crime-related variables from 47 U.S. states in the year of 1960 was analyzed in this project, to find the punishment regimes' effects on the criminal rates. This project took a linear regression approach, and adjusted the model with several diagnosis analysis. After a model was selected, the project predicted crime rate with the model.

Data Pre-processing

An example of the original dataset is provided in Table 1. Descriptions of the 14 potential independent variables and the independent variable included in the dataset are shown in Table 2.

Table 1

Example of the Original Dataset

Crime	Po1	Po2	Wealth	Prob	Pop	Ed
791	5.8	5.6	3940	0.084602	33	9.1
1635	10.3	9.5	5570	0.029599	13	11.3
578	4.5	4.4	3180	0.083401	18	8.9
1969	14.9	14.1	6730	0.015801	157	12.1
1234	10.9	10.1	5780	0.041399	18	12.1
682	11.8	11.5	6890	0.034201	25	11
963	8.2	7.9	6200	0.0421	4	11.1

Example of the Original Dataset (Continued)

U1	U2	LF	M.F	Ineq	Time	M
0.108	4.1	0.51	95	26.1	26.2011	15.1
0.096	3.6	0.583	101.2	19.4	25.2999	14.3
0.094	3.3	0.533	96.9	25	24.3006	14.2
0.102	3.9	0.577	99.4	16.7	29.9012	13.6
0.091	2	0.591	98.5	17.4	21.2998	14.1
0.084	2.9	0.547	96.4	12.6	20.9995	12.1
0.097	3.8	0.519	98.2	16.8	20.6993	12.7

Table 2

Variable Description in Original Dataset

Variable	Description
Crime	Crime rate: Number of offenses per 100,000 population in 1960
Po1	per capita expenditure in police protection in 1960
Po2	per capita expenditure in police protection in 1959
Wealth	Wealth: Median value of transferrable assets or family income
Prob	probability of imprisonment: ratio of number of commitment to number of offenses
Pop	state population in 1960 in hundred thousands
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 35-39-24
LF	labour force participation rate of civilian urban male in the age-group 14-24
M.F.	number of males per 100 females

Ineq	Income inequality: percentage of families earning below half the median income
Time	average time in months served by offenders in state prisons before their first release
M	percentage of males aged 14-24 in total state population

Descriptive Statistics on Dependent Variable

From the Table 2, it is clear that given the objectives of this project, the Crime should be treated as the depend variable. To better understand the dependent variable, Figure 1 and Figure 2 provide the information on the distribution of the dependent variable Crime, with probability plot and histogram respectively.

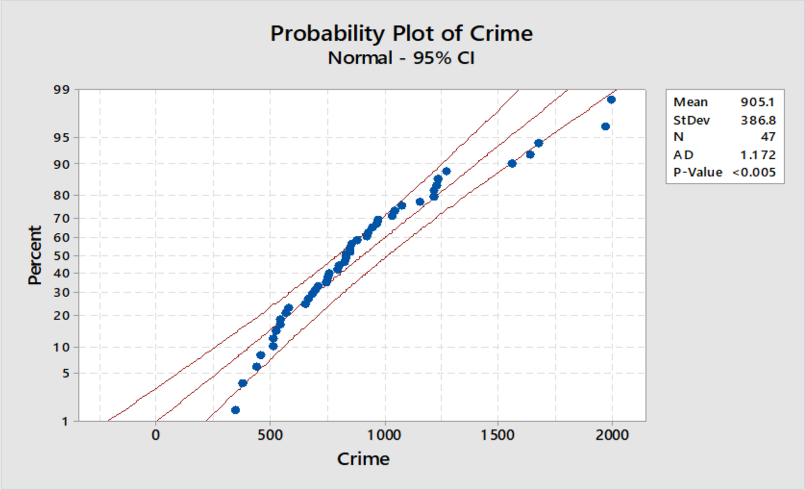


Figure 1. Probability plot of Crime. This figure shows the normality information on the dependent variable distribution.

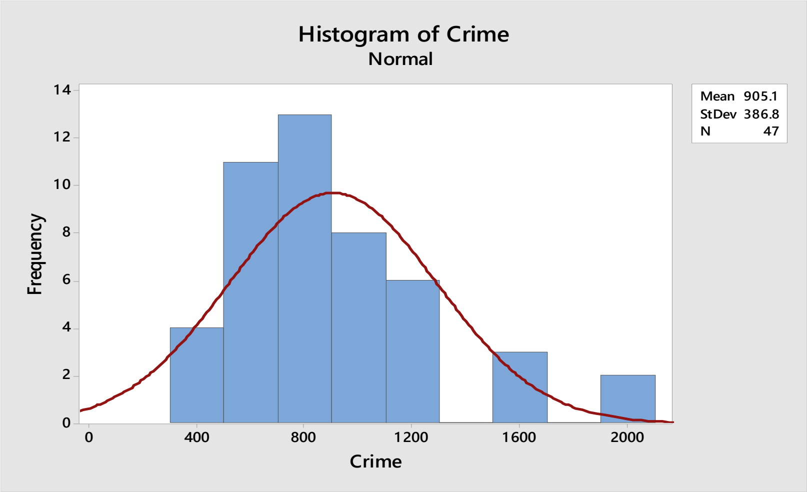


Figure 2. Histogram of Crime. This figure shows the histogram of crime in order to describe the distribution.

From Figure 1, a p-value less than 0.005 shows a significance that the distribution of the Crime is not following a normal distribution. Figure 2 confirms the non-normal distribution, showing a skewness in the Crime.

Log Transformation

In order to perform a regression on the Crime, the project transformed the dependent variable into a logged array. By doing so, the normality of the dependent variable, Logged Y, was improved. Figure 3 and Figure 4 shows the probability plot and histogram of the Logged Y.

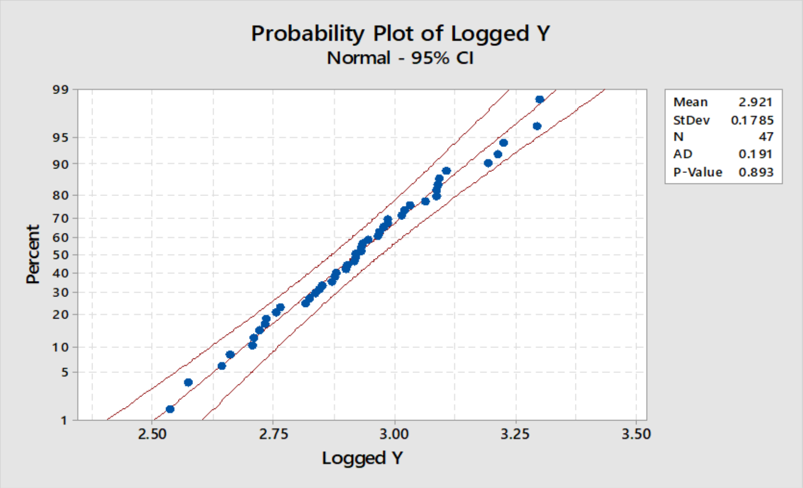


Figure 3. Probability plot for Logged Crime. This figure shows normality information on Logged Crime.

