M.Sc. PROJECT REPORT

BY

TAO FENG

SUPERVISOR: DOUGLAS WIENS

DATE: APRIL 29, 1992

M.Sc. PROJECT REPORT

SUPERVISOR: Dr. DOUGLAS WIENS.

STUDENT: Mr. TAO FENG.

DEPT. OF STATISTICS AND

APPLIED PROBABILITY.

UNIV. OF ALBERTA.

PROJECT 1

PART 1. INTRODUCTION

(1).The problem. There is a paper in 1958 (Journal of Geology 66, 114-150) by E.D.Sneed and R.L.Folk where the following result was demonstrated through an experiment at the Colorado River, Texas. The roundness of the stone along the river was thought to be dependent on the distance of transport and the lithology of the stone. The roundness was obtained by visual comparison of the silhouette of the maximum projection face of the pebble images developed by Krumbein (1941), based on the scale developed by Wadell (1934). The roundness has three levels:angular, regular, and rounded. The lithology has six levels: Limestone, Mudstone, Grantoid, Migmatite, Gabbro and Metasedimentary. The classification of the distance of transport is less objective than the one of roundness and lithology. In the paper, the authors used eight locations to pick up the stones. The following map indicates the eight locations and is helpful for us to understand the experiment.

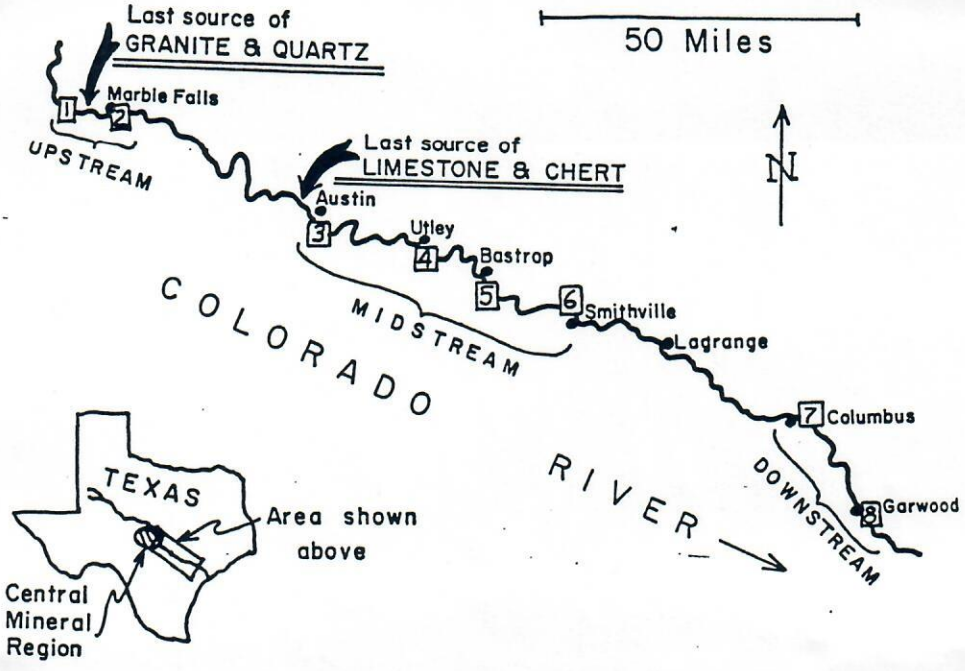


FIG. 1.—Location of sampling stations 1–8

Mr. Tracy Brennand, a graduate student from the Department of Geography, applied a similar experiment on another river. The difference is the classification of the distance of transport. He stopped at three locations which are 17.78, 28.07, 32.52 kilometers from the origin of the river respectively, and picked up randomly 60, 110, 360 stones respectively from the above three locations. Then he examined every stone to determine its appropriate levels of roundness and lithology. The classification of the roundness and lithology are the same as the ones used in the paper mentioned at the beginning.

With this experiment, he wants to somehow prove or disprove the generality of the result in the 1958 paper. Unfortunately, the analysis done by the authors was not clearly mentioned. Therefore, he turned to our department for help. As soon as I realized the problem, I divided the work into two parts: I would do the statistical data analysis; to prove or disprove statements and the other professional interpretations would be of course his responsibilities. Furthermore, we agree on the detailed objectives for my part of job(see (3).).

(2). The data. There are 530 observations and three variables:

Variable Y: the roundness
- angular
- regular
- rounded

Variable X1: the distance of transport
- 17.78 km from the origin
- 28.07 km from the origin
- 32.52 km from the origin

Variable X2: the lithology
- Limestone
- Mudstone
- Granitoid
- Migmatite
- Gabbro

Metasedmentary

Part of the data is exhibited below:

Observation	Y	X1	X2
1	angular	17.78	Limestone
2	rounded	17.78	Gabbro
3	rounded	17.78	Migmatite
...	...		
530	regular	32.52	Gabbro

(3). The objectives:

(i). Test to see if we may conclude that Y and X1, Y and X2, X1 and X2 are independent or not.

(ii). If Y and X1, Y and X2 are not independent , estimate the relationship that exists between Y and X1, X2.

PART 2. DATA ANALYSIS FOR OBJECTIVE 1

(1). Chi-square test of independence. First, I apply the chi-square method to test the pairwise independence of Y, X1, X2. The Frequency Procedure in SAS was used . However, since there are too many empty cells, the chi-square tests' results given by the SAS output may not be valid. Therefore, I turn to the "Fisher's exact test" method.

(2).Fisher's exact test of independence. I applied the Fisher's exact test method to test the following three null hypotheses:

(i). H01: Y and X1 are independent.

The result given by SAS: (Statistics for table of Y by X1)

Statistic	DF	Value	Prob
Chi-Square	10	19.070	0.039
Likelihood Ratio Chi-Square	10	21.583	0.017

(ii). H02: Y and X2 are independent.

The result given by SAS: (Statistics for table of Y by X2)

Statistic	DF	Value	Prob
Chi-Square	4	1060.000	0.000
Likelihood Ratio Chi-Square	4	885.828	0.000

(iii). H03: X1 and X2 are independent.

The result given by SAS: (Statistics for table of X1 by X2)

Statistic	DF	Value	Prob
Chi-Square	25	56.987	0.000
Likelihood Ratio Chi-Square	25	39.883	0.030

From the above, we conclude that all three null hypotheses are rejected at α=0.05. Y and X1, Y and X2, X1 and X2 are all dependent.

(3). Log-linear method. Actually I also tried the log-linear method to fit the data in order to have the pairwise independence tested. The loglinear model is:

$$\log m_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

where m_{ijk} =count of the cell (i,j,k).

α_i indicate the effect of Y(roundness).

β_j indicate the effect of X1 (distance of transport).

γ_k indicate the effect of X2 (lithology).

i=1,2,3. j=1,2,3. k=1,2,3,4,5,6.

