

Research on Administrative Cost of Pension Funds relating to the measures of size, turnover and administrative complexity of the funds

DAMILOLA TIJANI

1.0 Understanding the Data Set

The table in figure 1.1 shows the mean of each variable Y, A1, Per2 to Per7. The mean values of Per2 to Per7 are less than one and their respective standard deviations are less than one too. This indicates that the values are not widely spread from the mean. Variable A1 has a mean of 19338.822 and standard deviation of 22152.166 which is really large indicating that the values of A are large and it has a very wide spread from the mean. Variable Y whose mean is 25.568 has large values and it also has a wide spread from the mean.

The nmiss column indicates that there are no missing values in the above observations.

The histogram of the explanatory variable Y (figure 1.2) shows that the data is rightly skewed. This data is positively skewed and it shows that the percentage of people reduced as the total cost per active member increased.

Finding the mean, Standard Deviation and missing numbers of the variables

The MEANS Procedure

Variable	Label	Mean	Std Dev	N Miss
Y	total cost per active member	28.346	13.206	0
A1	Active members	19338.822	22152.166	0
Per2	Deferred pensioners per active member	0.389	0.351	0
Per3	pensioners per active member	0.480	0.247	0
Per4	starters in current year per active member	0.096	0.051	0
Per5	leavers in current year per active member	0.119	0.063	0
Per6	new pensioners in current year per active member	0.049	0.023	0
Per7	cessations in current year per active member	0.020	0.015	0

Figure 1.1

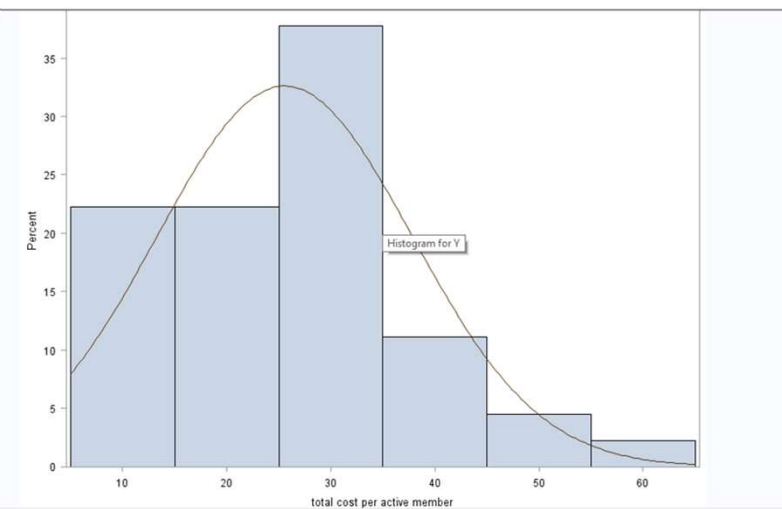


Figure 1.2

1.1 Understanding the Data Set

Explanatory variable C4 indicates that all but one customer can pay additional voluntary contribution towards their pension funds. This takes into consideration the majority of the people so there is no need to include it in regression as this data is not important to the prediction of the response variable. Because we are checking the administrative efficiency of the various schemes, the variable does not accurately take into account the members that cannot pay AVF.



A frequency distribution of C1 to C8

The FREQ Procedure

C1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	3	6.67	3	6.67
2	26	57.78	29	64.44
3	14	31.11	43	95.56
4	2	4.44	45	100.00

C2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5	11.11	5	11.11
1	40	88.89	45	100.00

C3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	7	15.56	7	15.56
1	38	84.44	45	100.00

C4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1	2.22	1	2.22
1	44	97.78	45	100.00

C5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4	8.89	4	8.89
1	41	91.11	45	100.00

C6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	33	73.33	33	73.33
1	12	26.67	45	100.00

C7	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	16	35.56	16	35.56
1	29	64.44	45	100.00

C8	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	25	55.56	25	55.56
1	20	44.44	45	100.00

2.0 Fitting a Possible Model

The data set contains one response variable which is Y and 16 explanatory variables out of which Con 1 to Con 8 are categorical variables. This means that we cannot perform a regression analysis directly on the variables without putting into consideration the different levels/categories of our variable. This is why we have created dummy variables as seen in Appendix and named them Con1, Con2 and Con3 for the 4 levels of C1. Our variable C1 has 4 levels so $4-1=3$ associated degrees of freedom.

The table produced from the fit of the model shows the analysis of variance. The F test statistics is 3.02 with 16 and 28 degrees of freedom. Its p-value = 0.0051 which is less than 0.01% level of significance. So we say that we have very strong evidence at a 0.01% level that there is a linear relationship between Y and at least one of the explanatory variables.

The adjusted Rsquare value is 0.42 which is not high enough to prove that the regression model may be useful in predicting the administrative efficiency of the various schemes from the total number of active members and other associated variables.