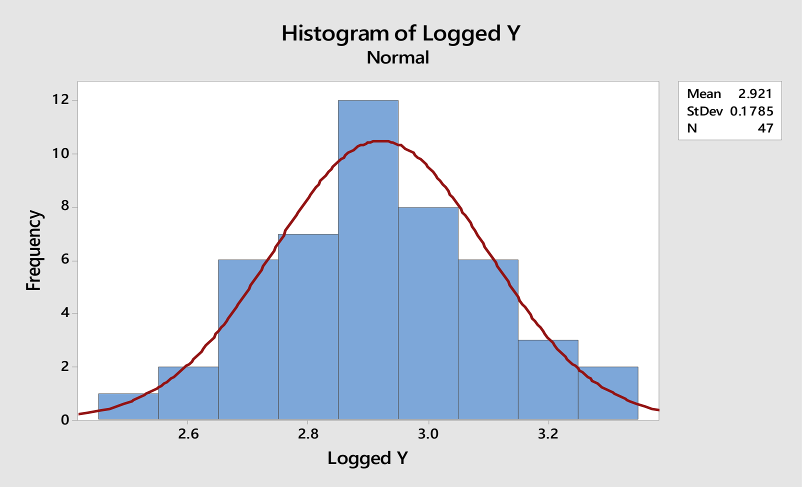


Figure 4. Histogram for Logged Crime. This figure shows normality information on Logged Crime.

From Figure 3, a p-value of 0.893 shows the normality in the Logged Y. Figure 4's histogram also shows an improvement of the normality for the dependent variable. Thus, the Logged Y was used in the project as the dependent variable.

Correlation Matrix

To obtain a general idea of what the model should be consist of, a correlation matrix was created with the original potential variables, and Table 3 shows the correlation matrix created based on the original dataset.

In Table 3, orange-highlighted cells represent correlations with absolute values larger than 0.4, which is the threshold indicating a sufficiently large correlation for consideration. There are two facts worthy of attention from Table 3. First, Po1, Po2, Wealth and Prob have relatively higher correlations with Logged Y, which is also why the correlation values for them are highlighted with red color font and a thick border box. Secondly, Po1, Po2, Wealth, Prob, Ed, and Ineq have different levels of correlations among them, which indicates potential multicollinearity when conducting regressions based on these variables.

Interactions

When considering the definitions of each variable, it is logical to possibly include interactions among variables. For example, by adding ed*Wealth to the dependent, the project would be able to examine the hypothesis of whether family income differs based on mean years of education received, while common sense suggests it might. Similarly with Ineq*Po1, the project would examine the hypothesis of whether poverty differs based on the investment in police enforcement, while common sense suggests it might.

In light of the possibilities for interaction terms, new variables were created to represent interactions. Table 4 shows a new correlation matrix containing new interactions. Table 4 orange-highlighted cells representing a correlation with an absolute value larger than 0.6. Compared to Table 3, it provided more options for potential regression dependent variables: Po1, Po2, Po1*Po2, Po1*Wealth, Ineq*Po1, Ineq*Po2, and potentially Po2*Wealth, ed*Wealth*Po1, ed*Wealth*Po2, given the correlations for these three variables just missed the cut by a small value.

Table 3

Correlation Matrix for the Logged Y and Original Independent Variables

	Logged Y	Po1	Po2	Wealth	Prob	Pop	Ed	U1	U2	LF	M.F	Ineq	Time	M
Logged Y	1													
Po1	0.6546315	1												
Po2	0.6373046	0.9935865	1											
Wealth	0.4266203	0.7872253	0.7942621	1										
Prob	-0.411892	-0.473247	-0.473027	-0.555335	1									
Pop	0.3373589	0.5262836	0.5137894	0.3082627	-0.347289	1								
Ed	0.3021452	0.4829521	0.4994096	0.735997	-0.389923	-0.017227	1							
U1	-0.074866	-0.043698	-0.051712	0.0448572	-0.007469	-0.03812	0.0181035	1						
U2	0.1674043	0.185093	0.1692242	0.0920717	-0.061592	0.2704216	-0.215682	0.7459248	1					
LF	0.1727319	0.1214932	0.1063496	0.2946323	-0.250086	-0.123672	0.561178	-0.2294	-0.420762	1				
M.F	0.1481607	0.0337603	0.0228425	0.1796086	-0.050858	-0.410628	0.4369149	0.3518919	-0.018692	0.5135588	1			
Ineq	-0.151693	-0.6305	-0.648152	-0.883997	0.4653219	-0.126294	-0.768658	-0.063832	0.0156782	-0.269886	-0.167089	1		
Time	0.1425776	0.1033577	0.0756267	0.0006486	-0.436246	0.4642105	-0.253974	-0.169853	0.1013583	-0.12364	-0.427697	0.1018228	1	
M	-0.056234	-0.505737	-0.513173	-0.670055	0.3611164	-0.280638	-0.53024	-0.224381	-0.244843	-0.160949	-0.02868	0.6392114	0.1145107	1