The log-linear model has very nice interpretations. However, the SAS output failed to give the estimates and standard deviations of all the parameters, or some chi-square statistics and p-values. Furthermore, the degrees of freedom of some effects are not valid because such effects contain one or more redundant or restricted parameters. The reason for the failure is that there are too many zero counts in the data set. Therefore, the Fisher's exact test method is the only appropriate method in this case.

PART 3. DATA ANALYSIS FOR OBJECTIVE 2

(1). Grizzle, Starmer, Koch method (weighted least square method). (See "Discrete Multivariate Analysis" by Bishop, Fienberg and Holland. P353-357). The CATMOD procedure on SAS was used. However, the SASLOG file shows that the G.S.K. method failed to analysis the data, because the response functions are linearly dependent due to too many zero counts. Therefore, I turned to the regression method.

(2). Regression method.

(i). Since Y is a nominal variable with three levels, first I transform Y according to the classification rule used in the G.S.K. method and the typical assignment follows:

nominal Y (roundness)	assigned value
level 1 (angular)	0
level 2 (regular)	0.5
level 3 (rounded)	1

Therefore, for each distance-lithology combination whose count is non-zero (14 of them), I calculate the weighted average. For example,if there are (a+b+c) observations in a distance-lithology combination, where "a" stones are angular (level1), "b" stones are regular (level2), and "c" stones are rounded (level3), so the weighted average for this combination is equal to (a*0+b*0.5+c*1)/(a+b+c)=(0.5b+c)/(a+b+c).

Altogether, we have 14 such weighted averages, denoted y1,y2,...., y14 along with their corresponding distance-lithology combinations. Here, yi's can be interpreted as a continuous measure of the average roundness of the stones from the ith distance-lithology combination. According to the above, I rearrange the original data by defining the following variables:

Dependent variable Y : The roundness measure as defined above;

Indep. variable X1: X1= 1----------if the distance of transporting is 28.07 km.
0----------if not.

Indep. variable X2: X2= 1----------if the distance of transport is 32.52 km.
0----------if not.

Indep. variable X3: X3= 1----------if it's Mudstone.
0----------if not.

Indep. variable X4: X4= 1----------if it's Granitoid.
0----------if not.

Indep. variable X5: X5= 1---------- if it's Migmatite.
0----------if not.

Indep. variable X6: X6= 1----------if it's Gabbro.
0----------if not.

Indep. variable X7: X7= 1----------if it's Metasedimentary.
0----------if not.

There are 14 observations having value on each of the above variables.

(ii). Fitting the following regression model:(Model 1)

$$Y=B0+B1X1+B2X2+B3X3+B4X4+B5X5+B6X5+B7X7+B8X13+B9X14+B10X!5$$
$$B11X16+B12X17+B13X23+B14X24+B15X25+B16X26+B17X27+\varepsilon$$

with the weight matrix whose diagonal elements are equal to the number of counts in the corresponding distance-lithology combinations(14 of them), and off-diagonal elements are equal to zero.

where X13=X1*X3, X14=X1*X4, X15=X1*X5, X16=X1*X6, X17=X1*X7, and X23=X2*X3, X24=X2*X4, X25=X2*X5, X26=X2*X6, X27=X2*X7.

(a). Although I am not able to obtain the summary statistics since the number of the observation is less than the number of parameters in the model 1, my interests are the following: (1). Applying stepwise regression to this starting model.

(2). Checking the linear regression model's assumptions by the residual analysis.

(3). Detecting the problem of multicollinearity by looking at the V.I.F.'s;

(b).For the first interest, SAS output gives the following result of the stepwise regression:

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 Procedure for Dependent Variable Y

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X23		1	0.5162	0.5162	.	12.8059	0.0038
2	X6		2	0.1719	0.6881	.	6.0636	0.0315
3	X5		3	0.1376	0.8257	.	7.8940	0.0185
4	X16		4	0.0505	0.8762	.	3.6684	0.0877

For the second interest, the normal probability plot of residuals and the Pearson Correlation Coefficient indicate serious departure from the normality assumption of the errors. The plot of the residuals vs. predicted values of Y indicates no systematic pattern, and therefore there is no strong evidence against the independence assumption of the errors.

For the third interest, the V.I.F.'s indicate only a slight problem of multicollinearity. . Therefore, I don't worry about it too much.

(iii). Fitting the following reduced model: (Model 2)

$$Y=B0+B1X5+B2X6+B3X16+B4X23+\varepsilon$$

with the same weight matrix as the one in model

(a). Summary statistics:

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	0.39872	0.09968	15.923	0.0004
Error	9	0.05634	0.00626		
C Total	13	0.45507			

Root MSE	0.07912	R-square	0.8762
Dep Mean	0.48720	Adj R-sq	0.8212
C.V.	16.24029		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for HO: Parameter=0	Prob > |T|
INTERCEP	1	0.490520	0.00355627	137.931	0.0001
X5	1	-0.058494	0.01849694	-3.162	0.0115
X6	1	-0.055860	0.02311577	-2.417	0.0388
X16	1	-0.097820	0.05107308	-1.915	0.0877
X23	1	0.480910	0.07920196	6.072	0.0002

From the above ANOVA of the model 2, we can see that the model fits the data very well. R-sq and R-sq adjusted are both considerably high for such a reduced model, all the independent variables in the model except X16 are significant at $\alpha=0.05$. Therefore I will drop the insignificant term X16 and fit the further reduced model.

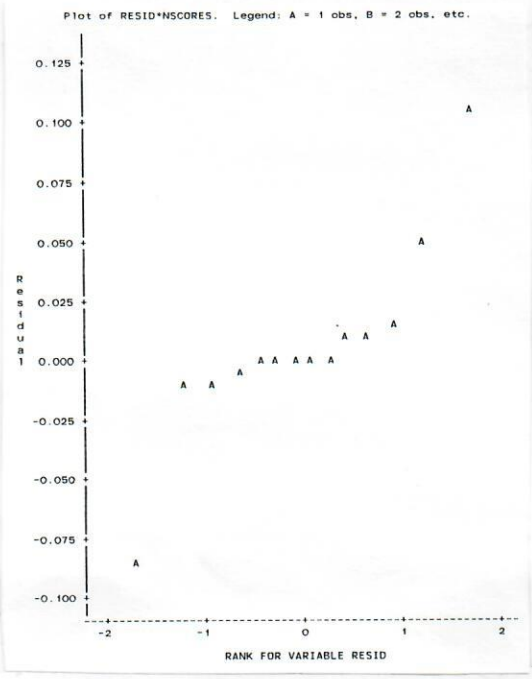
(b). V.I.F.'s:

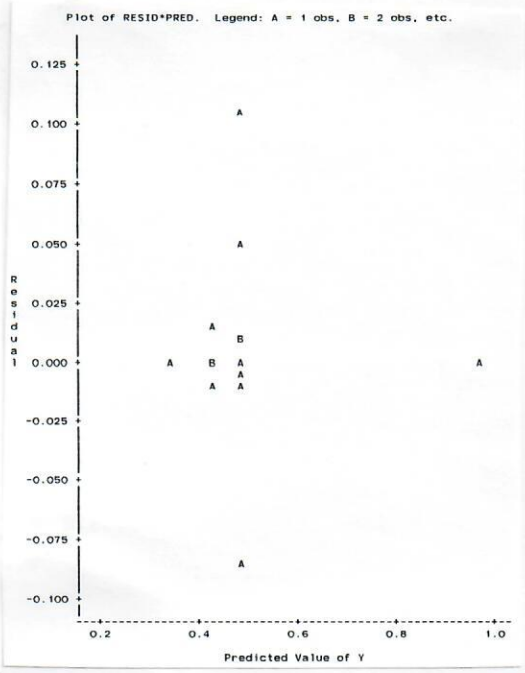
Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
X5	1	1.00115876
X6	1	1.24406804
X16	1	1.24292453
X23	1	1.00012960

All the VIF's are less than 2, which indicate no serious problem of multicollinearity.

(c). Residual analysis.

(1).Normal probability plot of the residuals and the Pearson Correlation Coefficient indicate serious departure from the normality assumption of the error term.(Shown below)





Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 14

	RESID	NSCORES
RESID	1.00000	0.87427
Residual	0.0	0.0001
NSCORES	0.87427	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

(2). Plot of residuals vs. predicted values of Y shows no systematic pattern, which provide no evidence against the constant variance and independence assumption of the errors.

(iv). Fitting the following reduced model: (model 3)

$$Y=B0+B1X5+B2X6+B3X23+\varepsilon$$

with the same weight matrix applied in the model 1

(a). Summary statistics:

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	0.37576	0.12525	15.793	0.0004
Error	10	0.07931	0.00793		
C Total	13	0.45507			

Root MSE	0.08905	R-square	0.8257
Dep Mean	0.48720	Adj R-sq	0.7734
C.V.	18.27906		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.490520	0.00400272	122.547	0.0001
X5	1	-0.058494	0.02081900	-2.810	0.0185
X6	1	-0.075424	0.02333967	-3.232	0.0090
X23	1	0.480910	0.08914480	5.395	0.0003

From the above ANOVA table, we can see that the weighted regression model fits the data very well. All of the independent variables are significant at $\alpha=0.05$.

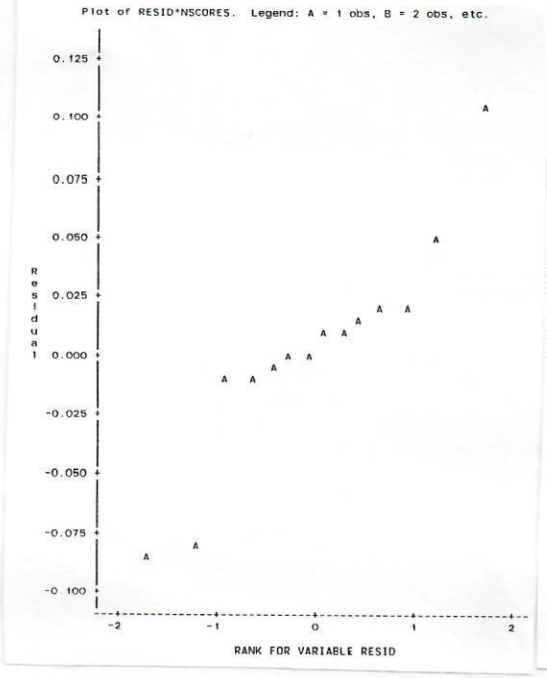
(b). VIF's:

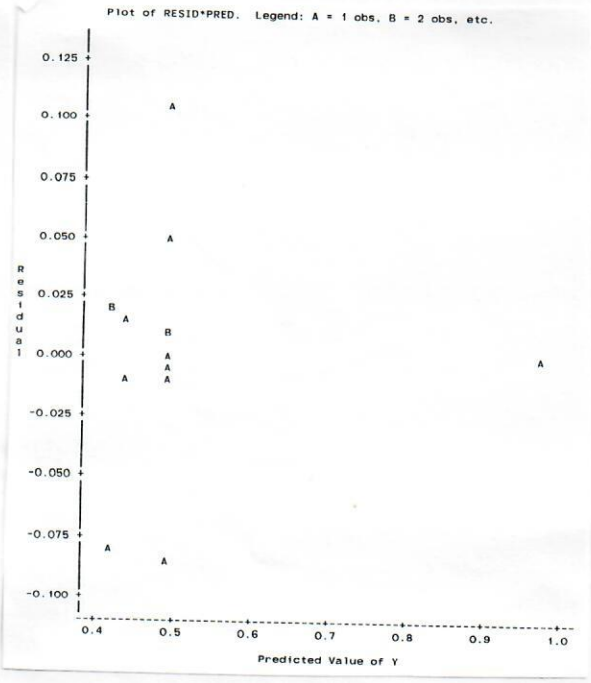
Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
X5	1	1.00115876
X6	1	1.00114351
X23	1	1.00012960

All the VIF's are considerably small, which indicate no problem of multicollinearity.

(c). Residual analysis.

(1). Normal probability plot of the residuals and the Pearson Correlation Coefficient indicate no serious departure from the normality assumption of the errors.





Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 14

	RESID	NSCORES
RESID	1.00000	0.93522
Residual	0.0	0.0001
NSCORES	0.93522	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

(2). Plot of the residuals vs. predicted values shows no evidence against the constant variance and independence assumptions of the errors.

PART 4. CONCLUSIONS AND COMMENTS

(1). The analysis in part 2 gives us a general idea about the dependence among the roundness, the distance of transport, and the lithology. We conclude that the roundness of the stone is dependent of its distance of transport from the origin of the river, and its lithology. The distance of transport of the stone is dependent on its lithology.

There are some aspects which ought to be mentioned.

(i).This kind of general relationship (independent or not) does not give the precise relationship and can be somewhat misleading. For example, since the distance of transport is dependent on the lithology, the dependence of the roundness on the distance may be mainly due to its dependence on the the lithology. For another example, we may ask, "Does the factor 'a stone is limestone or not' have a significant influence on the roundness of the stone?"

(ii).The tests for the independence does give us a general idea, and lead us to the appropriate detailed analysis. Besides, the tests are thought to be helpful for interpretations by the client.

(2). Mainly based on the model 3 in part 3, I draw the following conclusions:

(i). As we can see from the analysis in part 3, the model 3 fits the data very well. The proportionate reduction of the total variation in Y(defined in part 3) associated with the use of the indicator variables X1,X2....,X7 (defined in part 3) is over 80%. The adjusted coefficient of multiple determination is also considerably large. We do have strong evidence of the linear relationship between Y (the measure of the roundness defined in part 3) and the indicate variables which indicate the levels of the distance of transport and the lithology,according to the data.