Multiple Regression of Y on variables A1, Per2 to Per7, C1 to C3, and C5 to C8

The REG Procedure
Model: MODEL1
Dependent Variable: Y total cost per active member

Number of Observations Read	45
Number of Observations Used	45

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	4466.08973	279.13061	2.44	0.0189
Error	28	3206.96974	114.53463		
Corrected Total	44	7673.05947			

Root MSE	10.70209	R-Square	0.5820
Dependent Mean	28.34608	Adj R-Sq	0.3432
Coeff Var	37.75508		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	37.45115	13.34356	2.81	0.0090
A1	Active members	1	-0.00031733	0.00009461	-3.35	0.0023
Per2	Deferred pensioners per active member	1	-3.16805	5.52323	-0.57	0.5708
Per3	pensioners per active member	1	17.53385	16.27040	1.08	0.2904
Per4	starters in current year per active member	1	57.67256	44.75372	1.29	0.2081
Per5	leavers in current year per active member	1	-80.04668	34.35437	-2.33	0.0272
Per6	new pensioners in current year per active member	1	3.54760	150.41400	0.02	0.9814
Per7	cessations in current year per active member	1	-175.31612	249.66633	-0.70	0.4884
Con1	C1 level 1	1	-1.01461	11.90643	-0.09	0.9327
Con2	C1 level 2	1	7.13970	9.83785	0.73	0.4740
Con3	C1 level 3	1	13.25533	9.93673	1.33	0.1930
C2	Whether scheme is contracted out (0 = no, 1 = yes)	1	11.97310	6.59032	1.82	0.0800
C3	Whether scheme is contributory (0 = no, 1 = yes)	1	-13.94461	5.85597	-2.38	0.0243
C5	Whether all administration is based at one location (0 = no, 1 = yes)	1	-7.46431	7.26214	-1.03	0.3128
C6	Whether all administrative calculations are performed on one IT platform (0 = no, 1 = yes)	1	-5.34771	4.77075	-1.12	0.2718
C7	Whether special communications are sent to members at the year end (0 = no, 1 = yes)	1	-1.26036	4.20042	-0.30	0.7664
C8	Whether rule changes are communicated directly to members (0 = no, 1 = yes)	1	-2.10191	3.88996	-0.54	0.5932

Figure 2.1

2.1 Investigating the Fit of the Model

The plot in figure 2.2 shows that the relationship between the dependent and independent variables is non linear. We will further check if the model conforms to the assumptions of regression.

The plot in figure 2.3 is used to check if the model conforms to the assumption of regression which is homoscedasticity. The studentized residuals appear to be concentrated around the mean (mean = 0) and they do not have constant variance across the entire range of fitted values.

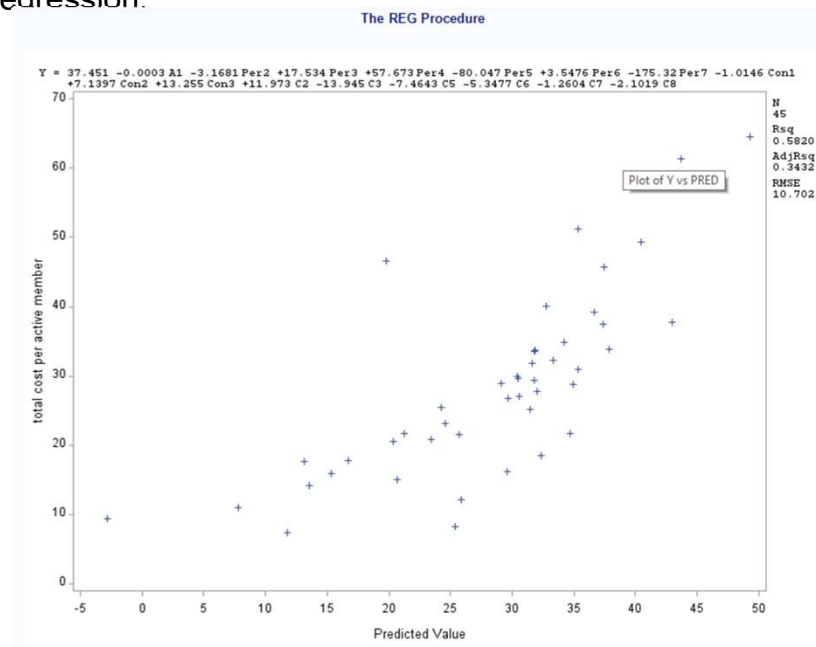


Figure 2.2

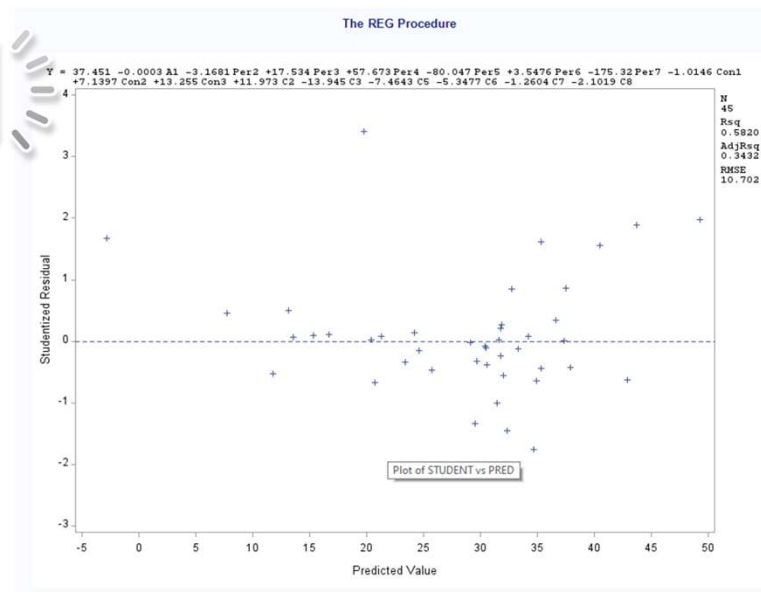


Figure 2.3

2.1 Investigating the Fit of the Model

The histogram plot is used to check the assumption of normality. The histogram is based on 45 observations. The histogram shows that the residuals produce an approximately normal distribution.

