

CASE STUDY: CRIME RATE EQUATION ESTIMATION IN DIFFERENT STATES OF USA IN THE 1960

CRIME DATA

Descripti on

Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The variables seem to have been re-scaled to convenient numbers. The data set contains the following columns:

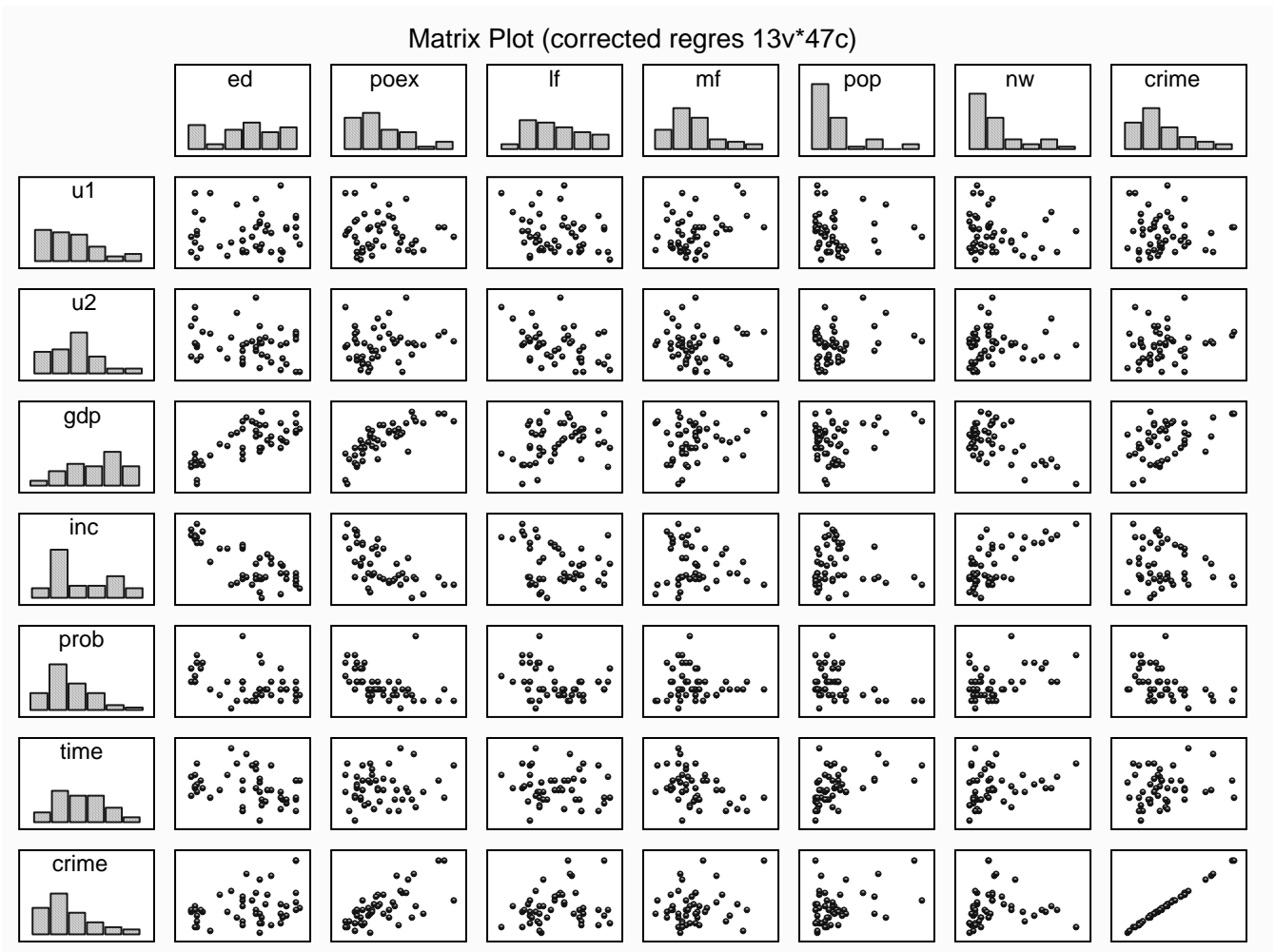
Vari abl e	Descri ption
Ed	mean years of schooling
PoEXP	police expenditure
LF	labor force participation rate
MF	number of males per 1000 females
Pop	state population
NW	number of nonwhites per 1000 people
U1	unemployment rate of urban males 14–24
U2	unemployment rate of urban males 35–39
GDP	gross domestic product per head
Inc	income inequality
Prob	probability of imprisonment
Unem	unemployment rate
Time	average time served in state prisons
Crime	rate of crimes in a particular category per head of population

Source: www.statsci.org

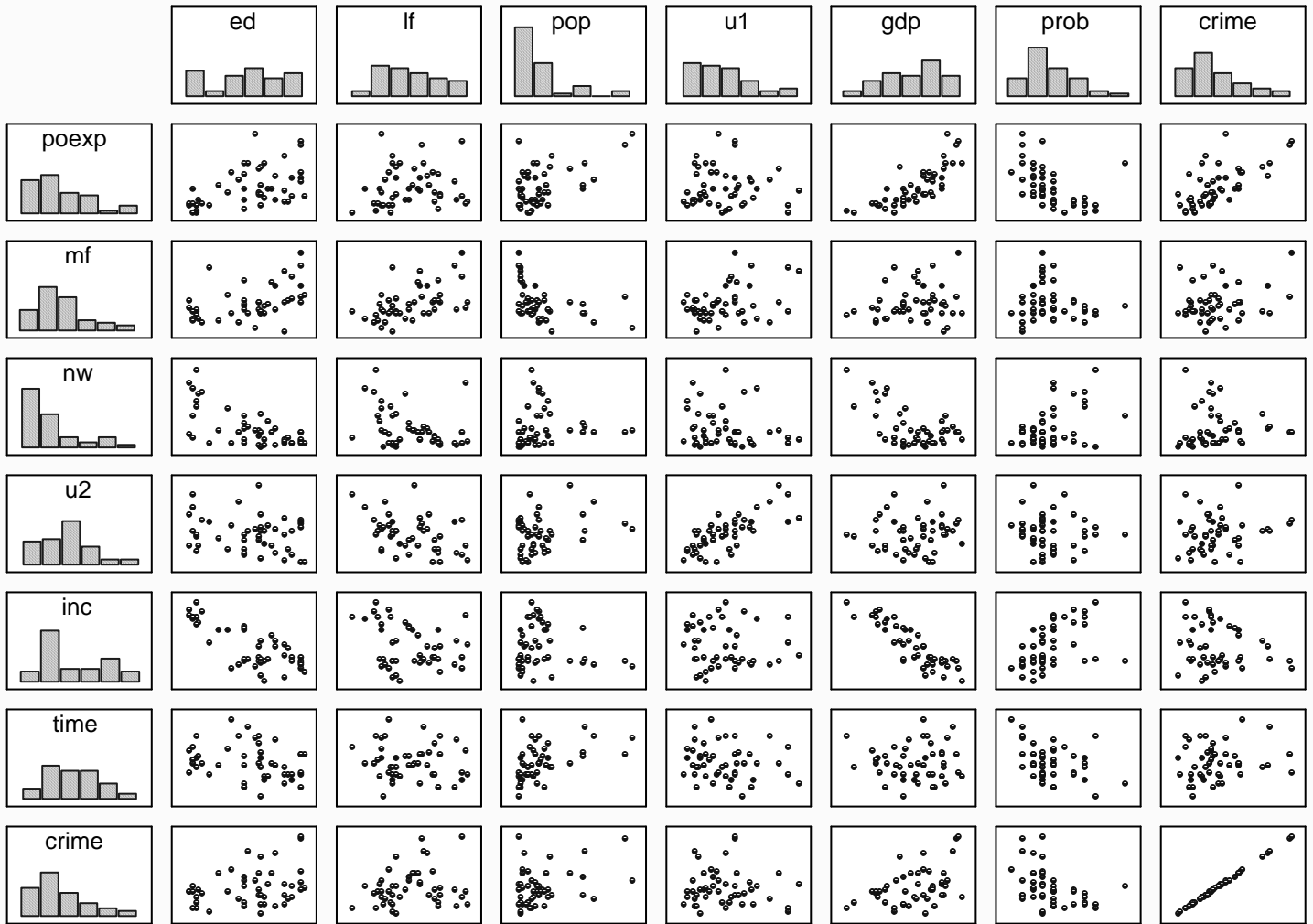
<i>Ed</i>	<i>PoEXP</i>	<i>LF</i>	<i>Mf</i>	<i>Pop</i>	<i>NW</i>	<i>U1</i>	<i>U2</i>	<i>GDP</i>	<i>INC</i>	<i>PROB</i>	<i>UNEM</i>	<i>TIME</i>	<i>CRIME</i>
91	114	510	950	33	301	108	41	394	261	0.08	149	26.2	791
113	198	583	1012	13	102	96	36	557	194	0.03	132	25.3	1635
89	89	533	969	18	219	94	33	318	250	0.08	127	24.3	578
121	290	577	994	157	80	102	39	673	167	0.02	141	29.9	1969
121	210	591	985	18	30	91	20	578	174	0.04	111	21.3	1234
110	233	547	964	25	44	84	29	689	126	0.03	113	21	682
111	161	519	982	4	139	97	38	620	168	0.04	135	20.7	963
109	224	542	969	50	179	79	35	472	206	0.04	114	24.6	1555
90	127	553	955	39	286	81	28	421	239	0.07	109	29.4	856
118	139	632	1029	7	15	100	24	526	174	0.04	124	19.6	705
105	237	580	966	101	106	77	35	657	170	0.02	112	41.6	1674
108	146	595	972	47	59	83	31	580	172	0.03	114	34.3	849
113	127	624	972	28	10	77	25	507	206	0.05	102	36.3	511
117	123	595	986	22	46	77	27	529	190	0.05	104	21.5	664
87	110	530	986	30	72	92	43	405	264	0.07	135	22.7	798
88	158	497	956	33	321	116	47	427	247	0.05	163	26.1	946
110	129	537	977	10	6	114	35	487	166	0.08	149	19.1	539
104	238	537	978	31	170	89	34	631	165	0.12	123	18.2	929
116	256	536	934	51	24	78	34	627	135	0.02	112	24.9	750
108	218	567	985	78	94	130	58	626	166	0.03	188	26.4	1225
108	141	602	984	34	12	102	33	557	195	0.02	135	37.6	742
89	91	512	962	22	423	97	34	288	276	0.09	131	37.1	439
96	170	564	953	43	92	83	32	513	227	0.03	115	25.2	1216
116	151	574	1038	7	36	142	42	540	176	0.04	184	17.6	968
116	120	641	984	14	26	70	21	486	196	0.07	91	21.9	523
121	303	631	1071	3	77	102	41	674	152	0.04	143	22.1	1993
109	140	540	965	6	4	80	22	564	139	0.04	102	28.5	342
112	158	571	1018	10	79	103	28	537	215	0.04	131	25.8	1216
107	323	521	938	168	89	92	36	637	154	0.02	128	36.7	1043
89	112	521	973	46	254	72	26	396	237	0.08	98	28.3	696
93	109	535	1045	6	20	135	40	453	200	0.04	175	21.8	373
109	171	586	964	97	82	105	43	617	163	0.04	148	30.9	754
104	127	560	972	23	95	76	24	462	233	0.05	100	25.5	1072
118	194	542	990	18	21	102	35	589	166	0.04	137	21.7	923
102	184	526	948	113	76	124	50	572	158	0.02	174	37.4	653
100	207	531	964	9	24	87	38	559	153	0.01	125	44	1272
87	114	638	974	24	349	76	28	382	254	0.05	104	31.7	831
104	98	599	1024	7	40	99	27	425	225	0.05	126	16.7	566
88	115	515	953	36	165	86	35	395	251	0.05	121	27.3	826
104	156	560	981	96	126	88	31	488	228	0.04	119	29.3	1151
122	138	601	998	9	19	84	20	590	144	0.03	104	30	880
109	110	523	968	4	2	107	37	489	170	0.09	144	12.2	542
99	145	522	996	40	208	73	27	496	224	0.05	100	32	823
121	191	574	1012	29	36	111	37	622	162	0.03	148	30	1030
88	87	480	968	19	49	135	53	457	249	0.06	188	32.6	455
104	203	599	989	40	24	78	25	593	171	0.05	103	16.7	508
121	181	623	1049	3	22	113	40	588	160	0.05	153	16.1	849