(ii). From the model 3, we can see that the significant indicator variables are X5, X6, and X23. Therefore, (1). The factor " the stone is migmatite or not" and the factor "the stone

is gabbro or not" affect the defined measure of the roundness of the stone significantly, while the variables indicating the other levels of the lithology do not have such significant impacts on the defined measure of the roundness. In fact, as we can see from the estimates of the parameters, the migmatite tends to be less rounded and so does the gabbro. However, the other lithological stones don't have such tendencies. (2). Since the interaction between X2 and X3 is significant, the mudstone tends to be more rounded than the other lithological kinds of stone, given the fact that the stones are collected 32.52 km away from the origin.

(3). In part 3, I applied weighted regression method to the defined measure Y. Since the variable Y actually contains the weighted averages of the 14 distance-lithology combinations, the analysis is much less sophisticated due to loss of information on the actual values of nominal variable Y.

PROJECT 2

PART 1. INTRODUCTION

(1). The problem. This consulting problem was introduced by Mr. John Kaul, a graduate student in the Department of Computer Science. He intends to design a computer package to improve the Acute Care Funding System in Alberta. The Acute Care Funding model is the basic funding mechanism that has been adopted to fund all hospitals in Alberta. Its premise is to promote an efficient distribution of funds based on the annual performance of all hospitals

within the project. Performance is being measured by the actual demonstrated costs incurred while treating a patient (case) mix of varying lengths of stay, age and severity. Before the design of the package, he wants to know whether the relationship between cost and the three variables (length of stay, age and severity) is a significant one or not. "The length of stay" and "the patient's age" are of course continuous variables, while "severity" is a nominal variable with four levels-----"minor", "moderate", "major", "catast.".

The data was collected by John. First, three hospitals within the project were randomly selected. And then within each hospital 30 cases (patients) were observed (about one-fourth of the 30 patients come from each one of the four levels indicating the severity). Altogether, there are 90 cases, along with their ages, length of stay, severity, cost and sex. The variable "sex" was not considered by the funding mechanism. However, John thought that "sex" might have an impact on the "cost".

After our meeting, I wrote down the following objectives in statistical language and had them confirmed by the client.

(2). The objectives:

(i). Test the significance of the "sex" effect on the "cost". If it is not significant, we might eliminate it and consider only the other variables thought to affect the "cost".

(ii). Test the validation of the relationship between the cost and the following criteria: Length of stay, Age, Severity , and sex (if it is found to be significant). In fact, only the linear relationship is of concern.

(iii). Determine if we can simplify the linear relationship mentioned above. The simplification was thought to be helpful for the design of the computer package. Interpret the variables in the simplified model in terms of explaining the variation of the cost.

(3). The method and the data structure. According to the above, multiple linear regression is the optimal method to apply. Therefore, I define the following variables:

Dependent variable Z: The cost of the patient observed (in dollars).

Independent variables:

X1= 1----------if the patient is male.
0----------if not.

X2= 1----------if the patient's severity level is "moderate".
0----------if not.

X3= 1----------if the patient's severity level is "major".
0----------if not.

X4= 1----------if the patient's severity level is "catastrophic"
0----------if not.

X5=Length of stay of the patient (in days).

X6=Age of the patient (in years).

There are 90 observations and a portion of the data follows:

Observation	Z	X1	X2	X3	X4	X5	X6	
1	10327.50	1	0	0	0	11	67	
2	20351.85	1	1	0	0	7	55	... ...
90	30789.20	1	0	0	1	34	70	

PART 2. DATA ANALYSIS.

(1). Fitting the following regression model: (model 1)

$$Z=B0+B1X1+B2X2+B3X3+B4X4+B5X5+B6X6+B7X12+B8X13+B9X14+B10X15$$
$$B11X16+B12X25+B13X26+B14X35+B15X36+B16X45+B17X46+B18X56+\varepsilon$$

Where X12=X1*X2, X13=X1*X3, X14=X1*X4, X15=X1*X5, X16=X1*X6,

X25=X2*X5, X26=X2*X6, X35=X3*X5, X36=X3*X6, X45=X4*X5,

X46=X4*X6, X56=X5*X6.

(i). Residual analysis. Because of my suspicion of the departure from the normality assumption of the error term, I first do the residual analysis of the model 1.

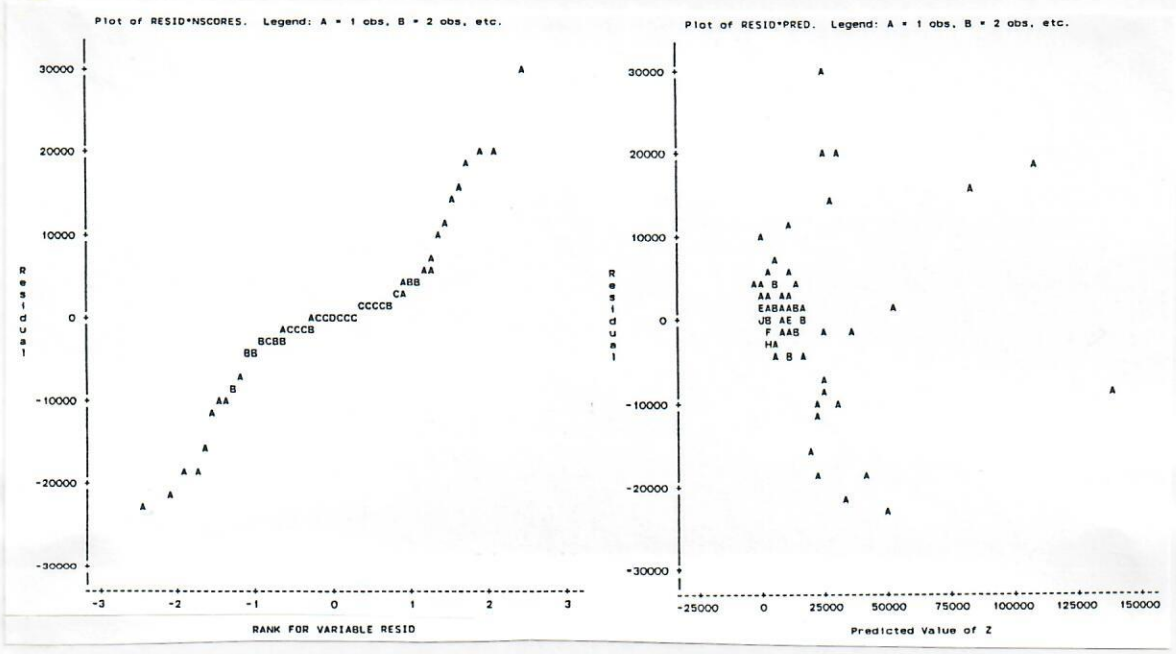
(a). Pearson Correlation Coefficient.