Table 4

New Correlation Matrix for Interactions

	Logged Y	*Po1*	*Po2*	*Wealth*	*Prob*	*Pop*	*Ed*	*U1*	*U2*	*LF*	*M.F*	*Ineq*	*Time*	*M*	*ed*ineq*	*ed*wealth*	*Po1*Po2*	*U1*LF*	*Po1*Wealth*	*Po2*Wealth*	*pop*Po1*	*pop*Po2*	*Ineq*Po1*	*Ineq*Po2*	*ed*wealth*Po1*	*ed*wealth*Po2*
Logged Y	1.00																									
Po1	0.65	1.00																								
Po2	0.64	0.99	1.00																							
Wealth	0.43	0.79	0.79	1.00																						
Prob	-0.41	-0.47	-0.47	-0.56	1.00																					
Pop	0.34	0.53	0.51	0.31	-0.35	1.00																				
Ed	0.30	0.48	0.50	0.74	-0.39	-0.02	1.00																			
U1	-0.07	-0.04	-0.05	0.04	-0.01	-0.04	0.02	1.00																		
U2	0.17	0.19	0.17	0.09	-0.06	0.27	-0.22	0.75	1.00																	
LF	0.17	0.12	0.11	0.29	-0.25	-0.12	0.56	-0.23	-0.42	1.00																
M.F	0.15	0.03	0.02	0.18	-0.05	-0.41	0.44	0.35	-0.02	0.51	1.00															
Ineq	-0.15	-0.63	-0.65	-0.88	0.47	-0.13	-0.77	-0.06	0.02	-0.27	-0.17	1.00														
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	1.00													
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	0.11	1.00												
*ed*ineq*	0.03	-0.54	-0.55	-0.69	0.35	-0.18	-0.31	-0.09	-0.17	0.07	0.11	0.84	-0.05	0.48	1.00											
*ed*wealth*	0.42	0.73	0.74	0.96	-0.53	0.21	0.89	0.04	-0.01	0.42	0.30	-0.89	-0.10	-0.65	-0.59	1.00										
*Po1*Po2*	0.61	0.98	0.98	0.73	-0.42	0.55	0.44	-0.04	0.19	0.08	0.02	-0.58	0.10	-0.48	-0.51	0.68	1.00									
*U1*LF*	0.02	0.03	0.01	0.18	-0.11	-0.08	0.26	0.92	0.59	0.16	0.58	-0.19	-0.23	-0.29	-0.08	0.23	0.01	1.00								
*Po1*Wealth*	0.60	0.98	0.98	0.87	-0.49	0.50	0.55	-0.01	0.18	0.17	0.08	-0.71	0.08	-0.57	-0.61	0.82	0.97	0.07	1.00							
*Po2*Wealth*	0.59	0.98	0.98	0.87	-0.49	0.49	0.56	-0.02	0.17	0.15	0.07	-0.73	0.06	-0.57	-0.62	0.82	0.97	0.06	1.00	1.00						
*pop*Po1*	0.37	0.65	0.64	0.41	-0.38	0.96	0.11	-0.01	0.24	-0.10	-0.32	-0.25	0.38	-0.35	-0.27	0.32	0.69	-0.05	0.62	0.62	1.00					
*pop*Po2*	0.37	0.65	0.65	0.41	-0.38	0.95	0.11	-0.02	0.23	-0.10	-0.32	-0.25	0.38	-0.35	-0.27	0.33	0.70	-0.05	0.63	0.63	1.00	1.00				
*Ineq*Po1*	0.79	0.83	0.81	0.41	-0.32	0.58	0.13	-0.11	0.22	0.01	-0.04	-0.12	0.18	-0.21	-0.06	0.34	0.82	-0.09	0.74	0.72	0.63	0.64	1.00			
*Ineq*Po2*	0.78	0.83	0.83	0.42	-0.32	0.57	0.15	-0.12	0.20	0.00	-0.05	-0.14	0.14	-0.22	-0.08	0.36	0.83	-0.10	0.75	0.74	0.64	0.64	0.99	1.00		
*ed*wealth*Po1*	0.60	0.96	0.96	0.88	-0.49	0.44	0.64	0.00	0.14	0.24	0.16	-0.74	0.03	-0.57	-0.58	0.86	0.95	0.12	0.99	0.99	0.58	0.58	0.70	0.71	1.00	
*ed*wealth*Po2*	0.58	0.96	0.97	0.88	-0.49	0.44	0.65	0.00	0.13	0.22	0.15	-0.75	0.01	-0.57	-0.59	0.87	0.95	0.11	0.99	0.99	0.58	0.58	0.68	0.70	1.00	1.00

Linear Regression

Methodologies

Model fitting. Based on the correlations, the project tested multiple combinations of independent variables, fitting them with linear regressions, in order to find the best-fitted model to estimate the dependent variable, i.e. Logged Y. The linear regression follows the Equation 1.

Equation 1

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{X}^{\boldsymbol{T}} + \boldsymbol{e}$$

y demotes the dependent variable matrix, and in this case, the dependent variable is Logged Y, indicating that the dimension of ***y*** is 1x1; ***A*** denotes the coefficient matrix, and in this case, it should be 1xn, where n represents the number of dependent variables; ***X*** denotes the dependent variable matrix, and in this case, the dimension of ***X*** is 1xn, while subscript ***T*** denotes the transposition; and ***e*** denotes the error term, with a dimension, specifically, of 1x1.

Model Selection and Diagnosis Analysis. Models are first selected based on the following criteria:
a) R-squared (R-sq), and adjusted R-squared (adjR-sq) values of the model to examine the variances coverage of the model;
b) Standard Error (SE) for coefficients to measure the precision of the prediction of the coefficients;
c) p-values for the coefficients to test the statistical significances for coefficients;
d) normality for residuals from the model;

Then, model will be adjusted and ultimately selected based on diagnosis analysis:
a) Multicollinearity measured by Variance inflation factor (VIF) for coefficients;
b) Heteroscedasticity measured by residual Vs fitted value (RVF);
c) Independence estimated by residual Vs observation order (RVO) plot;
d) Durbin–Watson statistic (DW) for autocorrelation;
e) and normality for ***X***.

Models

Initial Model. After cautiously examining various variables combinations, a model with an R-sq of 74.16% and an adjR-sq of 71.68% is selected. Independent variable matrix can be represented by Equation 2.

Equation 2

$$\boldsymbol{X}^{\boldsymbol{T}} = \begin{bmatrix} \text{Po1} \\ \text{Po2} \\ \text{Ineq} * \text{Po1} \\ \text{Ineq} * \text{Po2} \\ \text{ed} * \text{Wealth} \end{bmatrix}$$

Table 5 shows the coefficients statistics from this model. And Figure 5 and Figure 6 show the detailed residual probability plot for normality and a 4-in-one residual plot, respectively.

Table 5

Coefficient Statistics for the Regression Model

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	2.044	0.106	(1.830, 2.259)	19.24	0
Po1	0.489	0.226	(0.033, 0.944)	2.17	0.036
Po2	-0.573	0.238	(-1.053, -0.092)	-2.41	0.021
Ineq*Po1	-0.0257	0.0131	(-0.0522, 0.0007)	-1.96	0.056
Ineq*Po2	0.0332	0.0138	(0.0054, 0.0610)	2.41	0.02
ed*wealth	8E-06	0.000002	(0.000004, 0.000012)	3.93	0

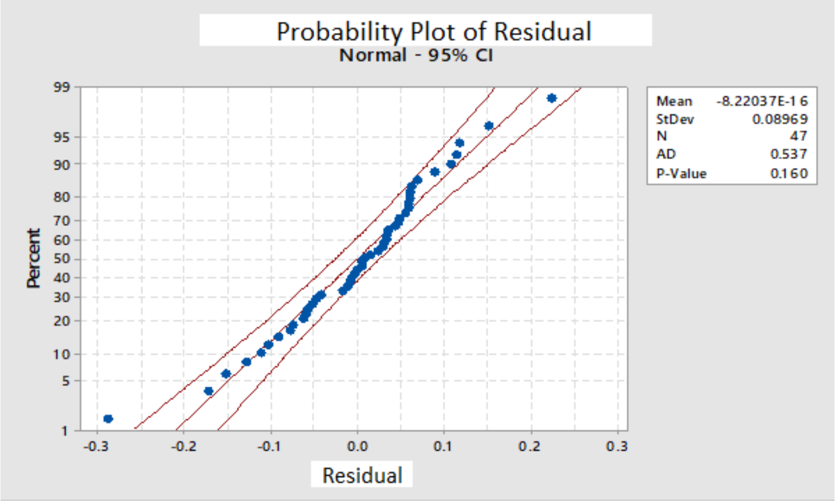


Figure 5. Detailed probability plot for residuals. This figure provides information on residuals' normality.