ANALYSIS

Matrix Plot for the given data



Matrix Plot (finals 13v*47c)



Comment on the graph:

- Matrix plot is one of the graphical tools used to see the relationship between several variables. By seeing the above graph we can observe that some variables are correlated.
- Crime rate and education, police expenditure, population are related.

Calculations:

Dependent Variable: CRIME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	5184514	432043	8.66	<.0001
Error	34	1696413	49895		
Corrected Total	46	6880928			

Root MSE	223.37080	R-Square	0.7535
Dependent Mean	905.08511	Adj R-Sq	0.6664
Coeff Var	24.67954		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-6193.41627	1640.63883	-3.78	0.0006
Ed	1	16.30232	6.50048	2.51	0.0171
Poexp	1	4.75097	1.32090	3.60	0.0010
Lf	1	-0.74083	1.30569	-0.57	0.5742
mf	1	3.47787	2.01142	1.73	0.0929
Pop	1	-0.98548	1.37304	-0.72	0.4778
NW	1	0.56134	0.55774	1.01	0.3213
u1	B	-6.59673	4.13239	-1.60	0.1197
u2	B	14.49021	8.45382	1.71	0.0956
GDP	1	0.65045	1.06825	0.61	0.5466
INCI NEQ	1	7.31227	2.23548	3.27	0.0025
PROB	1	-3523.29486	2251.24515	-1.57	0.1268
UNEMP	0	0			
TIME	1	4.14879	6.82938	0.61	0.5476

- Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
- The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

$$UNEMP = u1 + u2$$

By ignoring the variable unemployment rate, the calculations are as follows,

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	5184514	432043	8.66	<.0001
Error	34	1696413	49895		
Corrected Total	46	6880928			

Root MSE	223.37080	R-Square	0.7535
Dependent Mean	905.08511	Adj R-Sq	0.6664
Coeff Var	24.67954		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-6193.41627	1640.63883	-3.78	0.0006
Ed	1	16.30232	6.50048	2.51	0.0171
POEXP	1	4.75097	1.32090	3.60	0.0010
LF	1	-0.74083	1.30569	-0.57	0.5742

mF	1	3.47787	2.01142	1.73	0.0929
Pop	1	-0.98548	1.37304	-0.72	0.4778
NW	1	0.56134	0.55774	1.01	0.3213
u1	1	-6.59673	4.13239	-1.60	0.1197
u2	1	14.49021	8.45382	1.71	0.0956
GDP	1	0.65045	1.06825	0.61	0.5466
INCI NEQ	1	7.31227	2.23548	3.27	0.0025
PROB	1	-3523.29486	2251.24515	-1.57	0.1268
TIME	1	4.14879	6.82938	0.61	0.5476

Interpretation of the results:

ANOVA:

For the regression equation to be valid to be used for predictions, it must reflect the regression model for the population. The hypothesis used to test our assumptions will be based on the following null and alternative hypotheses

$H_0: b_1 = b_2 = \dots = b_p$ against H_1 : At least one of them is not equal to zero.

The results of the test summarized in the ANOVA table are used to reject or not reject the hypothesis that the regression coefficients are valid for prediction.

On the ANOVA table, the F-statistic and the significance F are used to test the validity of the regression. If the calculated F is higher than the significance F, we would reject the null hypothesis and conclude that at least one of independent variable is correlated to the dependent variable.

In this case, the calculated F is 8.66 and significance F is close to zero, therefore, we can reject the null hypothesis and conclude that at least one of the independent variable explains the variations in the dependent variable.

Interpreting R:

Multiple R:

It is used to measure the strength of the relationship between the independent variables and dependent variable.

R-SQUARE:

R-square is the squarer of the multiple R, it is called as coefficient of determination and it measures the proportion in the variation in the Y variable that is explained by the variations in the independent factors.

Note that in this case R-square=0.7535 means that 75.35% of the variations in the crime rates are explained by the independent variables.

Adjusted R-Square:

The adjusted R square takes the factors into account that can contribute to inflating the results; it is given by the formula

$$R^2_{Adj} = 1 - \frac{SS_{Res} / (n-p)}{SS_T / (n-1)}$$

In this case $R^2_{Adj} = 66.64\%$.

Standard Error:

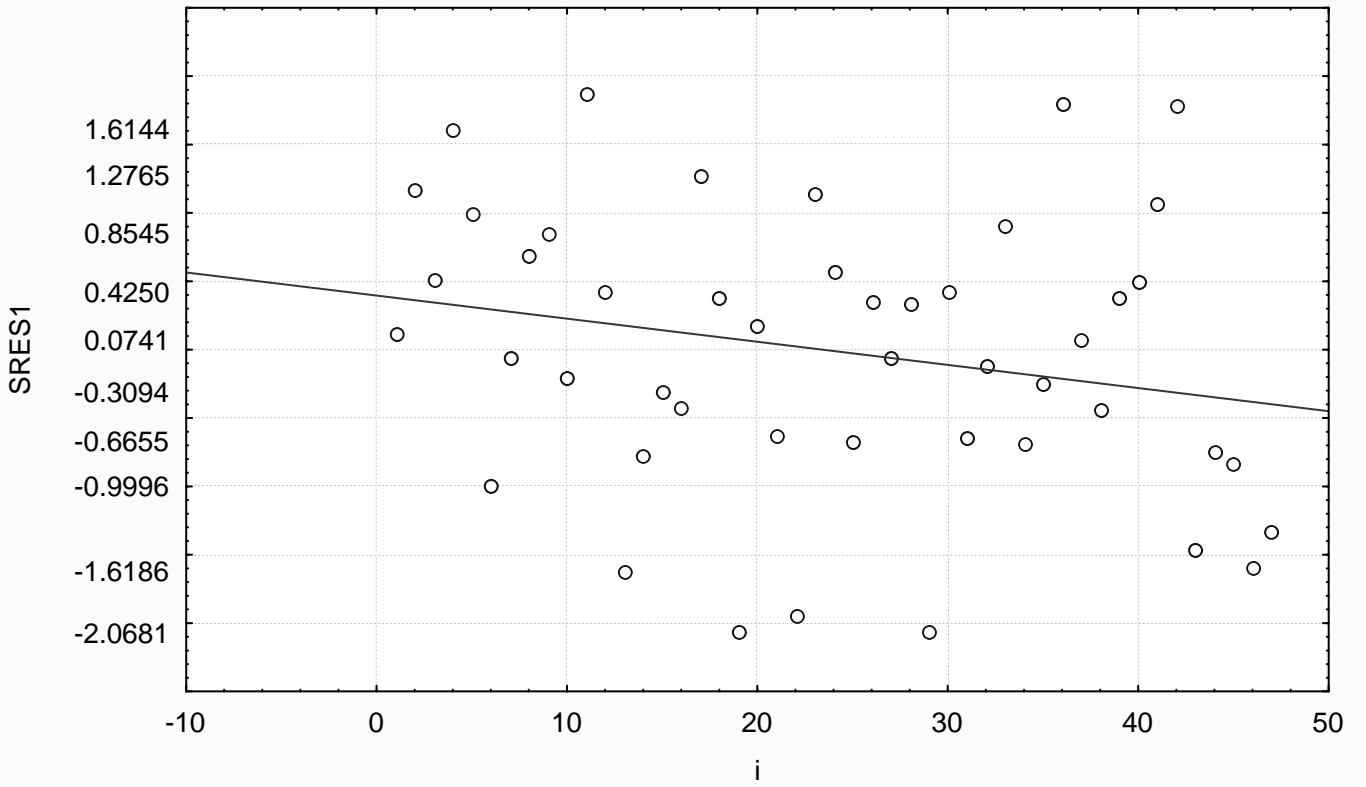
It is the standard deviation of the residuals. The residuals are the difference between predicted dependent variable and actual dependent variable.