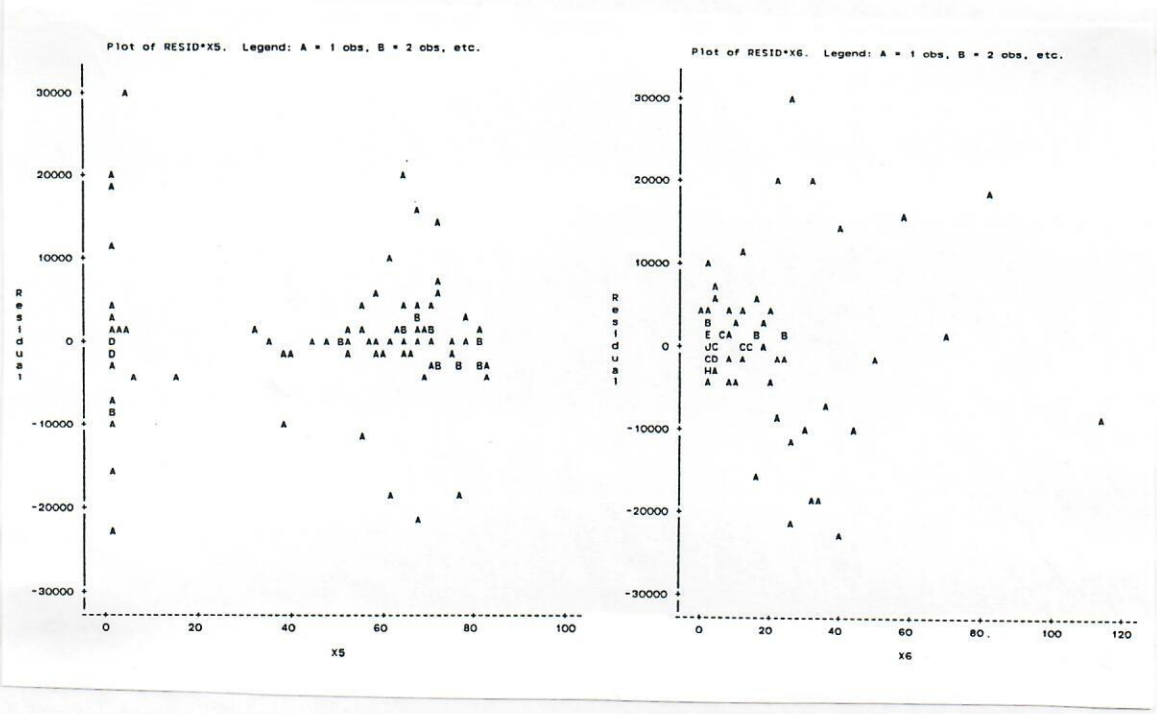
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 90

	RESID	NSCORES
RESID	1.00000	0.92641
Residual	0.0	0.0001
NSCORES	0.92641	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

Since 0.92641 is less than the corresponding critical value at α=0.05, there is strong evidence against the normality assumption of the error.

(b). Residual plots:





1). Normal probability plot indicates serious departure from the normality assumption of the error.

2). Plot of residuals vs. predicted Z shows a slightly systematic pattern which indicates a slight departure from the constant variance assumption of the error term.

3). Plot of residuals vs. X5 shows no systematic pattern, while the plot of residuals vs. X6 has a slightly systematic pattern.

In short, there is a serious departure from the normality assumption of the error, and there is a slight departure from the constant variance assumption of the error term. In addition, all of the plots mentioned in 2) and 3) show more or less random pattern around the base line 0, which indicate no departure from the independence assumption of the error term.

(ii). Transformation. Inspired by the normal prob. plot of the residuals, I tried two kinds of transformation "log(Z)" and "sqrt(Z)". The transformation "sqrt(Z)" turns out to be better in terms of remedying the departure from the normality assumption of the error. At this

stage, I don't worry about the slight departure from the constant variance assumption of the error too much, since it is better to do one thing at a time.

(2). Fitting the following transformed regression model: (model 2)

$$Y=B0+B1X1+B2X2+B3X3+B4X4+B5X5+B6X6+B7X12+B8X13+B9X14+B10X15$$
$$+B11X16+B12X25+B13X26+B14X35+B15X36+B16X45+B17X46+B18X56+\varepsilon$$

Where Y=sqrt(Z), and the independent variables as defined before.

(i). Summary statistics:

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	18	388131.21305	21562.84517	22.647	0.0001
Error	71	67601.67358	952.13625		
C Total	89	455732.88664			

Root MSE	30.85671	R-square	0.8517
Dep Mean	93.54559	Adj R-sq	0.8141
C.V.	32.98574		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for HO: Parameter=0	Prob > |T|
INTERCEP	1	62.580450	24.88818760	2.514	0.0142
X1	1	0.879865	24.00571219	0.037	0.9709
X2	1	38.478301	33.28933809	1.156	0.2516
X3	1	38.777681	26.97078571	1.438	0.1549
X4	1	1.931020	31.59799323	0.061	0.9514
X5	1	-0.557579	0.36650126	-1.521	0.1326
X6	1	2.504805	0.56115846	4.464	0.0001
X12	1	-14.975363	22.70679708	-0.660	0.5117
X13	1	-13.412778	19.58349442	-0.685	0.4956
X14	1	0.420266	24.12321981	0.017	0.9861
X15	1	-0.010964	0.34384000	-0.032	0.9747
X16	1	0.772118	0.46080439	1.676	0.0982
X25	1	-0.221687	0.39352745	-0.563	0.5750
X26	1	-1.613710	0.90430254	-1.784	0.0786
X35	1	-0.322152	0.32960195	-0.977	0.3317
X36	1	-0.854570	0.72322412	-1.182	0.2413
X45	1	0.161565	0.35758111	0.452	0.6528
X46	1	0.160785	0.54235357	0.296	0.7677
X56	1	0.024555	0.00802130	3.061	0.0031

(a). R-sq. and adj. R-sq. shown above indicate that the model 2 give a good fit to the data.

(b). The independent variable X1 and its associated interaction terms are not significant at α=0.05. Therefore, the addition of the "sex" variable is not necessary. We can delete X1, X12, X13, X14, X15, X16 from the model, according to the following test:

H0: B1=B7=B8=B9=B10=B11=0.

Fcal=[(SSEred-SSEfull)/6]/MSEfull=[(72850.29802-67601.67358)/6]/952.13625

=0.918745337 < Fcritical 6,71 (α=0.05).

Therefore, H0 is not rejected at α=0.05.