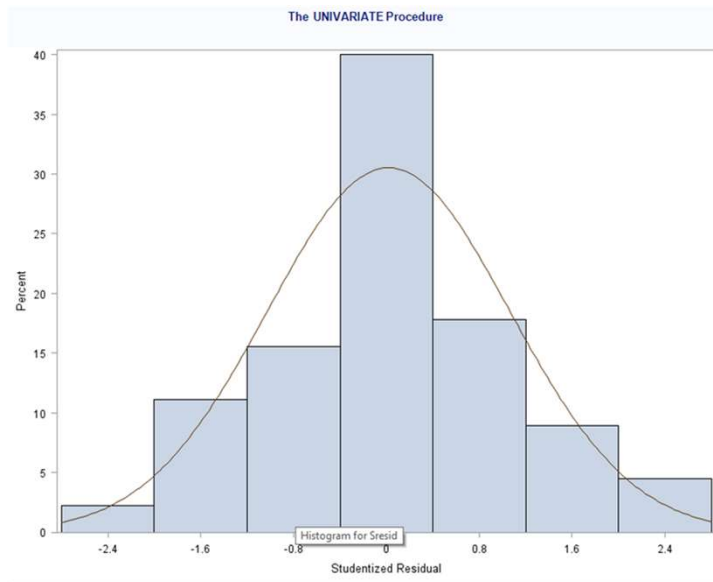


Figure 2.4

The final plot checks that the studentized residuals produce an approximately straight line. It can be seen that there are some irregularities on the plot but it can still be said to produce an approximately straight line. The assumption of near normality of the random errors is supported.

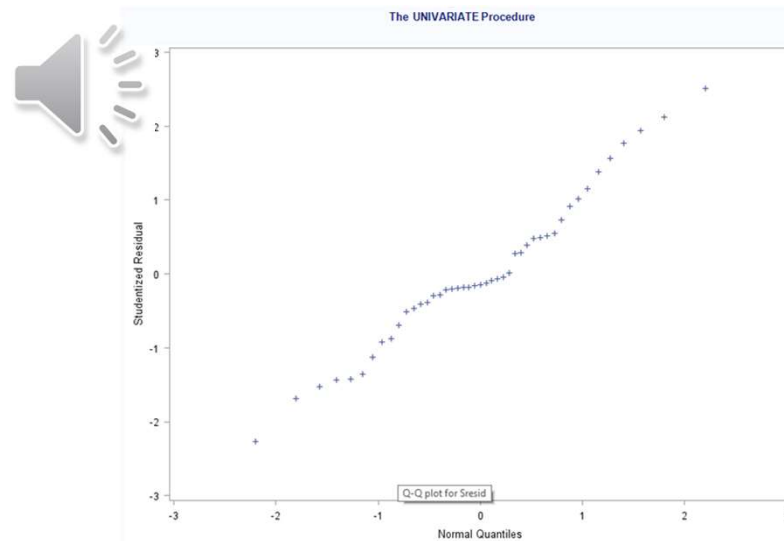



Figure 2.5

3.0 Finding a Better Model (Data Transformation)

Having fitted the above target and explanatory variables, we can see that the relationship between the dependent and independent variables is non linear so we will transform the data by finding the natural log of both response and explanatory variables. We will not transform variables C1 to C3 and C5 to C8 as these are categorical variables which should not be transformed. They are not numerical, their values are in form of an ID so they cannot be transformed.

Figure 3.1 shows the analysis of variance with the F statistic value at 6.21 with 16 and 28 degrees of freedom. The probability of the F statistic is <.0001 which is significant at a 0.01% level. This shows a strong evidence of a linear relationship between the transformed target variable Log Y and at least one of the transformed explanatory variables.

The high value of the adjusted Rsquare also indicates that multiple regression model will prove to be useful in predicting the company's administrative funds using the transformed explanatory variables.



The REG Procedure
Model: MODEL1
Dependent Variable: LY Log of variable Y

Number of Observations Read	45
Number of Observations Used	45

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	8.78434	0.54902	6.21	<.0001
Error	28	2.47453	0.08838		
Corrected Total	44	11.25887			

Root MSE	0.29728	R-Square	0.7802
Dependent Mean	3.22989	Adj R-Sq	0.6546
Coeff Var	9.20407		