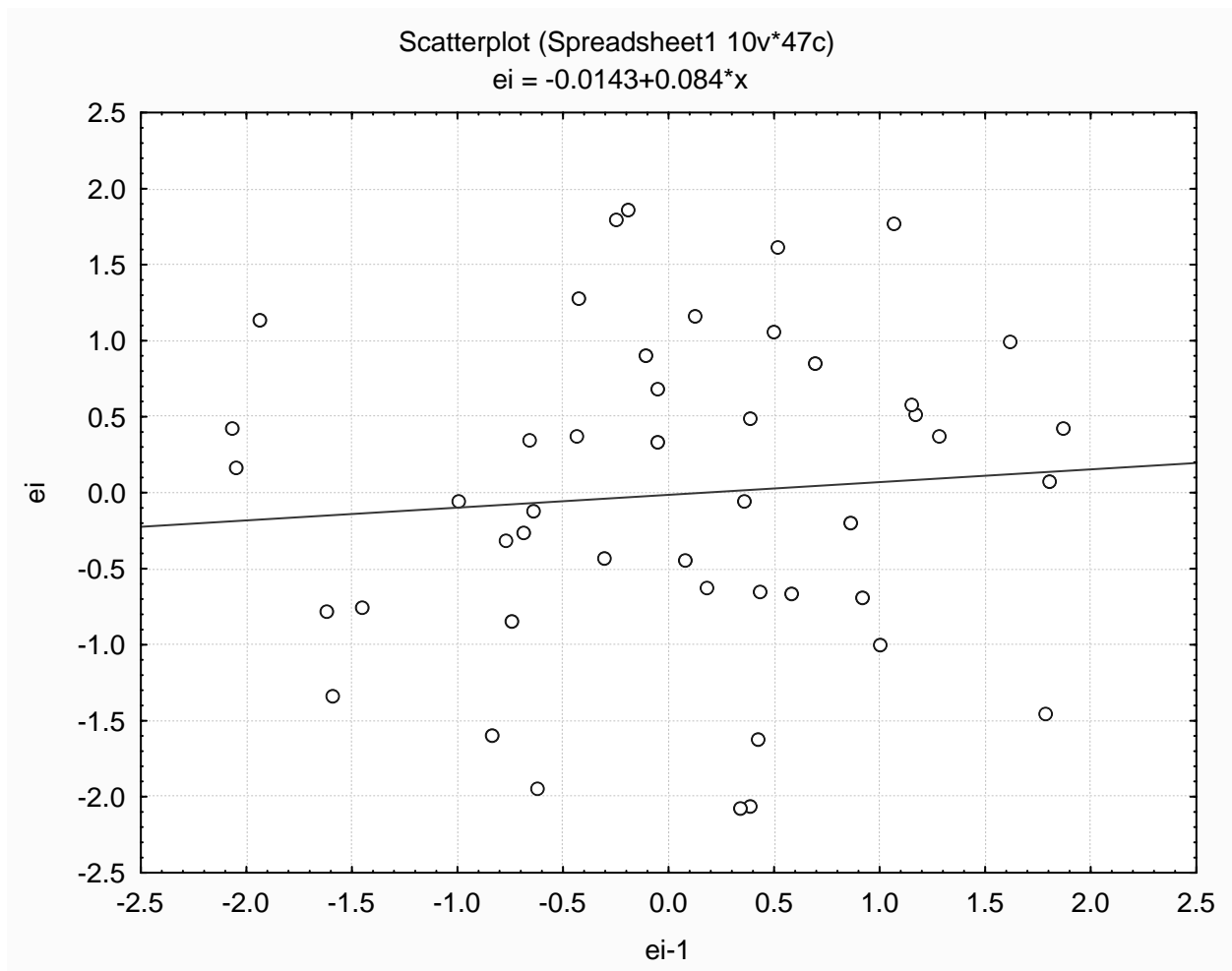
Here it is 223.3708.

RESIDUAL ANALYSIS:

INDEX PLOT

Scatterplot (Spreadsheet1 10v*47c)
SRES1 = 0.3953-0.0169*x



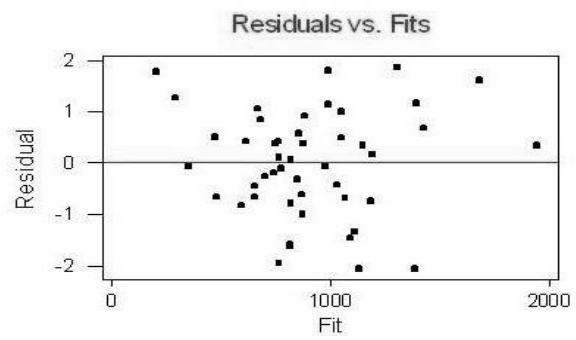
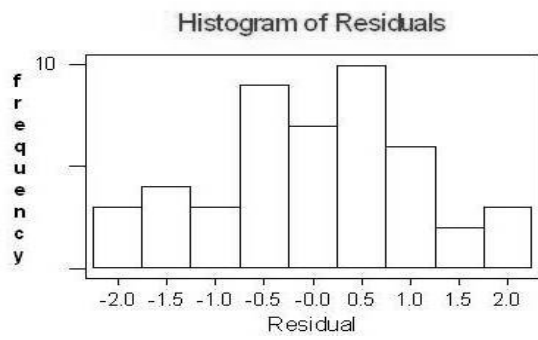
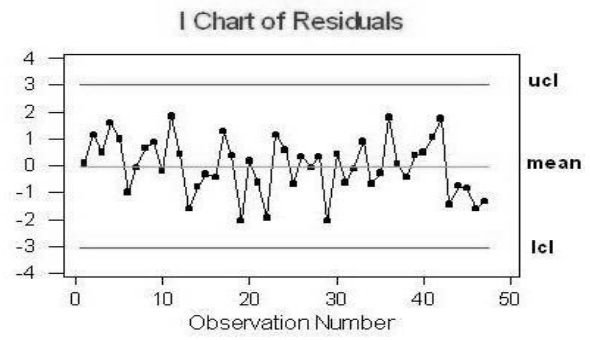
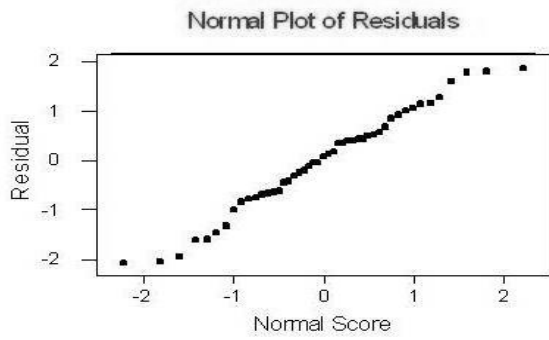


COMMENTS:

INDEX PLOT:

- This is the scatter plot of error Vs I showing random scatter with out any pattern with in 3-sigma limits around zero implies that uncorrelated structure and homoscedasticity of residuals.
- By observing second plot we can say that there is no serial correlation in e_i which represents uncorrelated structure in y_i 's.
- By observing residual plots e_i 's are not showing normality and graph of residuals and fitted values is not showing random pattern which implies that we need improvement in the model.

PLOT OF STD RESIDUALS AND FITTED VALUES



DELETION DIAGNOSTICS:

TABLE OF RESIDUALS, FITTED VALUES, LEVERAGES, COOK'S DISTANCES, DELETED RESIDUALS:

SL.NO	RESIDUALS	STD RESI	DEL RESI	LEVERAGES	COOK'SDIST	DFFITs	FITS
1	23.432	0.11733	0.11561	0.200568	0.000266	0.05791	767.57
2	244.909	1.16549	1.17187	0.115014	0.01358	0.42246	1390.09
3	105.338	0.51785	0.5122	0.170714	0.004246	0.23239	472.66
4	287.588	1.61435	1.65513	0.36395	0.114711	1.25201	1681.41
5	184.926	0.99803	0.99797	0.311898	0.03473	0.67189	1049.07
6	-189.33	-0.99957	-0.99956	0.28095	0.03003	-0.6248	871.33
7	-10.342	-0.05821	-0.05735	0.36735	0.000151	-0.0437	973.34
8	127.03	0.68637	0.68093	0.313486	0.016548	0.46014	1427.97
9	176.831	0.8545	0.85103	0.141703	0.009273	0.34579	679.17
10	-40.634	-0.19886	-0.19603	0.163164	0.000593	-0.08656	745.63
11	368.02	1.86679	1.94131	0.221069	0.076081	1.03421	1305.98
12	88.384	0.425	0.41982	0.133206	0.002135	0.16458	760.62
13	-302.199	-1.61857	-1.65982	0.301335	0.086916	-1.09006	813.2
14	-156.591	-0.77743	-0.77281	0.186863	0.010684	-0.37047	820.59
15	-54.482	-0.30943	-0.30528	0.378669	0.004489	-0.23832	852.48
16	-81.973	-0.43013	-0.42492	0.272081	0.00532	-0.25978	1027.97
17	249.607	1.27652	1.28888	0.233695	0.038226	0.71176	289.39
18	52.607	0.38262	0.37777	0.621127	0.018462	0.48369	876.39
19	-384.308	-2.05478	-2.16309	0.298906	0.138466	-1.41239	1134.31
20	32.801	0.17371	0.17121	0.285363	0.000927	0.10819	1192.2
21	-123.626	-0.62161	-0.61591	0.207252	0.007771	-0.31492	865.63
22	-328.126	-1.94277	-2.02997	0.428277	0.217489	-1.75695	767.13
23	222.305	1.1459	1.15137	0.245688	0.032899	0.6571	993.7
24	112.707	0.57767	0.57192	0.237058	0.007976	0.3188	855.29
25	-128.993	-0.6655	-0.65995	0.247024	0.011177	-0.378	651.99
26	54.834	0.35073	0.34616	0.51011	0.009853	0.35324	1938.17
27	-10.714	-0.05495	-0.05414	0.238063	0.000073	-0.03026	352.71
28	67.156	0.33846	0.33401	0.210943	0.002356	0.1727	1148.84
29	-340.132	-2.0681	-2.17913	0.457874	0.277873	-2.00265	1383.13
30	83.243	0.42918	0.42397	0.246012	0.004623	0.24218	612.76
31	-106.567	-0.64657	-0.64094	0.455542	0.026906	-0.58628	479.57
32	-21.846	-0.11249	-0.11084	0.244097	0.000314	-0.06299	775.85
33	190.006	0.91028	0.90792	0.12675	0.009252	0.3459	881.99
34	-144.895	-0.68733	-0.6819	0.10932	0.00446	-0.2389	1067.89
35	-47.855	-0.25406	-0.25053	0.288893	0.002017	-0.15969	700.86
36	281.441	1.80074	1.86523	0.510422	0.260055	1.90452	990.56
37	11.936	0.07407	0.07298	0.479513	0.000389	0.07005	819.06
38	-87.738	-0.43618	-0.43093	0.189072	0.003412	-0.20808	653.74
39	79.181	0.38161	0.37676	0.137114	0.00178	0.15019	746.82
40	100.846	0.49781	0.49223	0.177483	0.004113	0.22865	1050.15
41	210.253	1.06304	1.06514	0.215961	0.023944	0.55902	669.75
42	334.591	1.77932	1.84075	0.291289	0.100096	1.18011	207.41
43	-267.831	-1.45507	-1.48034	0.320947	0.076976	-1.01772	1090.83
44	-154.429	-0.74457	-0.73959	0.137829	0.006817	-0.29571	1184.43
45	-138.569	-0.83509	-0.83129	0.448162	0.043566	-0.74914	593.57
46	-305.325	-1.59193	-1.63028	0.262737	0.069471	-0.97322	813.33
47	-263.466	-1.33165	-1.34753	0.215458	0.037461	-0.70618	1112.47