From Table 5 "Coef" column, which shows the coefficients for each variable, it can be concluded that all the selected dependent variables have positive effects on the Logged Y, except for Po2 and Ineq*Po1. "SE Coef" column provides information on the SE of coefficient, and from which Po1 and Po2 are deemed to have less precisely predicted coefficient compared to other variables, given their higher values in SE coefficient. Moreover, p-values confirms that most of the coefficients are statistically significant at 0.05 level, except for Ineq*Po2, which is significant at 0.06 level.

From Figure 5, the residuals are showing normality, with a p-value at 0.16.

The results are suggesting that this is a fine model to start with, and it is the initial model chosen by the research.

However, when it comes to diagnosis analysis, this model became problematic.

First, VIF values were used to measure the multicollinearity. Table 6 shows the VIF values for the initial model chosen.

Table 6

VIF Values for Independent Variables in the Initial Model

	Po1	Po2	Ineq*Po1	Ineq*Po2	ed*wealth
VIF	2291.86	2255.26	1408.54	1338.87	4.21

From Table 6, even though with interaction terms in the model, the VIF is too large, indicating high correlations among the variables.

The project first took the most common method to solve the problem from high VIF, to drop one or several variables.

Model 1. After testing, in the end, a new model was chosen, called "model 1". Model 1 dropped Po2 and Ineq*Po2 from the initial model, while decreased the adjR-sq from 71.68% to 69.14%, indicating a 3.45% decrease. However, shown by Table 7, the VIF of the coefficient from this model are significantly smaller.

Table 7

Coefficient Statistics for Model 1 with VIF

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.029	0.110	(1.806, 2.251)	18.38	0	
Po1	-0.053	0.0169	(-0.0871, -0.0188)	-3.13	0.003	11.85
Ineq*Po1	0.005815	0.000908	(0.003983, 0.007646)	6.40	0	6.22
ed*wealth	0.000008	0.000002	(0.000003, 0.000012)	3.68	0.001	4.19

From Table 7, after dropping several variables, both the p-values and the SE coefficient values show improvements of the model, while VIF falls from an abnormal level to a normal level. However, Po1 and Ineq*Po1 still have VIF values larger than 5, which still indicates multicollinearity (Eberly College of Science, [PDF document]).

Model 2. As suggested, instead of utilizing the original data, the project centered the independent variables aiming at decreasing the VIF values. The centering is done by subtracting the mean of each independent variable from the original value.

After the centering, the project utilized the centered data to conduct linear regressions again, and selected a Model 2, whose features are shown in Table 8.

Table 8

Coefficient Statistics for Model 2 with VIF

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.4965	0.0704	(2.3545, 2.6385)	35.46	0	
Po1	0.04967	0.00779	(0.03395, 0.06538)	6.37	0	1.44

ed*Wealth C	0.00006	0.000027	(0.000006, 0.000115)	2.24	0.03	2.17
Ineq*Po1 C	0.00615	0.00275	(0.00061, 0.01169)	2.24	0.03	2.27

Note. "ed*Wealth C" and "Ineq*Po1 C" denote the centered ed*Wealth and centered Ineq*Po1 respectively.

From Table 8, when choosing Po1 and centered ed*Wealth and centered Ineq*Po1 as independent variables for Model 2, all VIF values are reduced below 2.30, while p-values for all the coefficients are showing significances.

However, the R-sq and adjR-sq also decreases to 49.67% and 46.16% respectively.

Model Comparisons

To summarize and compare the three potential models, Table 9 provides an overall model and diagnosis analysis. In Table 9, the best statistics for R-sq, adjR-sq, DW-statistics, and VIF values are highlighted, and noted that each model has highlighted advantages, at the same time, each of the models has cons.

Initial Model. Pros of the Initial Model:
a) highest R-sq and adjR-sq values, representing the largest coverage of the variances in the 3 models;
b) a DW-statistics close to 2, which indicates little autocorrection;
c) and has the best Heteroscedasticity among the 3 models based on the VRF plots.

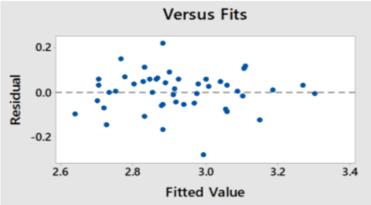
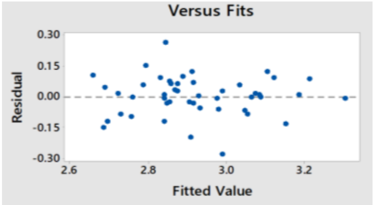
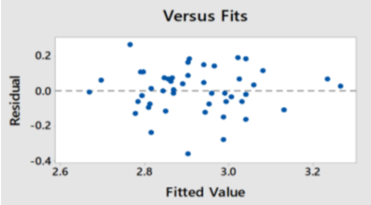
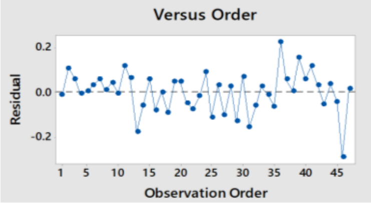
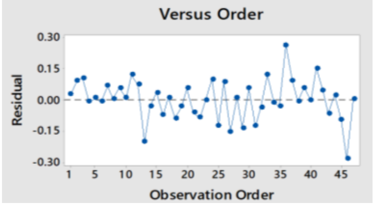
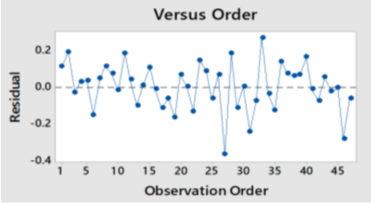
Cons of the Initial Model:
a) Po1 and Po2 have relatively high SE Coefficient which indicates less precision in the estimation of the coefficients;
b) Ineq*Po1 has a coefficient p-value of 0.056, which failed to show significance of a non-zero coefficient at a 0.05 level;
c) the incredibly high VIF values indicating high possibility of multicollinearity;
d) has the lowest p-value for residual normality test;
e) based on RVO plot, the first and last few observations show consecutive values above the reference line, which indicates the variables may not be independent;
f) and 2 out of 5 of the variables show small p-values, indicating non-normality.

Model 1. After dropping several variables, main pros for Model 1 are:
a) relatively high R-sq and adjR-sq values;
b) has a DW-statistics closest to 2.00;
c) SE of coefficient decreased compared to the Initial Model, implying more precisions in estimations of the coefficients;
d) every term has a significant evidence for non-zero coefficients at 0.05 level;
e) VIF values dropped dramatically compared to the Initial Model;
f) residuals normality p-value increase compared to that of the Initial Model;
g) and based on the RVO plot, the independence improved compared to the last few observations from the Initial Model.