Parameter Estimates					
Variable	Label	DF	Parameter Estimate	Standard Error	t Value Pr > t
Intercept	Intercept	1	6.94757	1.13084	6.14 <.0001
LA1	Log of variable A1	1	-0.44533	0.06481	-6.87 <.0001
LPer2	Log of variable Per2	1	-0.26961	0.07938	-3.40 0.0021
LPer3	Log of variable Per3	1	0.60941	0.25536	2.39 0.0240
LPer4	Log of variable Per4	1	0.11284	0.08815	1.28 0.2110
LPer5	Log of variable Per5	1	-0.08397	0.09538	-0.88 0.3861
LPer6	Log of variable Per6	1	-0.03906	0.15590	-0.25 0.8040
LPer7	Log of variable Per7	1	-0.14686	0.16705	-0.88 0.3868
Con1	C1 level 1	1	-0.59926	0.40178	-1.49 0.1470
Con2	C1 level 2	1	-0.12803	0.35599	-0.36 0.7218
Con3	C1 level 3	1	-0.00697	0.33842	-0.02 0.9837
C2	Whether scheme is contracted out (0 = no, 1 = yes)	1	0.57380	0.19661	2.92 0.0069
C3	Whether scheme is contributory (0 = no, 1 = yes)	1	-0.36133	0.17614	-2.05 0.0497
C5	Whether all administration is based at one location (0 = no, 1 = yes)	1	-0.19529	0.20106	-0.97 0.3397
C6	Whether all administrative calculations are performed on one IT platform (0 = no, 1 = yes)	1	-0.16002	0.13729	-1.17 0.2536
C7	Whether special communications are sent to members at the year end (0 = no, 1 = yes)	1	0.12436	0.11865	1.05 0.3036
C8	Whether rule changes are communicated directly to members (0 = no, 1 = yes)	1	0.06653	0.11066	0.60 0.5525

Figure 3.1

3.1 Fitting of the Model

The plot in figure 3.2 shows a better relationship between the target and response variables. The plot will produce a straight line which indicates a linear relationship between the total cost per active member and the explanatory variables. This shows the adequacy of the systematic component.

After establishing a linear relationship between our variables, we will produce appropriate plots to check if the fitted model conforms to the assumptions of multiple regression which are:

- Homoscedasticity
- Near Normality

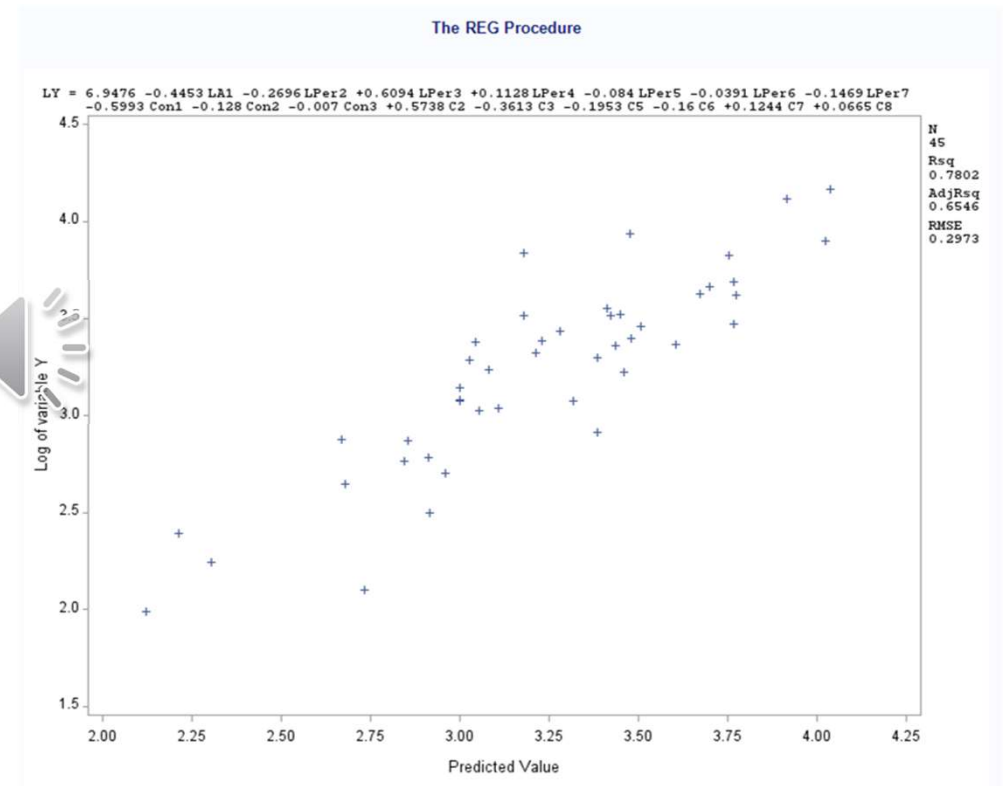


Figure 3.2

3.2 Investigating the Fit of the Model

The studentized residuals appear to now be reasonably randomly distributed about the mean value of zero and they show a constant variance across the entire range of fitted values. The plot therefore conforms to the assumption of constant variance.

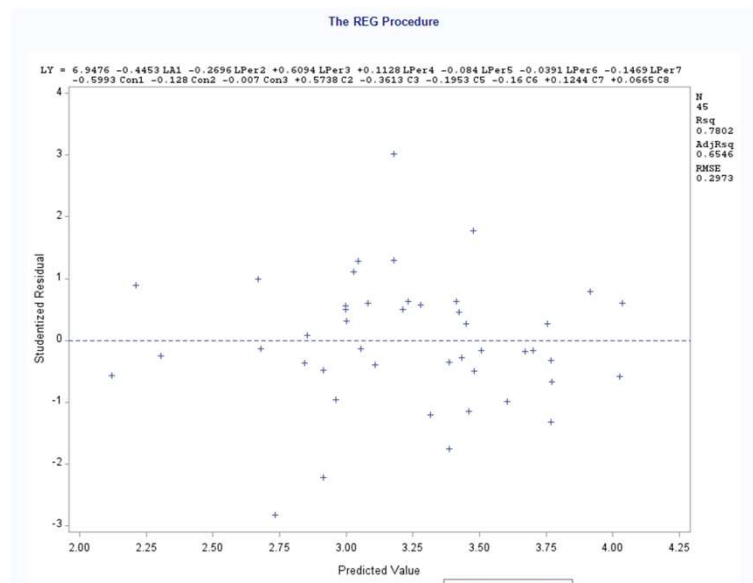


Figure 3.3

The histogram is produced from 45 observations and it produces an approximately normal probability curve which is fairly symmetrically distributed and unimodal as required.