COMMENTS:

There are two observations in the data that are influencing the fit. They are 19 and 29. after deleting those two observations the results are as follows,

The regression equation is

CRIME = - 3812 + 17.0 Ed + 6.81 PoEXP - 1.15 LF + 1.87 Mf - 0.38 Pop + 0.202 NW- 5.56 U1 + 11.0 U2 - 0.612 GDP + 6.37 INC - 4300 PROB + 2.72 TIME

Predictor	Coef	SE Coef	T	P
Constant	-3812	1565	-2.44	0.021
Ed	16.978	5.674	2.99	0.005
PoEXP	6.814	1.279	5.33	0.000
LF	-1.153	1.129	-1.02	0.315
Mf	1.872	1.800	1.04	0.306
Pop	-0.384	1.254	-0.31	0.761
NW	0.2018	0.4922	0.41	0.685
U1	-5.564	3.682	-1.51	0.141
U2	11.034	7.689	1.44	0.161
GDP	-0.6120	0.9790	-0.63	0.536
INC	6.375	1.940	3.29	0.002
PROB	-4300	1988	-2.16	0.038
TIME	2.718	5.993	0.45	0.653

S = 192.1 R-Sq = 82.7% R-Sq(adj) = **76.3%**

Analysis of Variance

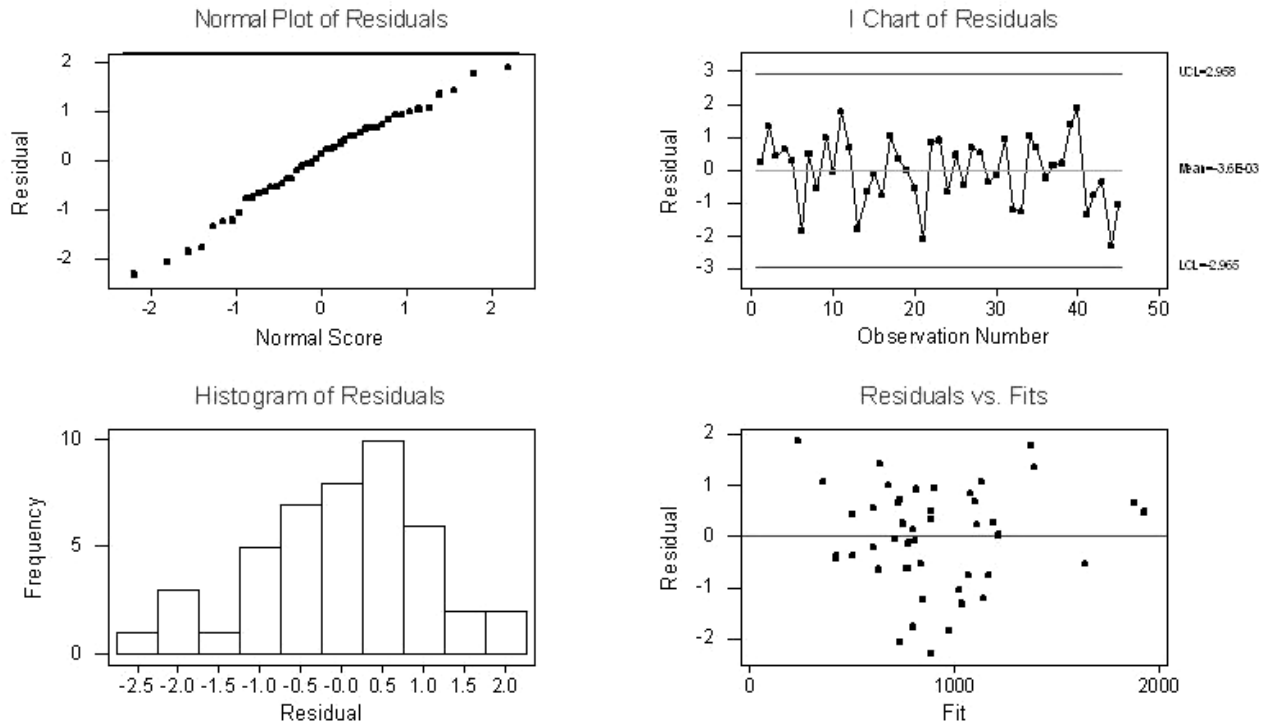
Source	DF	SS	MS	F	P
Regression	12	5656770	471397	12.77	0.000
Residual Error	32	1181080	36909		
Total	44	6837849			

Source	DF	Seq SS
Ed	1	764181
PoEXP	1	3276631
LF	1	37519
Mf	1	52042
Pop	1	66730
NW	1	255599
U1	1	51218
U2	1	128147
GDP	1	257795
INC	1	447162
PROB	1	312152
TIME	1	7594

COMMENTS:

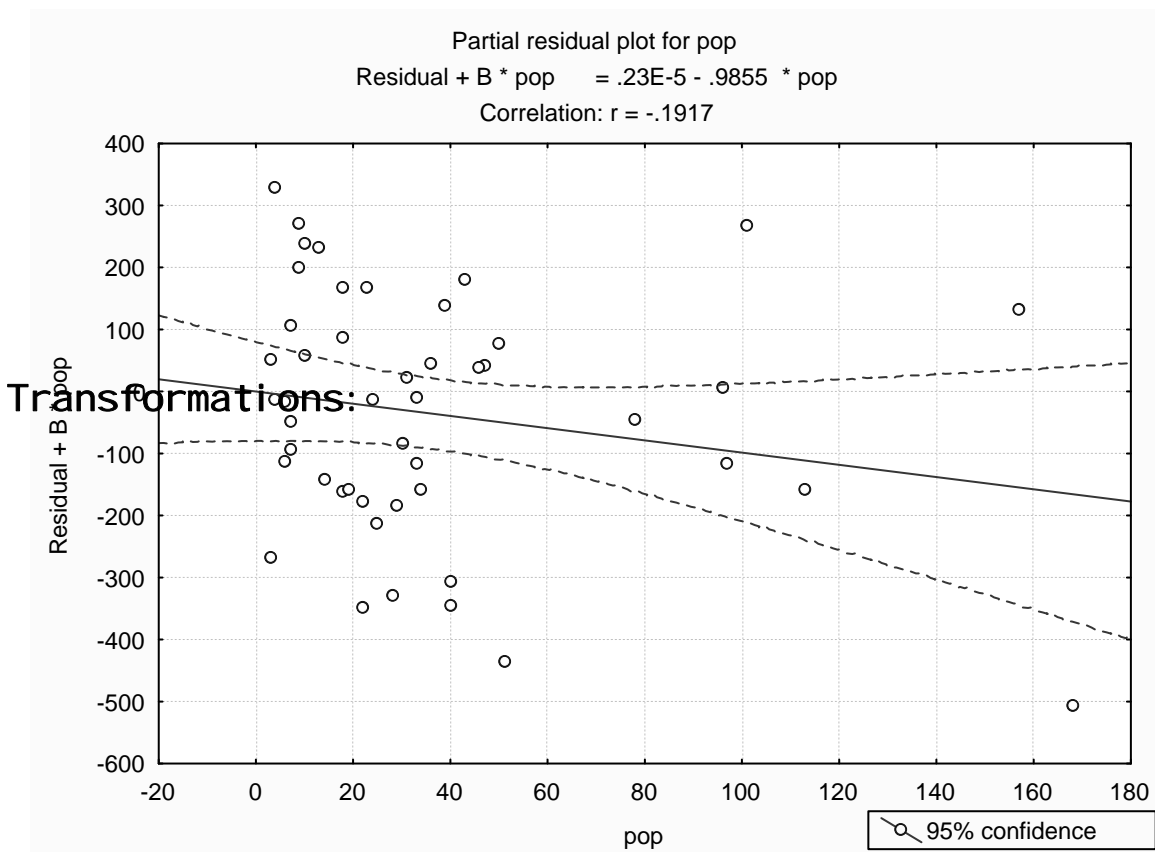
After deleting influential observations the residuals are showing normality and graph of standardized residuals and fitted values showing randomness in their pattern.

PLOT OF STD RESIDUALS AND FITTED VALUES



COMMENTS:

After deleting influential observations the residuals are showing normality and graph of standardized residuals and fitted values showing randomness in their pattern.



Comments:

- After plotting partial residuals plots for all regressor, it can be observed that the regressor 'population (pop)' and 'non white people (n w)' are showing non linear pattern .so they need some transformation in order to convert them to show linear pattern.
- After trying different patterns it is observed that the log to the base ten transformations for the regressor pop is better. And natural log transformations for the regressor 'nw' is better.