Cons for Model 2 are:

Table 9

Summary for 3 Models' Diagnosis Analysis

	Initial Model					Model 1					Model 2				
Model Summary	R-sq	74.76%				R-sq	71.15%				R-sq	49.67%			
	adjR-sq	71.68%				adjR-sq	69.14%				adjR-sq	46.16%			
	DW-Statistics	2.04173				DW-Statistics	2.01403				DW-Statistics	2.21954			
Coefficients	Term	Coef	SE Coef	P-Value	VIF	Term	Coef	SE Coef	P-Value	VIF	Term	Coef	SE Coef	P-Value	VIF
	Constant	2.044	0.106	0.000		Constant	2.029	0.110	0.000		Constant	2.497	0.070	0.000	
	Po1	0.489	0.226	0.036	2291.86	Po1	-0.053	0.017	0.003	11.85	Po1	0.050	0.008	0.000	1.44
	Po2	-0.573	0.238	0.021	2255.26	Ineq*Po1	0.006	0.001	0.000	6.22	Ineq*Po1 C	0.006	0.003	0.030	2.27
	Ineq*Po1	-0.026	0.013	0.056	1408.54	ed*wealth	0.000008	0.000002	0.001	4.19	Ed*Wealth C	0.00006	0.000027	0.030	2.17
	Ineq*Po2	0.033	0.014	0.020	1338.87										
	ed*wealth	0.000008	0.000002	0.000	4.21										
Residual	Normality p-value	0.160				Normality p-value	0.299				Normality p-value	0.639			
	Heteroscedasticity					Heteroscedasticity					Heteroscedasticity				
	Independence					Independence					Independence				
Normality		p-value					p-value					p-value			
	Po1	0.009				Po1	0.009				Po1	0.009			
	Po2	0.010				Ineq*Po1	0.709				Ineq*Po1 C	0.709			
	Ineq*Po1	0.709				ed*wealth	0.073				Ed*Wealth C	0.073			
	Ineq*Po2	0.148													
	ed*wealth	0.073													

Note. For Model 2, the Coef of the Constant term, which is the interception, is substituted with that from the Excel output, other than the Minitab (Minitab 18 Statistical Software, 2017.) output. This will be further discussed in the final subsection in the Conclusion and Discussion section.

a) Po1 and Ineq*Po1 still have VIF values larger than 5;
b) and based on RVO plot, there might still be dependence especially for the first few observations.

Model 2. This is the model based on the centered independent variables. And the pros for Model 1 are:
a) has the best SE of coefficients, which all show high precisions in estimation for coefficients;
b) VIF values are all below 5, which indicates little likelihood for multicollinearity;
c) has the highest residual normality p-value;
d) and based on the RVO plot, the independence improves.

While Model 2 tackled the problem of multicollinearity successfully, however, the cons of Model 2 are significant as well:
a) the R-sq and adjR-sq dropped below 50%, with values of 49.67% and 46.16% respectively, indicating poor coverage of the variances from the estimated values;
b) and RVF plot show huge heteroscedasticity from the residuals.

Model selection. Based on the comparisons, and taking the purpose of this project into account - to predict crime rate based on provided information, which should emphasize on the variance coverage, thus Model 1 is suggested.

The rationale of such suggestion lies in that the VIF number relatively acceptable while it still maintains high coverage of variances based on 69.14% in adjR-sq.

Selected Model

Interpretation. Model 1 is selected. And the model could be summarized as Equation 3.

Equation 3

*Logged Y = 2.029 - 0.0530 Po1 + 0.005815 Ineq*Po1 + 0.000008 ed*Wealth*

A simple interpretation of Equation 3 is that, for every unit of increase in per capita expenditure in police protection in 1960 (Po1), the logged crime rate (Logged Y) will decrease 0.0530 unit and increase 5.815% of the change of percentage of families earning below half the median income (Ineq); and when either mean years of schooling of the population aged 25 years or over (ed) or median value of transferrable assets or family income (Wealth) increases 1 unit, the logged crime rate will increase 0.0008% of the change in another variable.

Noted that even though the coefficient of ed*Wealth is extremely small, the 95% Confidence Interval (CI) is (0.000003, 0.000012), where zero is not included, and with a p-value of 0.

Figures and statistics. Figure 6 visualized the model with fitted and true values. And Figure 7 features a probability plot for the residual of chosen model – Model 1. As shown in Figure 7, other than the normality of the residual's distribution could be concluded, several potential influential observations could also be spotted. In order to identify the influential observations, a Fit and Diagnosis Table is output from Minitab 18. The detailed table is attached in the Appendix. One influential observation is statistically detected: with a leverage (Hi) of 0.28084, and DFITS of -1.01016. However, with only 47 observations in total, this project did not take actions on these outliers (Minitab, Inc.. n.d.).

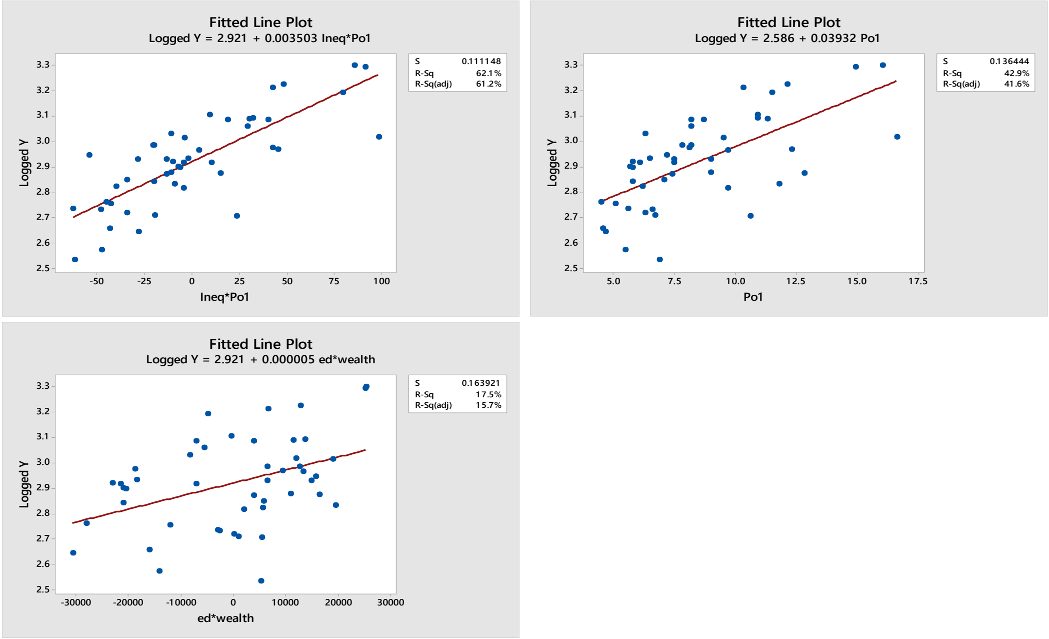


Figure 6. Fitted line Vs true values. This figure visualized the model.