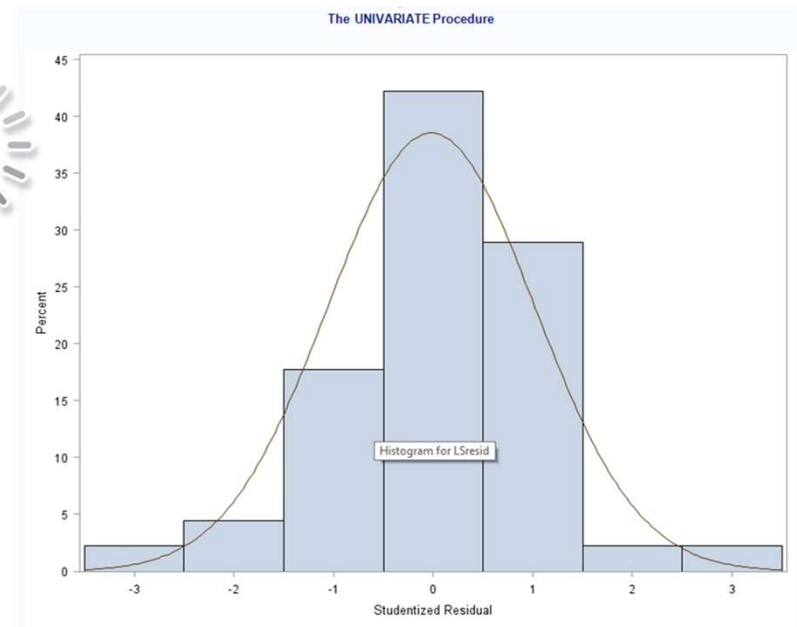


Figure 3.3

3.2 Investigating the Fit of the Model

The normal probability plot of the studentized residuals produces an approximate straight line. This conforms to the near normality of random errors as an assumption of regression.

This plot reveals a potential outlier as seen on the top right corner.

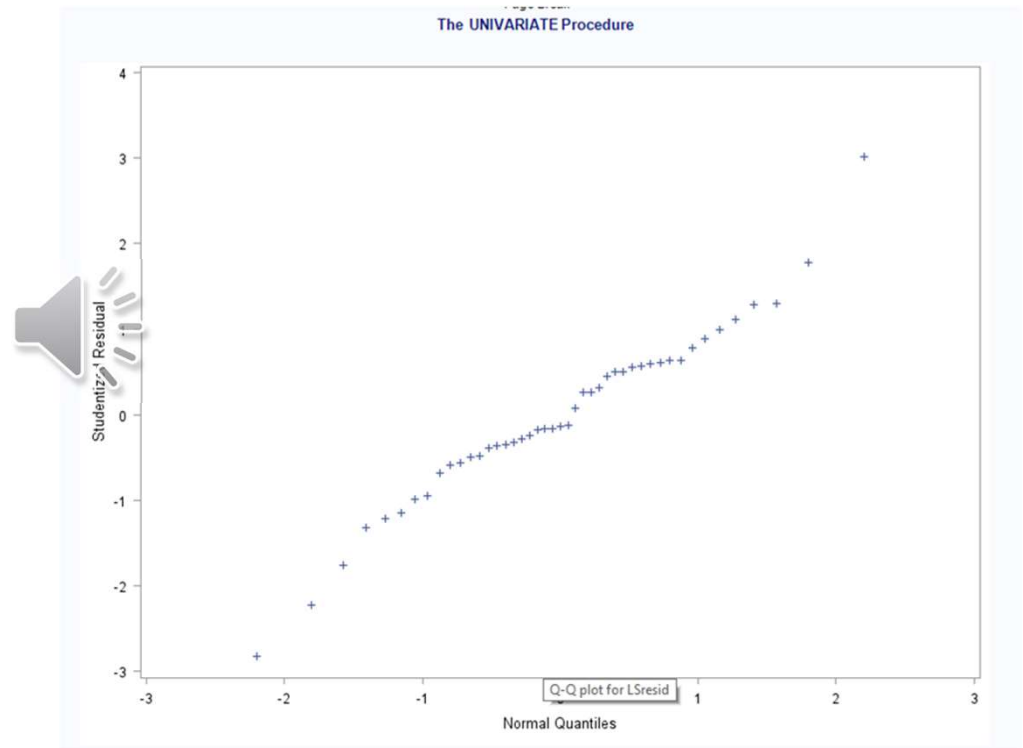


Figure 3.5

4.0 Finding the Best Model (Model Reduction/Selection)

It is possible that some of the explanatory variables make no effective contribution to the description of the response variable because it is unrelated to the response variable. This is why we will be performing model selection by fitting all possible models. This means we will produce all models containing one explanatory variable to a model with all 15 explanatory variables.