The regression equation is

$$\text{CRIME} = - 4947 + 16.3 \text{ Ed} + 4.86 \text{ PoEXP} - 0.34 \text{ LF} + 1.88 \text{ Mf} - 5.68 \text{ U1} + 13.1 \text{ U2} \\ + 0.63 \text{ GDP} + 7.52 \text{ INC} - 3000 \text{ PROB} + 5.81 \text{ TIME} + 130 \text{ log10NW} \\ - 83.5 \text{ LOGePOP}$$

Predictor	Coef	SE Coef	T	P
Constant	-4947	1794	-2.76	0.009
Ed	16.261	6.370	2.55	0.015
PoEXP	4.863	1.183	4.11	0.000
LF	-0.340	1.300	-0.26	0.795
Mf	1.880	2.237	0.84	0.407
U1	-5.675	4.073	-1.39	0.173
U2	13.058	8.356	1.56	0.127
GDP	0.633	1.068	0.59	0.557
INC	7.518	2.529	2.97	0.005
PROB	-3000	2083	-1.44	0.159
TIME	5.808	6.426	0.90	0.372
log10NW	129.8	105.4	1.23	0.227
LOGePOP	-83.46	56.24	-1.48	0.147

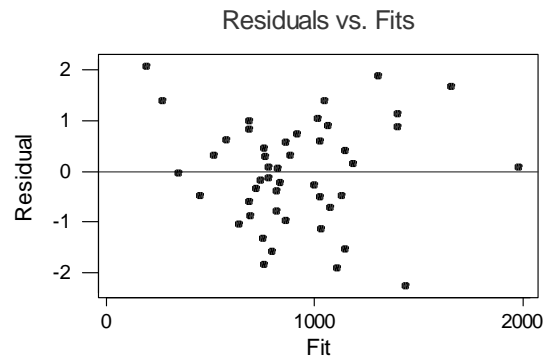
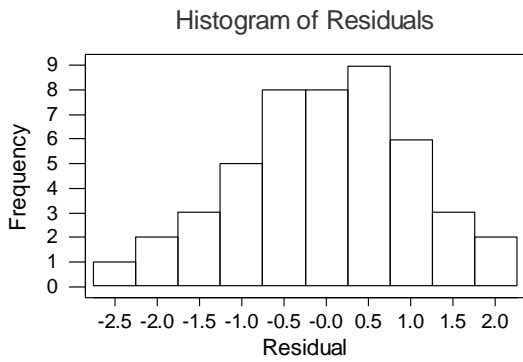
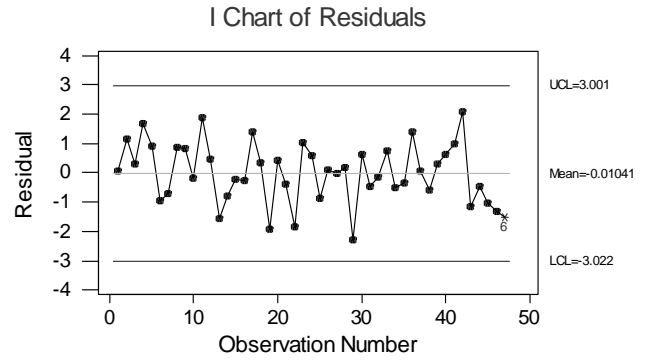
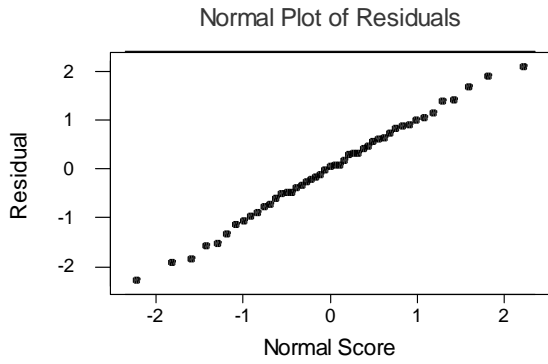
S = 218.8 R-Sq = 76.4% R-Sq(adj) = 68.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	12	5253818	437818	9.15	0.000
Residual Error	34	1627109	47856		
Total	46	6880928			

Source	DF	Seq SS
Ed	1	717146
PoEXP	1	2452248
LF	1	148491
Mf	1	254590
U1	1	51247
U2	1	300778
GDP	1	252209
INC	1	713402
PROB	1	186175
TIME	1	26388
log10NW	1	45742
LOGePOP	1	105402

PLOT OF STD RESIDUALS AND FITTED VALUES



Comments:

- The p-p plot of residuals shows normality which can also be observed from the histogram
- Random scatter in residuals vs. fits

Selection of variables:

1 stepwise method:

Dependent Variable: CRIME

Stepwise Selection: Step 1

Variable PoEXP Entered: R-Square = 0.4605 and C(p) = 31.4089

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3168330	3168330	38.40	<.0001
Error	45	3712597	82502		
Corrected Total	46	6880928			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	152.06737	128.53313	115480	1.40	0.2430
PoEXP	4.55728	0.73540	3168330	38.40	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable INC Entered: R-Square = 0.5709 and C(p) = 18.1797

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3928185	1964093	29.27	<.0001
Error	44	2952743	67108		
Corrected Total	46	6880928			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	-968.59581	352.63817	506290	7.54	0.0087
PoEXP	6.41662	0.86326	3707655	55.25	<.0001
INC	4.19298	1.24607	759855	11.32	0.0016

Stepwise Selection: Step 2

Bounds on condition number: 1.6941, 6.7763

Stepwise Selection: Step 3

Variable Ed Entered: R-Square = 0.6550 and C(p) = 8.5816

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4506867	1502289	27.21	<.0001
Error	43	2374061	55211		
Corrected Total	46	6880928			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	-3280.25888	782.39781	970475	17.58	0.0001
Ed	15.67394	4.84139	578682	10.48	0.0023
PoEXP	6.41810	0.78301	3709362	67.19	<.0001
INC	7.57262	1.53856	1337472	24.22	<.0001

Bounds on condition number: 3.1392, 21.832

Stepwise Selection: Step 4

Variable PROB Entered: R-Square = 0.6850 and C(p) = 6.4464

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4713192	1178298	22.83	<.0001
Error	42	2167736	51613		
Corrected Total	46	6880928			

Stepwise Selection: Step 4

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	-3054.19099	764.87816	822934	15.94	0.0003
Ed	15.03030	4.69204	529626	10.26	0.0026
PoEXP	5.98973	0.78680	2991183	57.95	<.0001
INC	7.96847	1.50070	1455182	28.19	<.0001
PROB	-3489.71060	1745.38817	206325	4.00	0.0521

Bounds on condition number: 3.1949, 35.488

Stepwise Selection: Step 5

Variable Mf Entered: R-Square = 0.7050 and C(p) = 5.6772

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4851358	970272	19.60	<.0001
Error	41	2029570	49502		
Corrected Total	46	6880928			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	-4658.36481	1217.82253	724301	14.63	0.0004
Ed	10.69832	5.27619	203521	4.11	0.0491
PoEXP	6.13296	0.77530	3097601	62.58	<.0001
Mf	2.17825	1.30382	138166	2.79	0.1024
INC	7.44565	1.50264	1215386	24.55	<.0001
PROB	-3526.11850	1709.45975	210618	4.25	0.0455

Bounds on condition number: 3.3397, 55.967

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PoEXP		1	0.4605	0.4605	31.4089	38.40	<.0001
2	INC		2	0.1104	0.5709	18.1797	11.32	0.0016
3	Ed		3	0.0841	0.6550	8.5816	10.48	0.0023
4	PROB		4	0.0300	0.6850	6.4464	4.00	0.0521
5	Mf		5	0.0201	0.7050	5.6772	2.79	0.1024

Analysis with the selected variables:

Dependent Variable: CRIME

Analysis of Variance

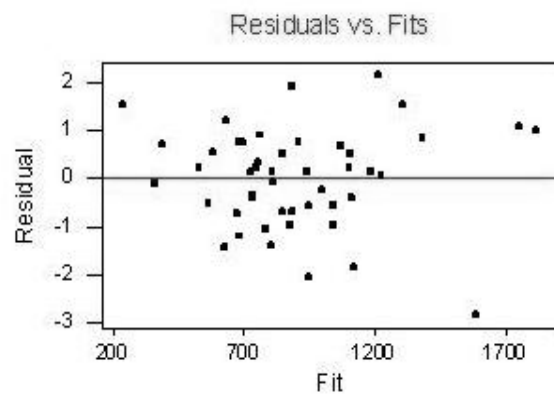
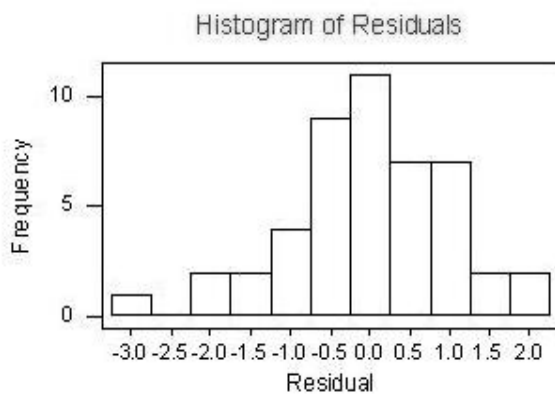
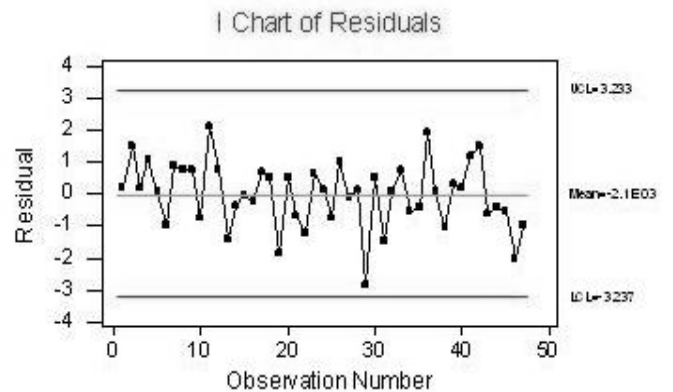
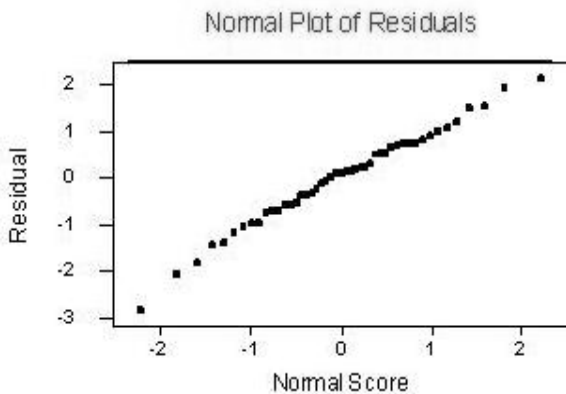
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4851358	970272	19.60	<.0001
Error	41	2029570	49502		
Corrected Total	46	6880928			