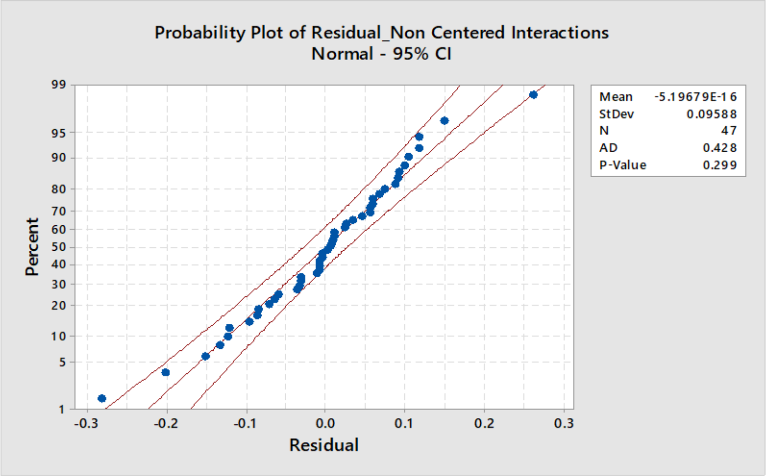


Figure 7. Residual probability plot for Model 1. This figure provides information on the normality of residual and influential observations.

Table 10 shows the contributions to the dependent variable from each term in Equation 3. From Table 10, Po1 is contributing to the Logged Y the most.

Table 10

Contributions from Terms in the Model

Po1	Ineq*Po1	ed*wealth	Error
42.85%	19.23%	9.07%	28.85%

Predictions and Discussion

Prediction

Recalling Equation 3, the project created interaction terms for Ineq*Po1 and ed*Wealth based on the information provided.

Table 11 provides independent variables information for the prediction.

Table 11

Information on Independent Variables for Prediction

Po1	Ineq*Po1	ed*Wealth
16	432	82680

And the predicted result is Logged Y equals to 4.319.

Noted that this model treated logged crime rate as the dependent variable, thus, the next step of the prediction should transform the predicted value back. The transformation should follow Equation 4.

Equation 4

$$\text{Crime Rate} = 10^{\text{Logged } Y}$$

Therefore, the predicted value for the crime rate is 20,839[1].

Uncertainty

Calculated in Excel, Table 12 reports the uncertainties from the prediction.

Table 12

Uncertainties Reported from the Prediction

	Interval	Logged Crime	Crime Rate
95% of Prediction Interval	Upper	4.516	32843.438
	Lower	4.121	13221.701

Note. A detailed table for calculation is provided in Appendix Table 2.

Since Excel does not feature a tool to estimate the SE for estimation, which will determine the prediction intervals. The project had to estimate the prediction interval manually. In order to simplify and estimate an

[1] $10^{4.32}$ equals to 20838.57, however, considering the context of the crime rate, that is the number of offenses per 100,000 population in 1960, the number is rounded when reported.

approximate value, the project treated the model SE as the SE for estimation (How to Make Predictions from a Multiple Regression Analysis. [Video file]). Together with estimated the t-value for the prediction, a margin of error was generated and then the upper bound and lower bound of the prediction intervals were predicted (For more detailed calculation rationale, please refer to the Excel File, under Excel_Uncertainty worksheet). Since the SE for estimation is always a little larger than the SE for the model, the prediction intervals estimated by this research is narrower than it should have been.

Conclusion and Discussion

Summary and discussion on model. The research selected a model with an adjR-sq of 0.69. This model predicted logged crime rate based on per capita expenditure in police protection in 1960, an interaction term between per capita expenditure in police protection in 1960 and percentage of families earning below half the median income, as well as a second interaction term between mean years of schooling of the population aged 25 years or over and median value of transferrable assets or family income. All independent variables show significant coefficients and the evidence for correlations within these variables is moderate. This model also performed relatively well with other diagnosis analysis.

The selected model can be expressed by Equation 3. Noted that the logarithm is a monotonic increasing function and is the inverse operation to exponentiation with a base of 10, meaning that even though when transforming the logged crime rate back with Equation 4, the model cannot be interpreted in the same way anymore, it is still safe to say that the variables will have the same positive or negative effects on the crime rate. When looking at the effect of per capita expenditure in police protection in 1960 (Po1), the total effect of Po1 can be expressed as a main negative effect, with a partial positive effect when the percentage of families earning below half the median income changes positively[2]. The crime rate performance based on these two variables is in line with the common sense: when investing more in police enforcement, the crime rate should go down, while the poverty increases, the crime rate increases regardless the police enforcement.

However, this model shows an abnormality in the second interaction term. In the model, the crime rate will increase when the education level and the wealth level increase. One possible explanation is that as the wealth level increase in the area, it is more and more likely to become a target for crime. Another explanation could be that the "wealth" variable from the original dataset is not a fair indicator in the first place, since it is only showing the median of the income. There could be situations that the median is high, but the income distribution is heavily skewed, that being said the poverty can still be severe and the poverty gap can be overwhelming.

Summary and discussion on prediction. The research predicted a crime rate of 20,839 offenses per 100,000 population in 1960, based on a selected model with 69.14% coverage of the variance.

As seen from the prediction, 20,839 of crime rate is incredibly high, especially considering the maximum value of crime in the original dataset is 1,993. However, after a cautious investigation with the data and

[2] The "negative" and "positive" effects on crime rate mean statistically the rate decreases and increases respectively, not in social terms.

methodology, the research deemed this situation as an abnormality from the provided prediction statistics, other than from the model itself.

The maximum value of Po1 in the original dataset is 16, for Ineq*Po1 is 255.64 and for ed*Wealth is 81,554. A more detailed description for these variables can be found in the Appendix Table 3. And the provided data for prediction has a Po1 of 16, an Ineq*Po1 of 432 and an ed*Wealth of 82680. As demonstrated, every variable is either equal to or much larger than the maximum value of the original dataset, which indicates that the model is likely to be predicting a data which it has not been trained on. To make the case worse, all the variables with larger than the maximum values are having positive effects on the logged crime rate. More accurately, from Table 10, the Ineq*Po1 contributes 19.23%, positively, on the dependent variable, and ed*Wealth contributes 9.07% positively. With such large contributions and large differences, it is not odd to predict with such a result.

Discussion on Excel Vs Minitab. When fitting the model with the same dataset, Minitab and Excel were returning the same result, except for the coefficient for the Constant term, which is the interception value. This situation was also mentioned in Table 9. The project suspected this as an error from Minitab, since it is absurd to observe all the exact same output from the p-values for the model and all the coefficients, the R-sq and adjR-sq, even each individual fitted value and residuals, except only for the coefficient for the interception. Moreover, when referring to Figure 6, which is the fitted Vs true data plot generated by Minitab itself, regressions for each variables all have constant values less than 3.0. Therefore, to further investigate the situation, the project tested the output interception coefficient from the Minitab with true values from the original dataset (circled in orange in Appendix Table 1), and in the end, the prediction using the coefficient did not match neither the true value nor the stated fitted value by the Minitab itself (Please find the detailed testing in the Excel file, under “Excel_Prediction” worksheet). Therefore, the project still believes the model is acceptable but the coefficient for interception needs to be substituted by that from Excel.