(c). VIF's:

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
X1	1	11.89639605
X2	1	19.34640633
X3	1	13.08121932
X4	1	16.31193307
X5	1	11.62183766
X6	1	10.55750809
X12	1	7.46692744
X13	1	4.18904106
X14	1	7.22554430
X15	1	10.98972676
X16	1	3.20611347
X25	1	11.08133957
X26	1	3.49698289
X35	1	6.85251534
X36	1	4.27184381
X45	1	5.99353401
X46	1	7.71464859
X56	1	3.36757434

There are 15 out of 18 independent variables with VIF larger than 4. Therefore, the problem of the multicollinearity is serious.

(iii). Residual analysis.

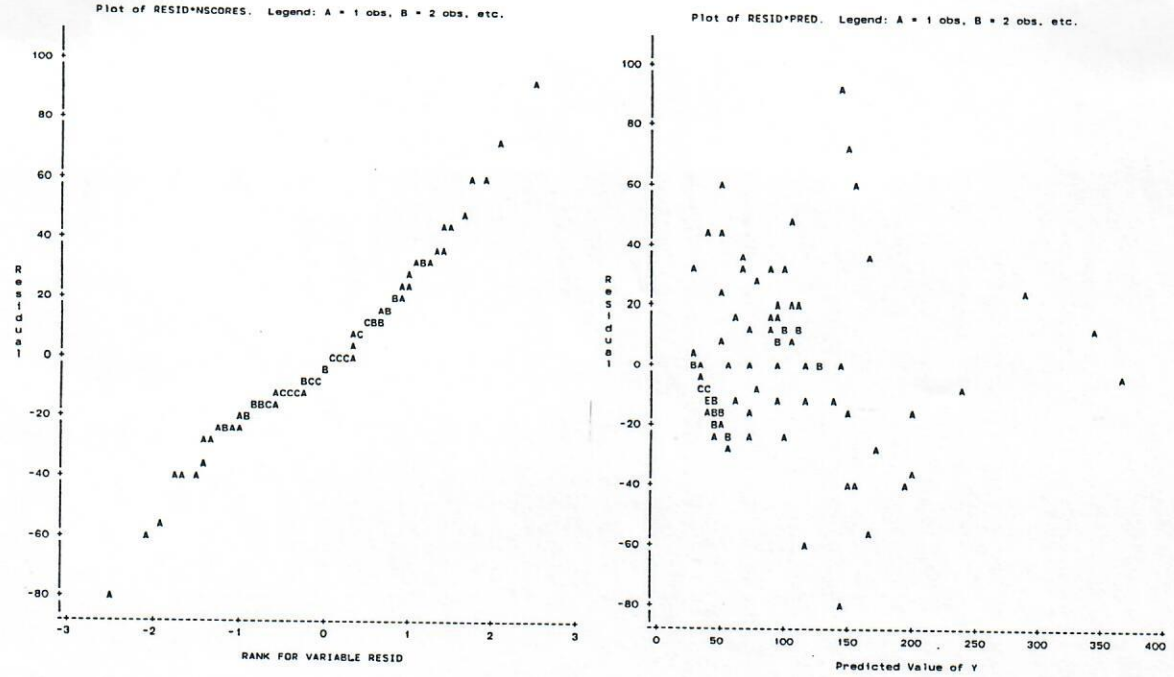
(a). Pearson Correlation Coefficient.

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 90

	RESID	NSCORES
RESID	1.00000	0.97951
Residual	0.0	0.0001
NSCORES	0.97951	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

Since 0.97951 is larger than the corresponding critical value at α=0.05, there is no serious departure from the normality assumption of the error in the transformed model.

(b). Residual plots:



Normal prob. plot of the residuals indicates no serious departure from the normality assumption of the error. The plot of residuals vs. predicted Y, X5, X6 look similar to the ones in the model 1. i.e., The slight departure from the constant variance assumption of the error still exists.

(3). Fitting the following reduced regression model: (model 3)

$$Y=B0+B1X2+B2X3+B3X4+B4X5+B5X6+B6X25+B7X26+B8X35+B9X36$$
$$+B10X45+B11X46+B12X56+\varepsilon$$

(i). Summary statistics:

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	12	382882.58862	31906.88239	33.724	0.0001
Error	77	72850.29802	946.10777		
C Total	89	455732.88664			

Root MSE	30.75886	R-square	0.8401
Dep Mean	93.54559	Adj R-sq	0.8152
C.V.	32.88115		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	67.312566	13.26987262	5.073	0.0001
X2	1	20.130235	24.26305953	0.830	0.4093
X3	1	28.267500	21.30694499	1.327	0.1885
X4	1	9.963311	19.48933143	0.511	0.6107
X5	1	-0.574998	0.22259480	-2.583	0.0117
X6	1	2.591068	0.45413577	5.705	0.0001
X25	1	-0.160165	0.37748759	-0.424	0.6725
X26	1	-1.210286	0.85177670	-1.421	0.1594
X35	1	-0.273043	0.31411593	-0.869	0.3874
X36	1	-0.789631	0.70479281	-1.120	0.2660
X45	1	0.025753	0.32250606	0.080	0.9366
X46	1	0.222130	0.51250878	0.433	0.6659
X56	1	0.026380	0.00726136	3.633	0.0005

(a). R-sq. and adj. R-sq. shown above indicate that the reduced model 3 still provide a good fit to the data. In fact, without X1, X12, X13, X14, X15, X16 in the model, the R-sq. and

adj. R-sq. remain almost the same. R-sq. is reduced by about 0.01, while the adj. R-sq. is increased by 0.001.

(b). F-value is 33.724, and its p-value is 0.0001, therefore the linear relationship between Y and the X's in the model is valid.

(c). There are still many insignificant independent variables in the model 3, thus a model reduction is necessary.

(ii). VIF's:

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
X2	1	10.34283518
X3	1	8.21602763
X4	1	6.24509143
X5	1	4.31432015
X6	1	6.95857490
X25	1	10.26138646
X26	1	3.12230991
X35	1	6.26338208
X36	1	4.08273307
X45	1	4.90645795
X46	1	6.93285699
X56	1	2.77729743

The values of VIF's are reduced, but there are still 10 out of 12 independent variables whose VIF's are larger than 4. Therefore, the problem of the multicollinearity remains.

(iii). Residual analysis.