Root MSE	222.48979	R-Square	0.7050
Dependent Mean	905.08511	Adj R-Sq	0.6691
Coeff Var	24.58220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-4658.36481	1217.82253	-3.83	0.0004
Ed	1	10.69832	5.27619	2.03	0.0491
PoEXP	1	6.13296	0.77530	7.91	<.0001
Mf	1	2.17825	1.30382	1.67	0.1024
INC	1	7.44565	1.50264	4.96	<.0001
PROB	1	-3526.11850	1709.45975	-2.06	0.0455

standardised residual plots



Comments:

- The p-p plot of residuals shows normality which can also be observed from histogram
- Non random scatter in residuals Vs fits

Which implies that model with these regressors is not advisable for the data.

2. R² selection method:

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	Variables in Model
1	0.4605	0.4485	31.4089	534.0234	PoEXP
1	0.1948	0.1769	68.0498	552.8424	GDP
1	0.1796	0.1614	70.1377	553.7178	PROB
1	0.1139	0.0942	79.2032	557.3406	Pop
1	0.1042	0.0843	80.5363	557.8505	Ed
1	0.0458	0.0246	88.5989	560.8220	Mf
1	0.0357	0.0142	89.9902	561.3163	LF
1	0.0320	0.0105	90.4896	561.4925	INC
1	0.0314	0.0099	90.5733	561.5219	U2
1	0.0225	0.0007	91.8125	561.9560	TIME
1	0.0025	-0.0196	94.5581	562.9036	U1
1	0.0011	-0.0211	94.7629	562.9735	NW

2	0.5709	0.5514	18.1797	525.2606	PoEXP INC
2	0.4983	0.4755	28.1840	532.6015	PoEXP Mf
2	0.4943	0.4713	28.7462	532.9819	PoEXP NW
2	0.4852	0.4618	29.9976	533.8177	PoEXP GDP
2	0.4730	0.4490	31.6785	534.9175	PoEXP LF
2	0.4722	0.4482	31.7900	534.9896	PoEXP PROB
2	0.4684	0.4442	32.3136	535.3265	PoEXP TIME
2	0.4638	0.4394	32.9503	535.7329	PoEXP U2
2	0.4608	0.4363	33.3599	535.9924	PoEXP Pop
2	0.4608	0.4363	33.3635	535.9947	PoEXP U1
2	0.4606	0.4361	33.3876	536.0100	Ed PoEXP
2	0.3987	0.3713	41.9288	541.1186	GDP INC

3	0.6550	0.6309	8.5816	517.0083	Ed PoEXP INC
3	0.6406	0.6155	10.5641	518.9269	PoEXP Mf INC
3	0.6127	0.5857	14.4086	522.4380	PoEXP LF INC
3	0.6080	0.5806	15.0613	523.0089	PoEXP INC PROB
3	0.6050	0.5774	15.4789	523.3706	PoEXP GDP INC
3	0.5879	0.5592	17.8286	525.3552	PoEXP Pop INC
3	0.5771	0.5476	19.3244	526.5763	PoEXP NW INC
3	0.5714	0.5415	20.1076	527.2033	PoEXP U1 INC
3	0.5713	0.5414	20.1228	527.2154	PoEXP INC TIME
3	0.5709	0.5409	20.1796	527.2605	PoEXP U2 INC
3	0.5673	0.5371	20.6776	527.6544	PoEXP Mf NW
3	0.5444	0.5126	23.8344	530.0770	PoEXP Mf GDP

4	0.6850	0.6550	6.4464	514.7352	Ed PoEXP INC PROB
4	0.6755	0.6446	7.7562	516.1312	PoEXP Mf INC PROB
4	0.6744	0.6434	7.8985	516.2804	Ed PoEXP Mf INC
4	0.6688	0.6372	8.6773	517.0886	Ed PoEXP GDP INC
4	0.6663	0.6345	9.0213	517.4413	Ed PoEXP INC TIME
4	0.6653	0.6334	9.1580	517.5806	Ed PoEXP U2 INC
4	0.6604	0.6281	9.8279	518.2579	PoEXP Mf GDP INC
4	0.6603	0.6280	9.8468	518.2769	PoEXP Mf INC TIME
4	0.6597	0.6272	9.9369	518.3672	Ed PoEXP Pop INC
4	0.6571	0.6244	10.2912	518.7206	Ed PoEXP LF INC
4	0.6564	0.6236	10.3919	518.8206	Ed PoEXP U1 INC
4	0.6553	0.6225	10.5344	518.9617	Ed PoEXP NW INC

5	0.7050	0.6691	5.6772	513.6398	Ed PoEXP Mf INC PROB
5	0.7000	0.6634	6.3694	514.4328	Ed PoEXP Mf INC TIME
5	0.6978	0.6610	6.6723	514.7757	Ed PoEXP Pop INC PROB
5	0.6929	0.6555	7.3473	515.5309	Ed PoEXP U2 INC PROB
5	0.6910	0.6533	7.6202	515.8328	Ed PoEXP GDP INC PROB
5	0.6901	0.6523	7.7381	515.9627	Ed PoEXP NW INC PROB
5	0.6863	0.6480	8.2660	516.5396	Ed PoEXP Mf GDP INC
5	0.6858	0.6475	8.3351	516.6146	Ed PoEXP U1 INC PROB
5	0.6853	0.6469	8.4020	516.6872	Ed PoEXP LF INC PROB
5	0.6853	0.6469	8.4037	516.6889	Ed PoEXP INC PROB TIME
5	0.6848	0.6464	8.4696	516.7603	PoEXP Mf GDP INC PROB
5	0.6819	0.6432	8.8623	517.1830	Ed PoEXP Pop INC TIME

6	0.7139	0.6710	6.4585	514.2101	Ed	PoEXP	Mf	NW	INC	PROB						
6	0.7108	0.6674	6.8848	514.7152	Ed	PoEXP	Mf	INC	PROB	TIME						
6	0.7096	0.6661	7.0443	514.9028	Ed	PoEXP	Mf	GDP	INC	PROB						
6	0.7088	0.6651	7.1641	515.0431	Ed	PoEXP	Mf	U2	INC	PROB						
6	0.7083	0.6646	7.2224	515.1114	Ed	PoEXP	Mf	Pop	INC	PROB						
6	0.7078	0.6640	7.2964	515.1977	Ed	PoEXP	Pop	U2	INC	PROB						
6	0.7062	0.6621	7.5176	515.4549	Ed	PoEXP	Mf	U1	INC	PROB						
6	0.7062	0.6621	7.5231	515.4614	Ed	PoEXP	Mf	GDP	INC	TIME						
6	0.7059	0.6618	7.5616	515.5060	Ed	PoEXP	Pop	GDP	INC	PROB						
6	0.7058	0.6617	7.5727	515.5188	Ed	PoEXP	Mf	U1	U2	INC						
6	0.7057	0.6615	7.5904	515.5394	Ed	PoEXP	LF	Mf	INC	PROB						
6	0.7050	0.6607	7.6901	515.6547	Ed	PoEXP	Mf	U2	INC	TIME						