References

Eberly College of Science, The Pennsylvania State University. *Detecting multicollinearity using variance inflation factors.* [PDF document]. Retrieved from https://onlinecourses.science.psu.edu/stat501/node/347

How to Make Predictions from a Multiple Regression Analysis. [Video file]. Retrieved from https://www.youtube.com/watch?v=E73AJ73-S6g

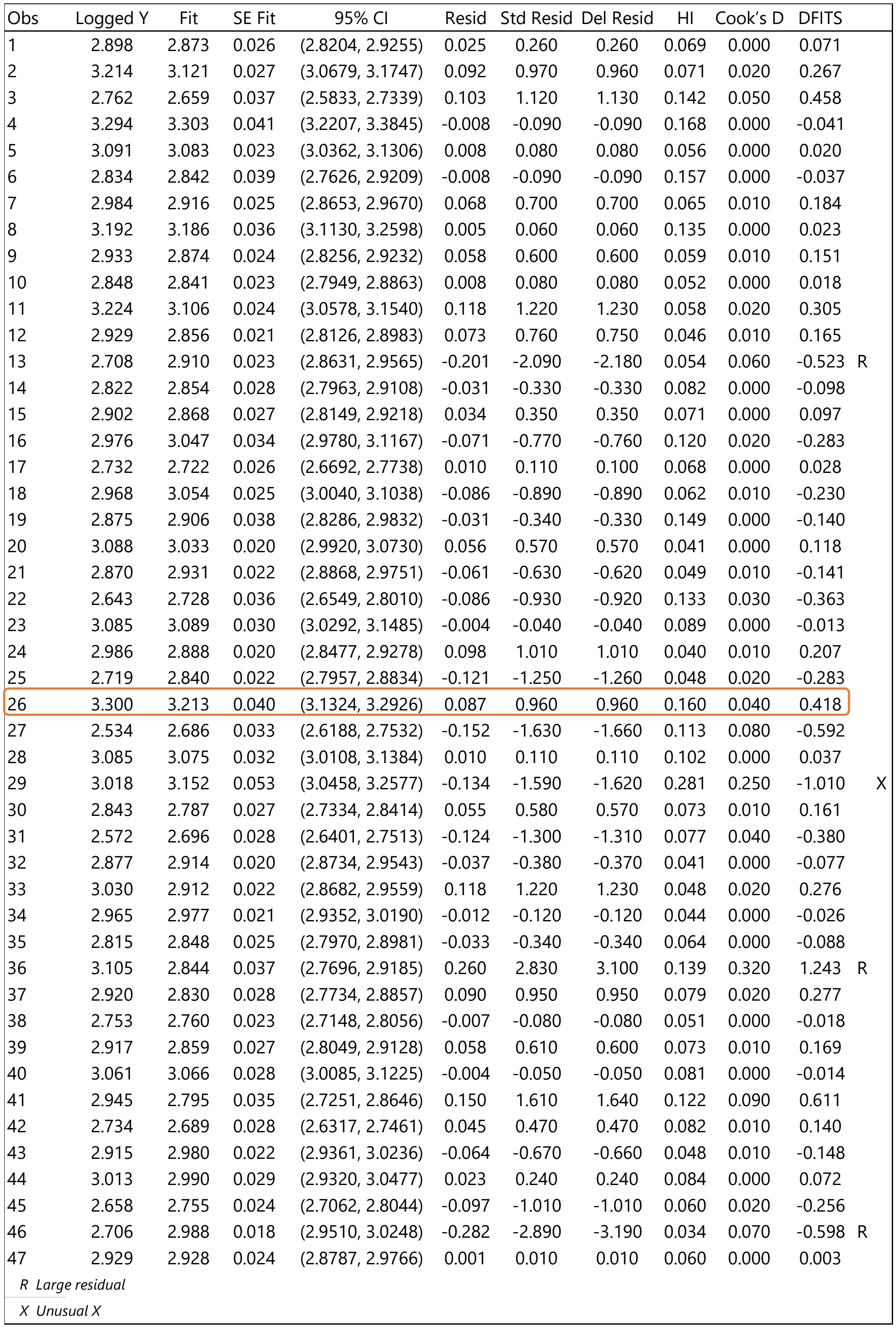
Minitab 18 Statistical Software (2017). [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com)

Minitab Inc. (n.d.) *Fits and diagnostics table for Fit Regression Model*. Retrieved from Minitab: https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/fit-regression-model/interpret-the-results/all-statistics-and-graphs/fits-and-diagnostics-table/