4.1 Error Mean Square Method

This table is shown to produce only the best Rsquare values. Then we will use the Mean Square Error to determine the best model. As the size of the model decreases the MSE values fluctuate between 0.088 to 0.079, until k=5 where it jumps to 0.089. We will now pick the closest value before the jump which gives the model to contain **LA1, LPer2, LPer3, Con1, Con2 and C2**

Figure 4.1 shows a reduced table that only contains the best Rsquare value for each k number of variables.

Multiple Regression of Y on variables A1, Per2 to Per7, C1 to C3, and C5 to C8

The REG Procedure
Model: MODEL1
Dependent Variable: LY

R-Square Selection Method

Number of Observations Read	45
Number of Observations Used	45

Number in Model	R-Square	MSE	Variables in Model
1	0.5291	0.12328	LA1
2	0.5899	0.10995	LA1 Con1
3	0.6168	0.10523	LA1 LPer3 Con1
4	0.6626	0.09497	LA1 LPer2 LPer3 Con1
5	0.6903	0.08942	LA1 LPer2 LPer3 Con1 C2
6	0.7113	0.08553	LA1 LPer2 LPer3 Con1 Con2 C2
7	0.7335	0.08109	LA1 LPer2 LPer3 Con1 Con2 C2 C3
8	0.7441	0.08003	LA1 LPer2 LPer3 Con1 Con2 Con3 C2 C3
9	0.7527	0.07956	LA1 LPer2 LPer3 LPer4 LPer7 Con1 C2 C3 C7
10	0.7607	0.07926	LA1 LPer2 LPer3 LPer4 LPer7 Con1 Con2 C2 C3 C7
11	0.7660	0.07983	LA1 LPer2 LPer3 LPer4 LPer7 Con1 Con2 C2 C3 C6 C7
12	0.7710	0.08055	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6
13	0.7770	0.08098	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6 C7
14	0.7797	0.08267	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6 C7 C8
15	0.7802	0.08533	LA1 LPer2 LPer3 LPer4 LPer5 LPer6 LPer7 Con1 Con2 C2 C3 C5 C6 C7 C8
16	0.7802	0.08838	LA1 LPer2 LPer3 LPer4 LPer5 LPer6 LPer7 Con1 Con2 Con3 C2 C3 C5 C6 C7 C8

Figure 4.1

4.0 Finding the Best Model (Model Reduction/Selection)

4.2 The Cp Method

Using the Mallows' Cp method, a measure of the overall sum of squares of the differences between the fitted regression model and the true model is used to generate a table. A smaller table is derived from the table of all possible fits by selecting the ones with the best Rsquare values.

From figure 4.2 we can see that as the size of k decreases the Cp value is approximately equals to or less than p ($p = k+1$) until $k = 5$ ($p = k+1=6$) where $C(p)$ value becomes 6.46 which is greater than 6. So we pick the best of the larger values which is $k=6$ which build its model using **LA1, LPer2, LPer3, Con1, Con2 and C2**.



The REG Procedure			
Model: MODEL1			
Dependent Variable: LY			
R-Square Selection Method			
Number of Observations Read		45	
Number of Observations Used		45	
Number in Model	R-Square	C(p)	Variables in Model
1	0.5291	18.9851	LA1
2	0.5899	13.2510	LA1 Con1
3	0.6168	11.8176	LA1 LPer3 Con1
4	0.6626	7.9841	LA1 LPer2 LPer3 Con1
5	0.6903	6.4591	LA1 LPer2 LPer3 Con1 C2
6	0.7113	5.7772	LA1 LPer2 LPer3 Con1 Con2 C2
7	0.7335	4.9480	LA1 LPer2 LPer3 Con1 Con2 C2 C3
8	0.7441	5.6013	LA1 LPer2 LPer3 Con1 Con2 Con3 C2 C3
9	0.7527	6.5093	LA1 LPer2 LPer3 LPer4 LPer7 Con1 C2 C3 C7
10	0.7607	7.4912	LA1 LPer2 LPer3 LPer4 LPer7 Con1 Con2 C2 C3 C7
11	0.7660	8.8096	LA1 LPer2 LPer3 LPer4 LPer7 Con1 Con2 C2 C3 C6 C7
12	0.7710	10.1679	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6
13	0.7770	11.4059	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6 C7
14	0.7797	13.0634	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6 C7 C8
15	0.7802	15.0004	LA1 LPer2 LPer3 LPer4 LPer5 LPer6 LPer7 Con1 Con2 C2 C3 C5 C6 C7 C8
16	0.7802	17.0000	LA1 LPer2 LPer3 LPer4 LPer5 LPer6 LPer7 Con1 Con2 Con3 C2 C3 C5 C6 C7 C8

Figure 4.2

4.0 Finding the Best Model (Model Reduction/Selection)

4.2 Adjusted Rsquare Method

The adjusted Rsquare method measures the proportion of variation in the target variable that is shown by the regression model. Models with larger values of adjusted Rsquare are considered to be better.

The adjusted Rsquare values remain fairly stable as k reduces between 0.6546 and 0.6903 until it reaches k=5 where the Rsquare value drops to 0.6506. So we pick the closest of the larger values which is k=6 that models **LA1, LPer2, LPer3, Con1, Con2 and C2**.