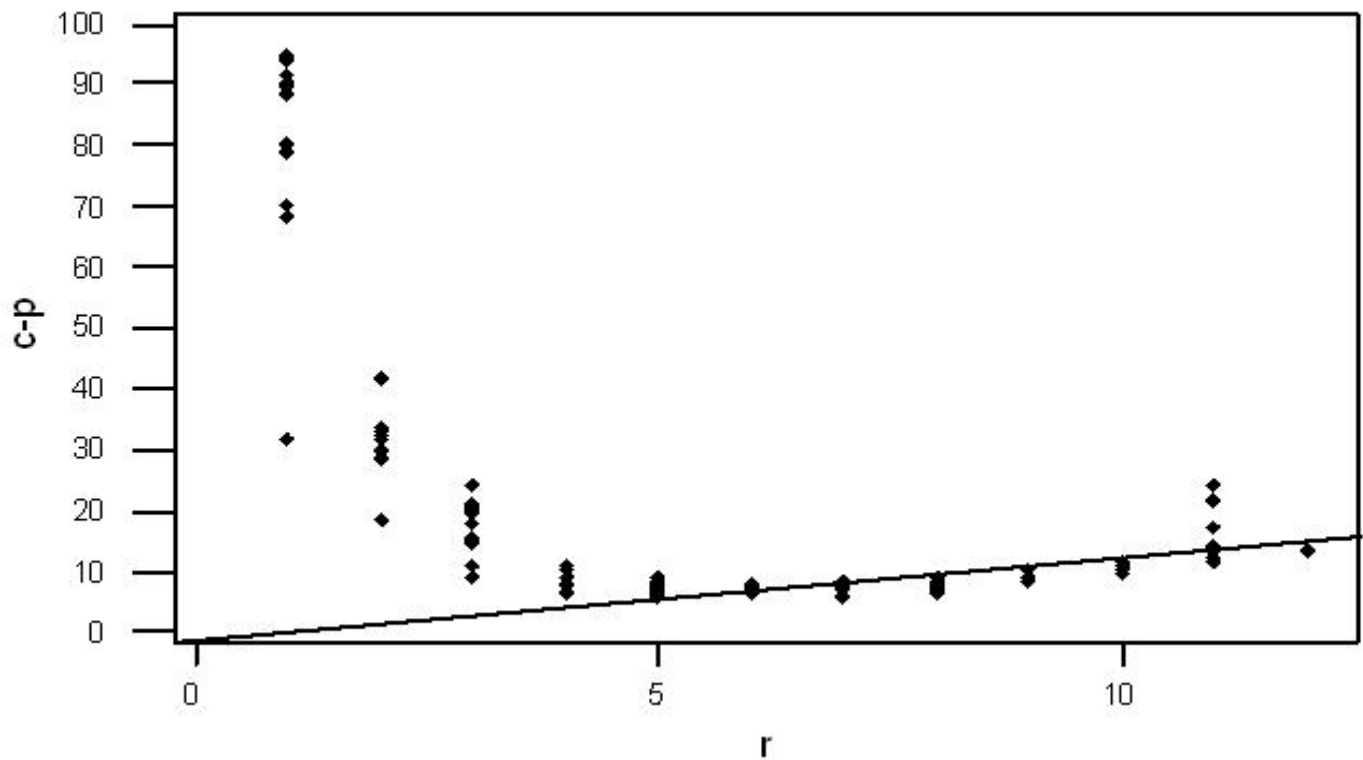
7	0.7342	0.6865	5.6509	512.7410	Ed	PoEXP	Mf	U1	U2	INC	PROB					
7	0.7258	0.6766	6.8090	514.2031	Ed	PoEXP	Mf	U1	U2	INC	TIME					
7	0.7212	0.6712	7.4478	514.9906	Ed	PoEXP	Mf	NW	GDP	INC	PROB					
7	0.7185	0.6680	7.8166	515.4392	Ed	PoEXP	Mf	NW	U2	INC	PROB					
7	0.7182	0.6676	7.8596	515.4913	Ed	PoEXP	Pop	U1	U2	INC	PROB					
7	0.7169	0.6661	8.0424	515.7118	Ed	PoEXP	Mf	NW	INC	PROB	TIME					
7	0.7168	0.6660	8.0574	515.7299	Ed	PoEXP	Mf	Pop	INC	PROB	TIME					
7	0.7168	0.6659	8.0610	515.7342	Ed	PoEXP	Mf	Pop	NW	INC	PROB					
7	0.7152	0.6641	8.2712	515.9865	Ed	PoEXP	Pop	NW	U2	INC	PROB					
7	0.7149	0.6637	8.3175	516.0418	Ed	PoEXP	LF	Mf	NW	INC	PROB					
7	0.7147	0.6635	8.3452	516.0750	Ed	PoEXP	Mf	U2	INC	PROB	TIME					
7	0.7146	0.6634	8.3535	516.0849	Ed	PoEXP	Pop	NW	GDP	INC	PROB					
8	0.7430	0.6889	6.4450	513.1686	Ed	PoEXP	Mf	NW	U1	U2	INC	PROB				
8	0.7375	0.6823	7.1945	514.1520	Ed	PoEXP	Mf	U1	U2	INC	PROB	TIME				
8	0.7374	0.6821	7.2199	514.1850	Ed	PoEXP	Mf	Pop	U1	U2	INC	PROB				
8	0.7367	0.6813	7.3087	514.3000	Ed	PoEXP	LF	Mf	U1	U2	INC	PROB				
8	0.7350	0.6792	7.5464	514.6067	Ed	PoEXP	Mf	U1	U2	GDP	INC	PROB				
8	0.7301	0.6733	8.2216	515.4672	Ed	PoEXP	Mf	Pop	U1	U2	INC	TIME				
8	0.7276	0.6703	8.5647	515.8985	Ed	PoEXP	Mf	U1	U2	GDP	INC	TIME				
8	0.7276	0.6702	8.5714	515.9069	Ed	PoEXP	LF	Mf	U1	U2	INC	TIME				
8	0.7272	0.6698	8.6208	515.9686	Ed	PoEXP	Mf	NW	U1	U2	INC	TIME				
8	0.7257	0.6679	8.8318	516.2315	Ed	PoEXP	Mf	Pop	NW	GDP	INC	PROB				
8	0.7243	0.6662	9.0254	516.4714	Ed	PoEXP	Pop	NW	U1	U2	INC	PROB				
8	0.7243	0.6662	9.0275	516.4740	Ed	PoEXP	Mf	NW	U2	GDP	INC	PROB				

9	0.7459	0.6841	8.0453	514.6355	Ed	PoEXP	Mf	Pop	NW	U1	U2	INC	PROB			
9	0.7458	0.6840	8.0528	514.6456	Ed	PoEXP	LF	Mf	NW	U1	U2	INC	PROB			
9	0.7450	0.6829	8.1695	514.8019	Ed	PoEXP	Mf	NW	U1	U2	GDP	INC	PROB			
9	0.7443	0.6821	8.2613	514.9243	Ed	PoEXP	Mf	NW	U1	U2	INC	PROB	TIME			
9	0.7429	0.6804	8.4504	515.1757	Ed	PoEXP	Mf	Pop	U1	U2	INC	PROB	TIME			
9	0.7408	0.6777	8.7525	515.5746	Ed	PoEXP	LF	Mf	U1	U2	INC	PROB	TIME			
9	0.7389	0.6753	9.0144	515.9176	Ed	PoEXP	LF	Mf	Pop	U1	U2	INC	PROB			
9	0.7386	0.6751	9.0428	515.9546	Ed	PoEXP	Mf	Pop	U1	U2	GDP	INC	PROB			
9	0.7381	0.6744	9.1146	516.0481	Ed	PoEXP	Mf	U1	U2	GDP	INC	PROB	TIME			
9	0.7377	0.6739	9.1691	516.1191	Ed	PoEXP	LF	Mf	U1	U2	GDP	INC	PROB			
9	0.7328	0.6678	9.8514	516.9973	Ed	PoEXP	Mf	Pop	U1	U2	GDP	INC	TIME			
9	0.7310	0.6656	10.0979	517.3108	Ed	PoEXP	Mf	Pop	NW	U1	U2	INC	TIME			

10	0.7487	0.6789	9.6537	516.1074	Ed	PoEXP	Mf	Pop	NW	U1	U2	INC	PROB	TIME		
10	0.7487	0.6789	9.6612	516.1176	Ed	PoEXP	Mf	Pop	NW	U1	U2	GDP	INC	PROB		
10	0.7483	0.6784	9.7110	516.1850	Ed	PoEXP	LF	Mf	NW	U1	U2	GDP	INC	PROB		
10	0.7477	0.6776	9.7949	516.2985	Ed	PoEXP	LF	Mf	Pop	NW	U1	U2	INC	PROB		
10	0.7476	0.6775	9.8049	516.3120	Ed	PoEXP	LF	Mf	NW	U1	U2	INC	PROB	TIME		
10	0.7460	0.6754	10.0293	516.6140	Ed	PoEXP	Mf	NW	U1	U2	GDP	INC	PROB	TIME		
10	0.7448	0.6739	10.1948	516.8356	Ed	PoEXP	LF	Mf	Pop	U1	U2	INC	PROB	TIME		
10	0.7441	0.6730	10.2913	516.9642	Ed	PoEXP	Mf	Pop	U1	U2	GDP	INC	PROB	TIME		
10	0.7416	0.6698	10.6402	517.4267	Ed	PoEXP	LF	Mf	U1	U2	GDP	INC	PROB	TIME		
10	0.7403	0.6682	10.8153	517.6570	Ed	PoEXP	LF	Mf	Pop	U1	U2	GDP	INC	PROB		
10	0.7349	0.6613	11.5579	518.6216	Ed	PoEXP	Mf	Pop	NW	U2	GDP	INC	PROB	TIME		
10	0.7346	0.6609	11.5959	518.6704	Ed	PoEXP	Mf	Pop	NW	U1	U2	GDP	INC	TIME		

11	0.7511	0.6729	11.3219	517.6552	Ed	PoEXP	Mf	Pop	NW	U1	U2	GDP	INC	PROB	TIME	
11	0.7508	0.6725	11.3690	517.7197	Ed	PoEXP	LF	Mf	Pop	NW	U1	U2	GDP	INC	PROB	
11	0.7508	0.6724	11.3707	517.7221	Ed	PoEXP	LF	Mf	Pop	NW	U1	U2	INC	PROB	TIME	
11	0.7497	0.6711	11.5151	517.9191	Ed	PoEXP	LF	Mf	NW	U1	U2	GDP	INC	PROB	TIME	
11	0.7461	0.6663	12.0129	518.5921	Ed	PoEXP	LF	Mf	Pop	U1	U2	GDP	INC	PROB	TIME	
11	0.7357	0.6526	13.4494	520.4818	Ed	PoEXP	LF	Mf	Pop	NW	U1	U2	GDP	INC	TIME	
11	0.7350	0.6517	13.5483	520.6092	Ed	PoEXP	LF	Mf	Pop	NW	U2	GDP	INC	PROB	TIME	
11	0.7322	0.6480	13.9379	521.1076	Ed	PoEXP	LF	Mf	Pop	NW	U1	GDP	INC	PROB	TIME	
11	0.7318	0.6475	13.9897	521.1734	Ed	PoEXP	LF	Pop	NW	U1	U2	GDP	INC	PROB	TIME	
11	0.7079	0.6160	17.2894	525.1895	PoEXP	LF	Mf	Pop	NW	U1	U2	GDP	INC	PROB	TIME	
11	0.6759	0.5740	21.6995	530.0716	Ed	PoEXP	LF	Mf	Pop	NW	U1	U2	GDP	PROB	TIME	
11	0.6597	0.5527	23.9367	532.3670	Ed	LF	Mf	Pop	NW	U1	U2	GDP	INC	PROB	TIME	

12	0.7535	0.6664	13.0000	519.2123	Ed	PoEXP	LF	Mf	Pop	NW	U1	U2	GDP	INC	PROB	TIME



Comments:

By observing above results and graph of r and C-p the best set of regressors is Ed, PoEXP, LF, Mf, Pop, U1, U2, INC, PROB

Analysis with the best set of regressors:

The regression equation is

$$\text{CRIME} = - 5438 + 14.9 \text{ Ed} + 5.37 \text{ PoEXP} - 0.59 \text{ LF} + 3.21 \text{ Mf} - 0.72 \text{ Pop} - 7.46 \text{ U1} + 16.2 \text{ U2} + 7.35 \text{ INC} - 3848 \text{ PROB}$$