Appendix

Table 1

Fit and Diagnosis Table Output from Minitab



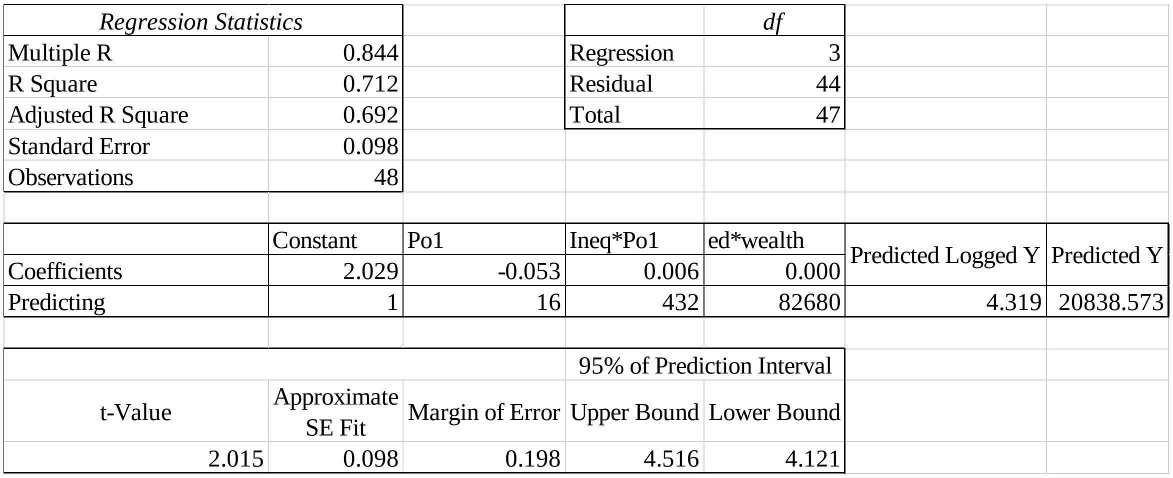
Obs	Logged Y	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D	DFITS	
1	2.898	2.873	0.026	(2.8204, 2.9255)	0.025	0.260	0.260	0.069	0.000	0.071	
2	3.214	3.121	0.027	(3.0679, 3.1747)	0.092	0.970	0.960	0.071	0.020	0.267	
3	2.762	2.659	0.037	(2.5833, 2.7339)	0.103	1.120	1.130	0.142	0.050	0.458	
4	3.294	3.303	0.041	(3.2207, 3.3845)	-0.008	-0.090	-0.090	0.168	0.000	-0.041	
5	3.091	3.083	0.023	(3.0362, 3.1306)	0.008	0.080	0.080	0.056	0.000	0.020	
6	2.834	2.842	0.039	(2.7626, 2.9209)	-0.008	-0.090	-0.090	0.157	0.000	-0.037	
7	2.984	2.916	0.025	(2.8653, 2.9670)	0.068	0.700	0.700	0.065	0.010	0.184	
8	3.192	3.186	0.036	(3.1130, 3.2598)	0.005	0.060	0.060	0.135	0.000	0.023	
9	2.933	2.874	0.024	(2.8256, 2.9232)	0.058	0.600	0.600	0.059	0.010	0.151	
10	2.848	2.841	0.023	(2.7949, 2.8863)	0.008	0.080	0.080	0.052	0.000	0.018	
11	3.224	3.106	0.024	(3.0578, 3.1540)	0.118	1.220	1.230	0.058	0.020	0.305	
12	2.929	2.856	0.021	(2.8126, 2.8983)	0.073	0.760	0.750	0.046	0.010	0.165	
13	2.708	2.910	0.023	(2.8631, 2.9565)	-0.201	-2.090	-2.180	0.054	0.060	-0.523	R
14	2.822	2.854	0.028	(2.7963, 2.9108)	-0.031	-0.330	-0.330	0.082	0.000	-0.098	
15	2.902	2.868	0.027	(2.8149, 2.9218)	0.034	0.350	0.350	0.071	0.000	0.097	
16	2.976	3.047	0.034	(2.9780, 3.1167)	-0.071	-0.770	-0.760	0.120	0.020	-0.283	
17	2.732	2.722	0.026	(2.6692, 2.7738)	0.010	0.110	0.100	0.068	0.000	0.028	
18	2.968	3.054	0.025	(3.0040, 3.1038)	-0.086	-0.890	-0.890	0.062	0.010	-0.230	
19	2.875	2.906	0.038	(2.8286, 2.9832)	-0.031	-0.340	-0.330	0.149	0.000	-0.140	
20	3.088	3.033	0.020	(2.9920, 3.0730)	0.056	0.570	0.570	0.041	0.000	0.118	
21	2.870	2.931	0.022	(2.8868, 2.9751)	-0.061	-0.630	-0.620	0.049	0.010	-0.141	
22	2.643	2.728	0.036	(2.6549, 2.8010)	-0.086	-0.930	-0.920	0.133	0.030	-0.363	
23	3.085	3.089	0.030	(3.0292, 3.1485)	-0.004	-0.040	-0.040	0.089	0.000	-0.013	
24	2.986	2.888	0.020	(2.8477, 2.9278)	0.098	1.010	1.010	0.040	0.010	0.207	
25	2.719	2.840	0.022	(2.7957, 2.8834)	-0.121	-1.250	-1.260	0.048	0.020	-0.283	
26	3.300	3.213	0.040	(3.1324, 3.2926)	0.087	0.960	0.960	0.160	0.040	0.418	
27	2.534	2.686	0.033	(2.6188, 2.7532)	-0.152	-1.630	-1.660	0.113	0.080	-0.592	
28	3.085	3.075	0.032	(3.0108, 3.1384)	0.010	0.110	0.110	0.102	0.000	0.037	
29	3.018	3.152	0.053	(3.0458, 3.2577)	-0.134	-1.590	-1.620	0.281	0.250	-1.010	X
30	2.843	2.787	0.027	(2.7334, 2.8414)	0.055	0.580	0.570	0.073	0.010	0.161	
31	2.572	2.696	0.028	(2.6401, 2.7513)	-0.124	-1.300	-1.310	0.077	0.040	-0.380	
32	2.877	2.914	0.020	(2.8734, 2.9543)	-0.037	-0.380	-0.370	0.041	0.000	-0.077	
33	3.030	2.912	0.022	(2.8682, 2.9559)	0.118	1.220	1.230	0.048	0.020	0.276	
34	2.965	2.977	0.021	(2.9352, 3.0190)	-0.012	-0.120	-0.120	0.044	0.000	-0.026	
35	2.815	2.848	0.025	(2.7970, 2.8981)	-0.033	-0.340	-0.340	0.064	0.000	-0.088	
36	3.105	2.844	0.037	(2.7696, 2.9185)	0.260	2.830	3.100	0.139	0.320	1.243	R
37	2.920	2.830	0.028	(2.7734, 2.8857)	0.090	0.950	0.950	0.079	0.020	0.277	
38	2.753	2.760	0.023	(2.7148, 2.8056)	-0.007	-0.080	-0.080	0.051	0.000	-0.018	
39	2.917	2.859	0.027	(2.8049, 2.9128)	0.058	0.610	0.600	0.073	0.010	0.169	
40	3.061	3.066	0.028	(3.0085, 3.1225)	-0.004	-0.050	-0.050	0.081	0.000	-0.014	
41	2.945	2.795	0.035	(2.7251, 2.8646)	0.150	1.610	1.640	0.122	0.090	0.611	
42	2.734	2.689	0.028	(2.6317, 2.7461)	0.045	0.470	0.470	0.082	0.010	0.140	
43	2.915	2.980	0.022	(2.9361, 3.0236)	-0.064	-0.670	-0.660	0.048	0.010	-0.148	
44	3.013	2.990	0.029	(2.9320, 3.0477)	0.023	0.240	0.240	0.084	0.000	0.072	
45	2.658	2.755	0.024	(2.7062, 2.8044)	-0.097	-1.010	-1.010	0.060	0.020	-0.256	
46	2.706	2.988	0.018	(2.9510, 3.0248)	-0.282	-2.890	-3.190	0.034	0.070	-0.598	R
47	2.929	2.928	0.024	(2.8787, 2.9766)	0.001	0.010	0.010	0.060	0.000	0.003	

R Large residual

X Unusual X

Table 2

Calculation of Uncertainties from Excel



Regression Statistics				*df*
Multiple R	0.844		Regression	3
R Square	0.712		Residual	44
Adjusted R Square	0.692		Total	47
Standard Error	0.098			
Observations	48			

	Constant	Po1	Ineq*Po1	ed*wealth	Predicted Logged Y	Predicted Y
Coefficients	2.029	-0.053	0.006	0.000		
Predicting	1	16	432	82680	4.319	20838.573

			95% of Prediction Interval	
t-Value	Approximate SE Fit	Margin of Error	Upper Bound	Lower Bound
2.015	0.098	0.198	4.516	4.121

Table 3

Description of the Variables

Po1		Ineq1*Po1		ed*Wealth	
Mean	8.5	Mean	157.5834	Mean	56278.128
Median	7.8	Median	150.48	Median	60144
Minimum	4.5	Minimum	95.2	Minimum	25632
Maximum	16.6	Maximum	255.64	Maximum	81554