After trying to remove unnecessary variables and factors using the “all possible models” approach, we have seen that each process produced 241 observations each out of which we had to select the best of each K values. This method is considered time wasting and it also wastes computer resources as many models are fitted that are not even considered for selection.

After trying the above methods, we will use another elimination procedure to select an appropriate model.



The REG Procedure			
Model: MODEL1			
Dependent Variable: LY			
R-Square Selection Method			
Number of Observations Read		45	
Number of Observations Used		45	
Number in Model	R-Square	Adjusted R-Square	Variables in Model
1	0.5291	0.5182	LA1
2	0.5899	0.5703	LA1 Con1
3	0.6168	0.5888	LA1 LPer3 Con1
4	0.6626	0.6289	LA1 LPer2 LPer3 Con1
5	0.6903	0.6506	LA1 LPer2 LPer3 Con1 C2
6	0.7113	0.6657	LA1 LPer2 LPer3 Con1 Con2 C2
7	0.7335	0.6831	LA1 LPer2 LPer3 Con1 Con2 C2 C3
8	0.7441	0.6872	LA1 LPer2 LPer3 Con1 Con2 Con3 C2 C3
9	0.7527	0.6891	LA1 LPer2 LPer3 LPer4 LPer7 Con1 C2 C3 C7
10	0.7607	0.6903	LA1 LPer2 LPer3 LPer4 LPer7 Con1 Con2 C2 C3 C7
11	0.7660	0.6880	LA1 LPer2 LPer3 LPer4 LPer7 Con1 Con2 C2 C3 C6 C7
12	0.7710	0.6852	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6
13	0.7770	0.6835	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6 C7
14	0.7797	0.6769	LA1 LPer2 LPer3 LPer4 LPer5 LPer7 Con1 Con2 C2 C3 C5 C6 C7 C8
15	0.7802	0.6665	LA1 LPer2 LPer3 LPer4 LPer5 LPer6 LPer7 Con1 Con2 C2 C3 C5 C6 C7 C8
16	0.7802	0.6546	LA1 LPer2 LPer3 LPer4 LPer5 LPer6 LPer7 Con1 Con2 Con3 C2 C3 C5 C6 C7 C8

Figure 4.3

4.1 Backward Elimination Procedure

The backward elimination model in comparison the the above elimination methods (all possible models) is better because it does not use as much CPU space . Also, early decisions are made to remove the variables with the highest non significant p-value. Instead of having to deal with removing each non-significant model manually, this is done by the computer and this produces a faster process. Furthermore, the backward selection model produces only 5 explanatory variables and it is easier to determine a final model using this procedure.

From the model selection in the previous section, we have gotten our final model to be

$$LY = LA1 + LPer2 + LPer3 + Con1 + C2$$

From figure 4.4, the error mean square represented as s^2 remains fairly stable between 0.0900 and 0.075. When the variables are reduced to 4 it increases to 0.094, showing that the model is inferior to the other ones so we take the good model with the lowest explanatory variable 5.

Now, we will fit the regression model to obtain our parameter estimates.

k	s^2
16	0.08838
15	0.08533
14	0.08267
13	0.08098
12	0.08055
11	0.08037
10	0.07966
9	0.07962
8	0.08028
7	0.08377
6	0.08733
5	0.08942
4	0.09497

Figure 4.4

4.2 Understanding the Parameters

From fitting the final regression model , we have obtained our parameter estimates as shown in figure 4.5. However, these figures we have gotten are in natural log form so we need to convert them before using.

$$LY = 7.160 - 0.444\text{LA1} - 0.182\text{LPer2} + 0.262\text{LPer3} - 0.521\text{Con1} + 0.285\text{C2}$$

From the formula, we can say that the regression line has an intercept of 7.160 (the point where it meets line Y).

Variables LA1, LPer2 and Con1 have a negative relationship with the response variables ie when Y increases, they decrease and vice versa.

Variables C2 and LPer3 have positive relationships with the response variable Y ie as Y increases, they increase and vice versa.

The REG Procedure
Model: MODEL1
Dependent Variable: LY Log of variable Y

Number of Observations Read	45
Number of Observations Used	45

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7.77163	1.55433	17.38	< .0001
Error	39	3.48724	0.08942		
Corrected Total	44	11.25887			

Root MSE	0.29903	R-Square	0.6903
Dependent Mean	3.22989	Adj R-Sq	0.6506
Coeff Var	9.25808		

Parameter Estimates					
Variable	Label	DF	Parameter Estimate	Standard Error	t Value Pr > t
Intercept	Intercept	1	7.16004	0.45299	15.81 < .0001
LA1	Log of variable A1	1	-0.44446	0.04795	-9.27 < .0001
LPer2	Log of variable Per2	1	-0.18152	0.06332	-2.87 0.0067
LPer3	Log of variable Per3	1	0.26208	0.08811	2.97 0.0050
Con1	C1 level 1	1	-0.52082	0.18878	-2.76 0.0088
C2	Whether scheme is contracted out (0 = no, 1 = yes)	1	0.28542	0.15291	1.87 0.0695

Figure 4.5

5.0 Investigating the Fit of the Model

After producing the best model, we will investigate the model to see if it conforms to the assumptions of regression.

Figure 5.1 shows that the studentized residuals appear to be randomly scattered across the mean value of 0 and it has a constant variance (spread) across the entire range of values.