Predictor	Coef	SE Coef	T	P
Constant	-5438	1459	-3.73	0.001
Ed	14.923	6.098	2.45	0.019
PoEXP	5.371	1.084	4.96	0.000
LF	-0.590	1.283	-0.46	0.649
Mf	3.208	1.944	1.65	0.107
Pop	-0.715	1.301	-0.55	0.586
U1	-7.457	3.978	-1.87	0.069
U2	16.183	8.041	2.01	0.051
INC	7.349	1.596	4.60	0.000
PROB	-3848	1783	-2.16	0.037

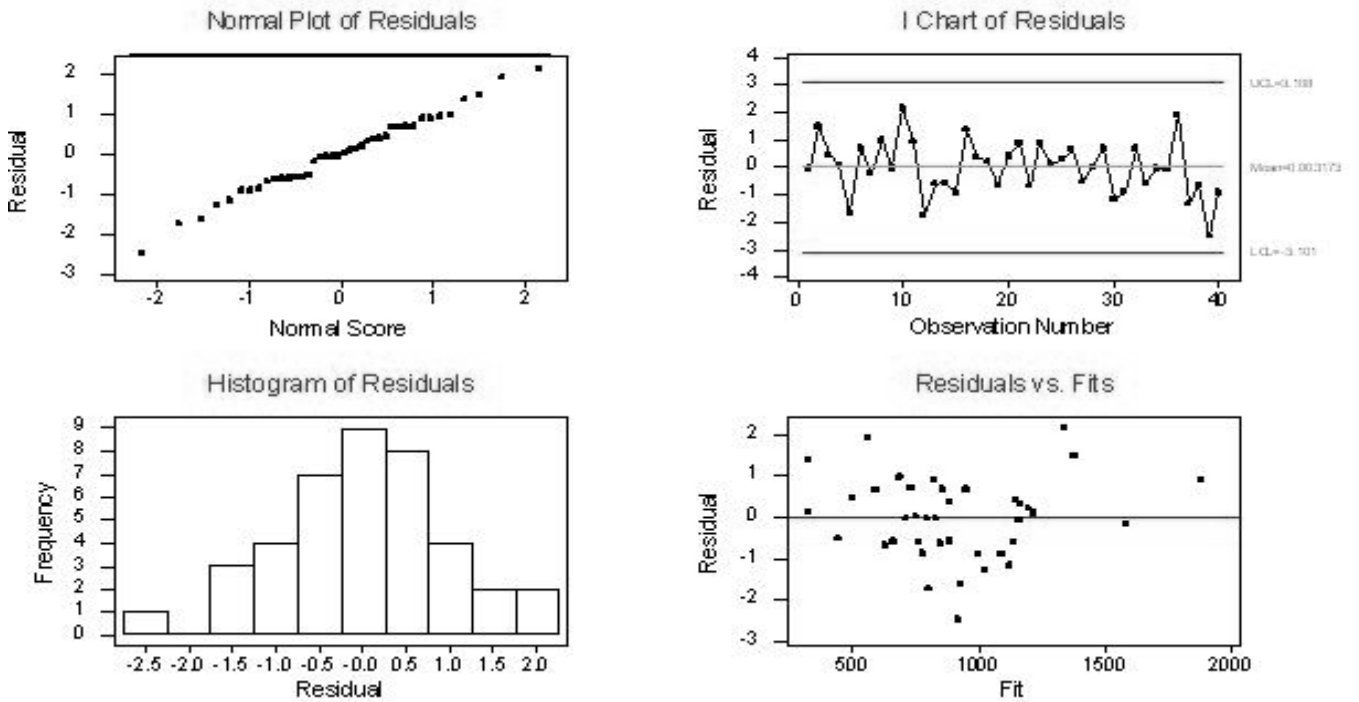
S = 220.4 R-Sq = 73.9% R-Sq(adj) = 67.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	9	5084006	564890	11.63	0.000
Residual Error	37	1796922	48565		
Total	46	6880928			

Source	DF	Seq SS
Ed	1	717146
PoEXP	1	2452248
LF	1	148491
Mf	1	254590
Pop	1	25153
U1	1	81146
U2	1	287450
INC	1	891584
PROB	1	226197

plot of standardised residuals and fitted values



Comments:

- The p-p plot of residuals shows normality which can also be observed from the histogram.
- Non-random scatter in residuals vs. fits.
- So this model is also not recommended.

Variance Inflation Factors:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Tolerance	Variance Inflation
Intercept	1	-5142.41763	1596.94064	-3.22	0.0033	.	0
Ed	1	18.93551	5.68901	3.33	0.0025	0.20417	4.89793
PoEXP	1	6.74590	1.31758	5.12	<.0001	0.21062	4.74796
LF	1	-1.12838	1.07352	-1.05	0.3025	0.45893	2.17897
MF	1	2.33027	1.71904	1.36	0.1865	0.31847	3.14005
Pop	1	-0.69689	1.69959	-0.41	0.6850	0.34583	2.89162
NW	1	0.71361	0.53918	1.32	0.1968	0.30360	3.29385
U1	1	-3.22842	3.59844	-0.90	0.3776	0.19050	5.24925
U2	1	6.81274	7.64601	0.89	0.3808	0.19037	5.25295
GDP	1	-0.32127	1.07124	-0.30	0.7665	0.08594	11.63583
INC	1	7.38184	2.11865	3.48	0.0017	0.12714	7.86559
PROB	1	-3621.71447	1981.93658	-1.83	0.0787	0.49772	2.00915
TIME	1	8.85591	7.37273	1.20	0.2401	0.38377	2.60575

Collinearity Diagnostics

Number	Eigen value	Condition Index	-----Proportion of Variation-----				
			Intercept	Ed	PoEXP	LF	MF
1	11.74350	1.00000	0.00000223	0.00001519	0.00011041	0.00001509	0.00000196
2	0.52345	4.73653	0.00000467	0.00016563	0.00102	0.00006980	0.00000649
3	0.43728	5.18227	0.00000322	0.00001583	0.00027102	0.00002190	0.00000417
4	0.10550	10.55067	5.966725E-7	0.00001812	0.01367	1.418719E-7	1.417286E-7
5	0.08243	11.93583	0.00006888	0.00017839	0.05906	0.00088473	0.00004808
6	0.07020	12.93354	0.00000414	0.00057632	0.04436	0.00037845	7.118048E-7
7	0.01779	25.69542	0.00023232	0.00059058	0.01673	0.00430	0.00032921
8	0.01022	33.89334	0.00024698	0.02115	0.23792	0.00583	0.00027504
9	0.00464	50.33403	0.00046051	0.00680	0.44552	0.01606	0.00004410
10	0.00244	69.39677	0.00546	0.05207	0.06692	0.36865	0.00062335
11	0.00164	84.65730	0.00004321	0.84908	0.00003949	0.35069	0.00202
12	0.00074846	125.26026	0.14986	0.06141	0.03093	0.11822	0.10901
13	0.00016498	266.80101	0.84362	0.00793	0.08345	0.13487	0.88764

Number	-----Proportion of Variation-----						
	Pop	NW	U1	U2	GDP	INC	PROB
1	0.00068427	0.00060429	0.00004577	0.00008022	0.00001699	0.00003110	0.00046583
2	0.02227	0.19103	0.00027347	0.00002539	0.00023789	0.00007780	0.00269
3	0.24504	0.01601	0.00003521	0.00009397	0.00000774	0.00020961	0.02700
4	0.15445	0.18397	0.00081507	0.00296	0.00030332	0.00031802	0.42672
5	0.00832	0.01029	0.00056048	0.01253	0.00042808	0.00751	0.07951
6	0.00012264	0.01956	0.02649	0.06096	0.00205	0.00127	0.00765
7	0.16548	0.00498	0.00341	0.02002	0.00379	0.07991	0.22915
8	0.22609	0.48908	0.09597	0.15137	0.01538	0.08681	0.03373
9	0.01108	0.03429	0.55658	0.58050	0.07604	0.00790	0.04581
10	0.00035840	0.00616	0.01331	0.09798	0.34958	0.14329	0.01541
11	0.00611	0.00890	0.04094	0.03815	0.07394	0.01135	0.00059700
12	0.00420	0.00558	0.18518	0.03280	0.45520	0.65917	0.10268
13	0.15580	0.02955	0.07639	0.00253	0.02302	0.00215	0.02859

-Proportion of Variation-
Number TIME

1	0.00014284
2	0.00000629
3	0.00102
4	0.00141
5	0.07024
6	0.00795
7	0.51506
8	0.20395
9	0.14528
10	0.02451
11	0.00395
12	0.00274
13	0.02375

Comments:

- **Variance Inflation Factors (VIF_i'S):**

VIF's corresponding to the regressors 'gdp' and 'inc' are very large which indicates that there is a near dependence of those two regressors and the other regressor which are highlighted in the table.

- **Collinearity Diagnostics :**

By observing the tables of Proportion of Variations there are two large condition indices indicate dependency large proportions with in the corresponding rows indicates the X-columns that are candidates for dependency.

FINAL REPORT

After applying different techniques on the data is observed that

- a. Transformations are needed for the regressors 'pop' and 'nw'.
- b. And it happened to delete two observations which are influencing the model very much. The deletion shown better result in fitting appropriate model.
- c. After doing 'selection of variables' it happened to delete some variables for better fit. Even though we obtained a better model it is not satisfying some model assumptions like normality. So it is not best. So we added them back and continued the analysis.
- d. Finally we suggested the model with all regressors given and deleting only two observations.

Suggested model for the given data;

$$\begin{aligned} \text{CRIME} = & - 4947 + 16.3 \text{ Ed} + 4.86 \text{ PoEXP} - 0.34 \text{ LF} + 1.88 \text{ Mf} - 5.68 \text{ U1} + 13.1 \text{ U2} \\ & + 0.63 \text{ GDP} + 7.52 \text{ INC} - 3000 \text{ PROB} + 5.81 \text{ TIME} + 130 \log_{10}\text{NW} \\ & - 83.5 \ln\text{POP} \end{aligned}$$