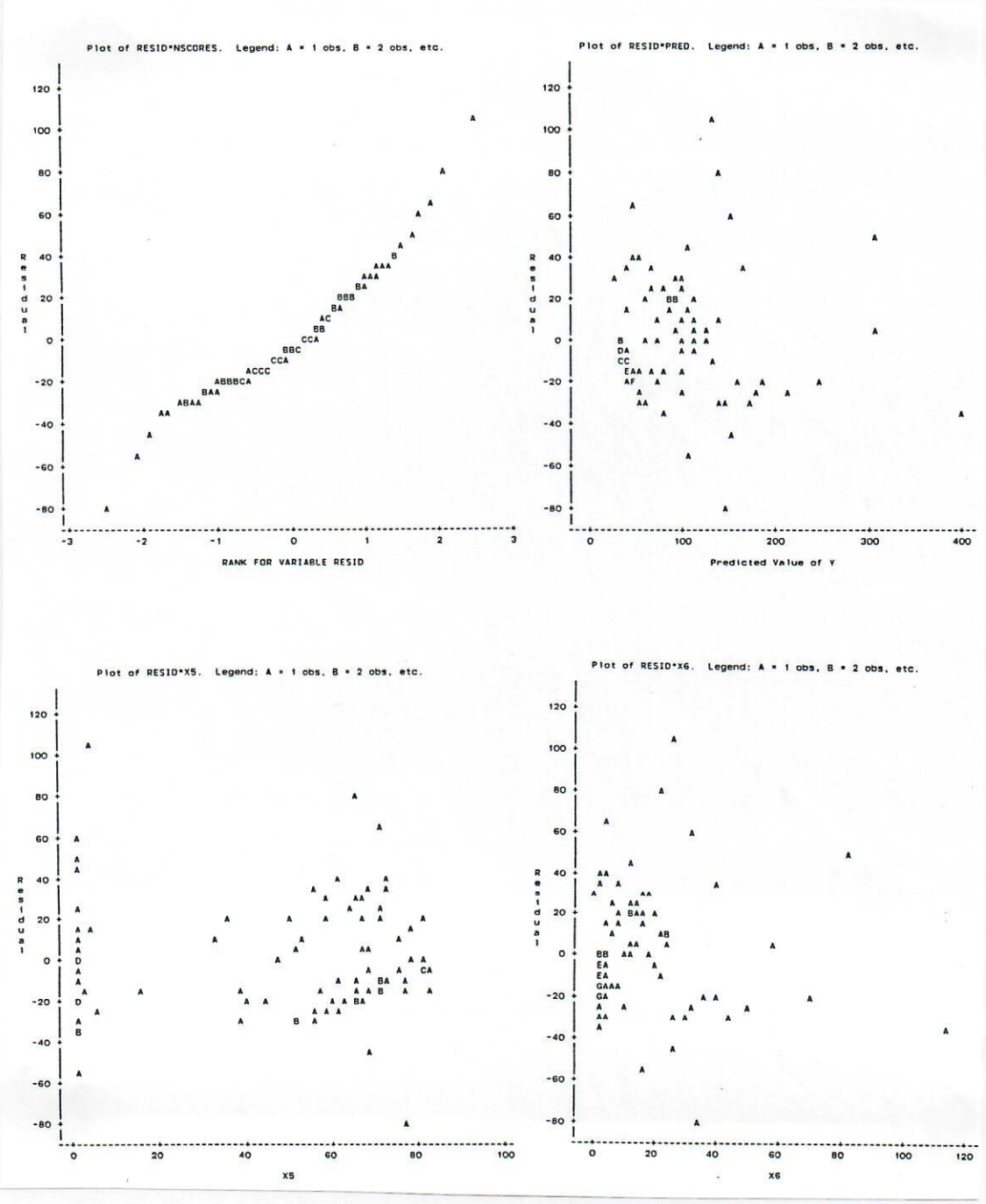
(a). Pearson Correlation Coefficient.

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 90

	RESID	NSCORES
RESID	1.00000	0.97143
Residual	0.0	0.0001
NSCORES	0.97143	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

Since 0.97143 is larger than the corresponding critical value at α=0.05, there is no serious departure from the normality assumption of the error term.

(b). Residual plots:



The normal prob. plot of the residuals indicate no serious departure from the normality assumption of the error. The plots of residuals vs. predicted Y, X5, X6 show no systematic pattern, which indicate no serious departure from the constant variance assumption of the error. i.e., we remedied the slight departure which exists in the model 2.

(4). Model reduction of the model 3.

The stepwise regression method was applied and the result follows:

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 Procedure for Dependent Variable Y

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X6		1	0.7843	0.7843	17.8948	319.9985	0.0001
2	X5		2	0.0160	0.8003	12.2075	6.9516	0.0099
3	X56		3	0.0206	0.8208	4.3016	9.8713	0.0023
4	X46		4	0.0108	0.8317	1.0911	5.4616	0.0218

From the above, we ought to fit the reduced model with only X5, X6, X46, X56 included.

(5). Fitting the following reduced regression model: (model 4)

$Y=B0+B1X5+B2X6+B3X46+B4X56+\varepsilon$

(i). Summary statistics:

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	379011.93933	94752.98483	104.978	0.0001
Error	85	76720.94730	902.59938		
C Total	89	455732.88664			

Root MSE	30.04329	R-square	0.8317
Dep Mean	93.54559	Adj R-sq	0.8237
C.V.	32.11620		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for HO: Parameter=0	Prob > |T|
INTERCEP	1	78.845675	8.14950661	9.675	0.0001
X5	1	-0.640156	0.13963994	-4.584	0.0001
X6	1	2.186751	0.32664632	6.695	0.0001
X46	1	0.692928	0.29650201	2.337	0.0218
X56	1	0.022509	0.00596634	3.773	0.0003

(a). R-sq., adj. R-sq., and F-value shown above indicate that the model 4 provides a good fit to the data. Compared to the model 3, F-value is largely increased; the adj. R-sq. is increased although the R-sq. is slightly decreased.

(b). All of the independent variables in the model are significant at α=0.05. In fact, they are the only significant variables out of the 18 variables.

(ii). VIF's:

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
X5	1	1.77970190
X6	1	3.77354900
X46	1	2.43226277
X56	1	1.96538784

All the VIF's are less than 4, which indicate no serious problem of the multicollinearity.

(iii). Residual analysis.

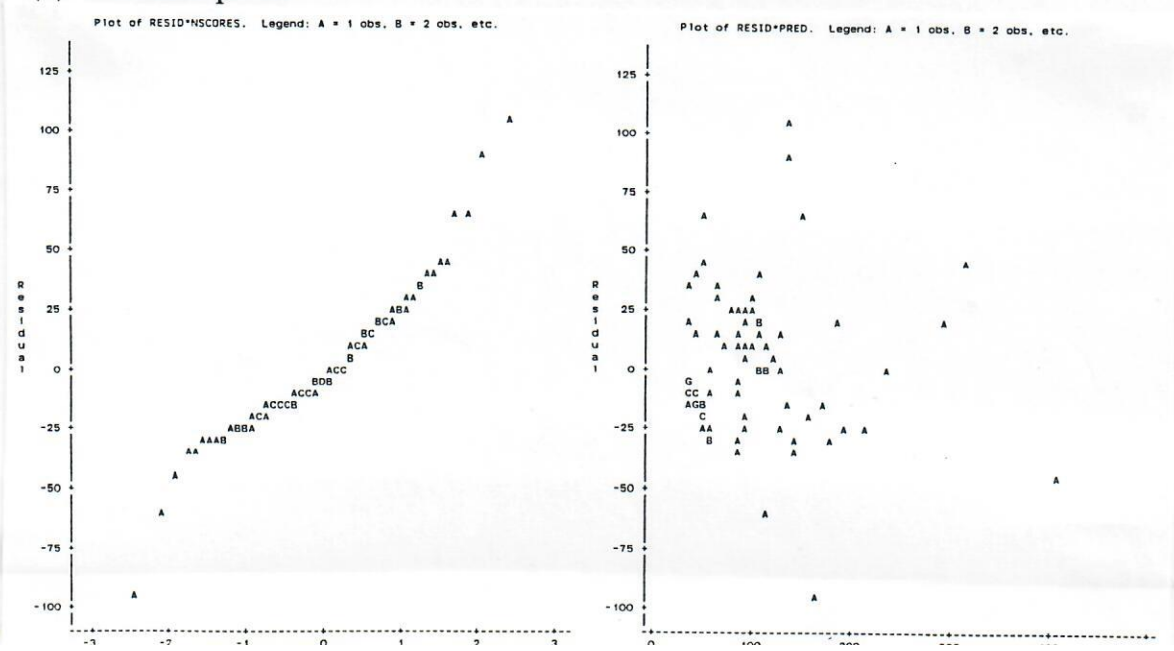
(a). Pearson Correlation Coefficient:

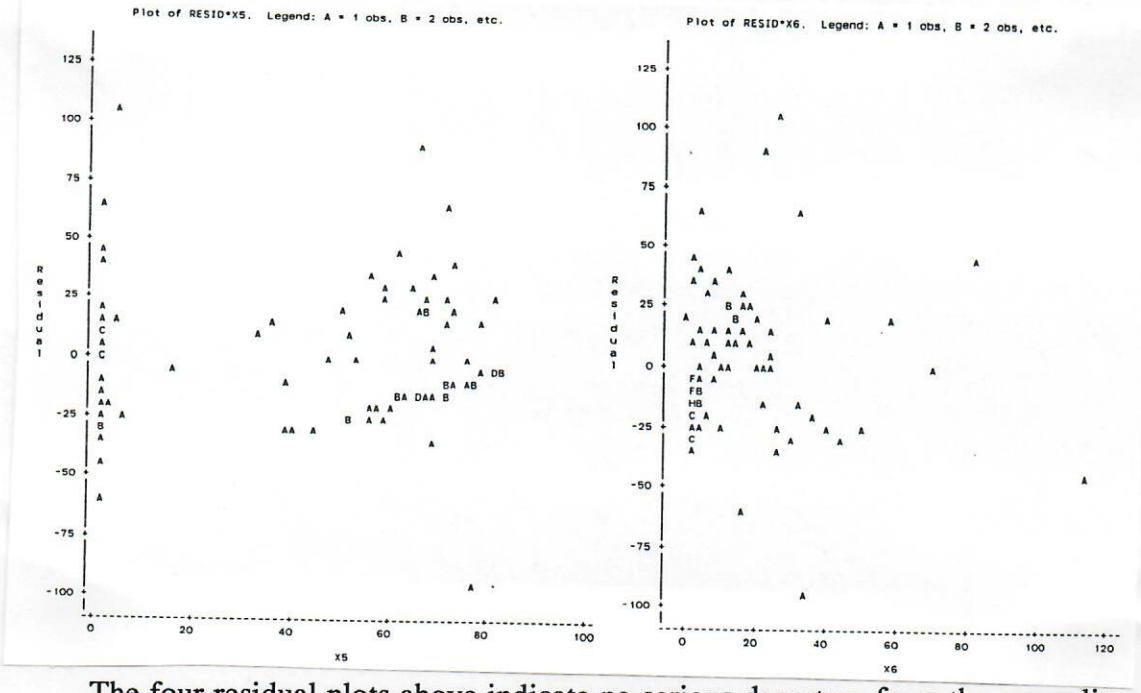
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 90

	RESID	NSCORES
RESID	1.00000	0.96664
Residual	0.0	0.0001
NSCORES	0.96664	1.00000
RANK FOR VARIABLE RESID	0.0001	0.0

Since 0.96664 is still large than the corresponding critical value at α=0.05, there is no serious departure from the normality assumption of the error.

(b). Residual plots:





The four residual plots above indicate no serious departure from the normality, independence and constant variance assumptions of the error term.

From the above (i), (ii), (iii), we may easily see that the model 4 is not only the simpler but also more appropriate than the model 3, in that we are happier with the assumptions made in the analysis.

PART 3. COMMENTS AND CONCLUSIONS.

(1). As we can see from part 2, there are several problems regarding the departures from the model assumptions and the multicollinearity in the beginning model 1.
I succeeded to remedy one of them at a time, without much sacrifice of simplicity. Finally, model 4 is ready for interpretations.

(2). The final model (model 4) is the appropriate model according to the data observed by the client. It might not, however, be appropriate in the overall sense. Furthermore, since only the linear relationship is concerned, we are not able to comment on a possible nonlinear relationship between Y and X's.

(3). From (2) in the part 2, we conclude that the "sex" of the patient doesn't affect his (or her) (square root of) cost in a hospital significantly. i.e., The client's addition of "sex" variable is not necessary.

(4). From (3) in the part 2, we conclude that the linear relationship between square root of the cost and the length of stay, the patient' s age, the severity is a reasonable one. However, there is only a small portion of the independent variables in the model 3 which are significant at α=0.05. Furthermore, there is still a problem of the multicollinearity in the model 3. Therefore, the model reduction in (4) of part 2 is necessary.

(5). From (4) and (5) in the part 2, we obtained the final model (model 4), which was claimed to be not only the simplest but also the most appropriate model for interpretations. According to the model 4, X5, X6, X46, X56 are significant. Thus we conclude the following:

(i). The length of stay of a patient affects the square root of his (or her) cost significantly. In fact, the longer a patient stays in the hospital, the less he tends to pay. This conclusion doesn't make sense to me. This may be due to large variation in the response variable caused by varying age groups of patients. Note that we have few observations from teenage and mid-aged patients.

(ii). A patient's age affects square root of his (or her) cost significantly. In fact, the older a patient is , the more he or she tends to pay.

(iii). The interaction term between X4 and X6 affects the square root of a patient's cost significantly. e.g., Given the fact that the patients are of same age, the patient whose severity level is "catast." tends to pay more than the patients whose severity level is not "catast." do.

(iv). The interaction term between X5 and X6 affects square root of the cost significantly. e.g., Given the fact that the patients are of same age, the patient with longer staying time tends to pay more than the patients with shorter staying time.

(v). The patient's severity level does not affect square root of the cost significantly. It is surprising to me , too. The reason might be the "bad" classification rule of the severity.

(6).From the above, we may see that a "good" data analysis results in some discouraging conclusions. I suggested to the client that he should observe more data according to a statistical sampling scheme.