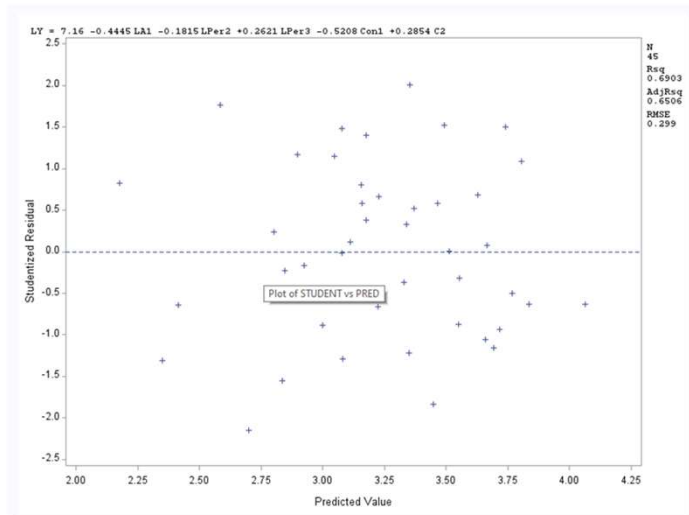


Figure 5.1

The histogram of residuals conform to the assumption of normality of the studentized residuals.

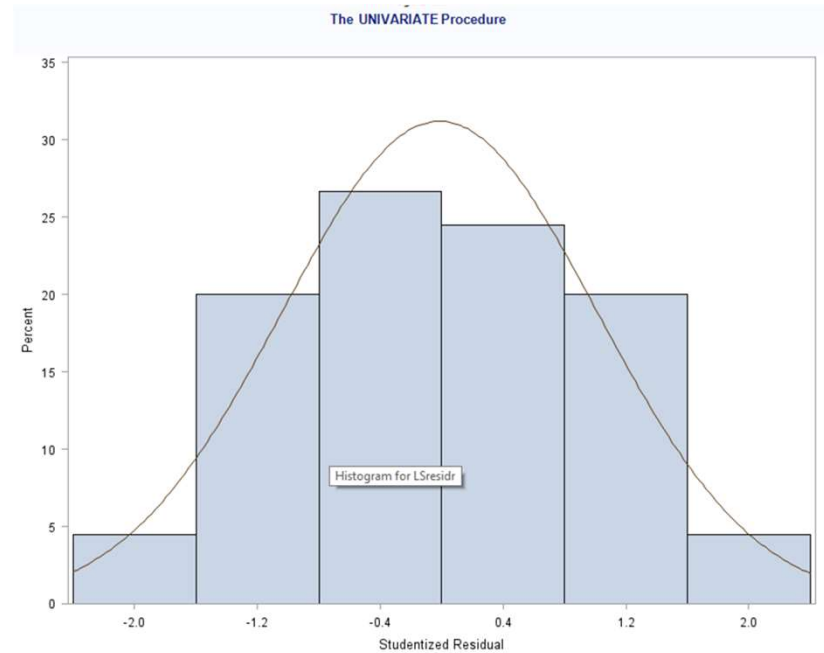


Figure 5.2

5.0 Investigating the Fit of the Model

The plot of studentized residuals produce an approximately straight line of unit slope that will pass near the origin.

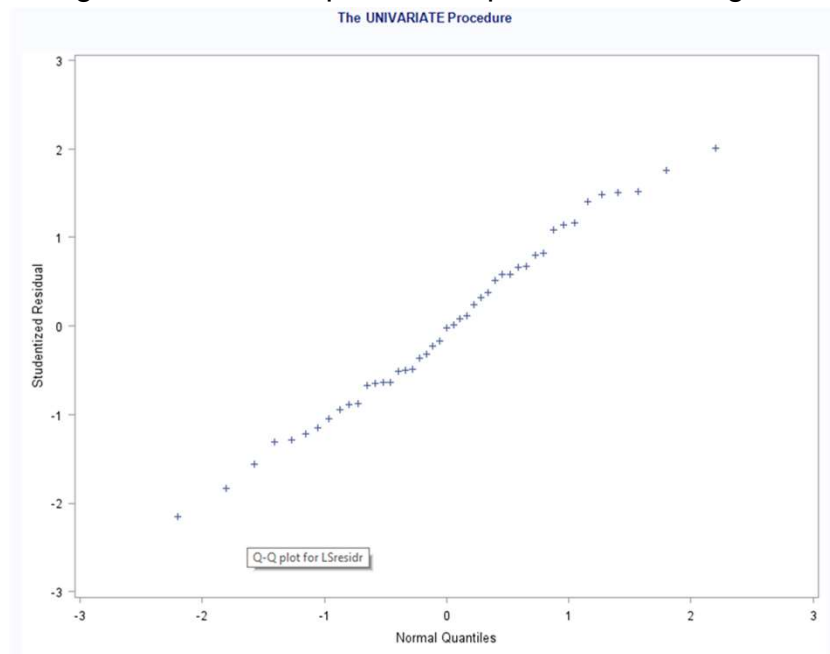


Figure 5.3

We will also check the fit of the response variable with each explanatory variable. This is shown below:

The plot produced shows that the studentized residuals appear to be randomly scattered around the mean value of zero and with constant variance across the entire range of explanatory variable **LA1**. The plot therefore satisfies the assumption of linear association and constant variation between **LY** and **LA1**.

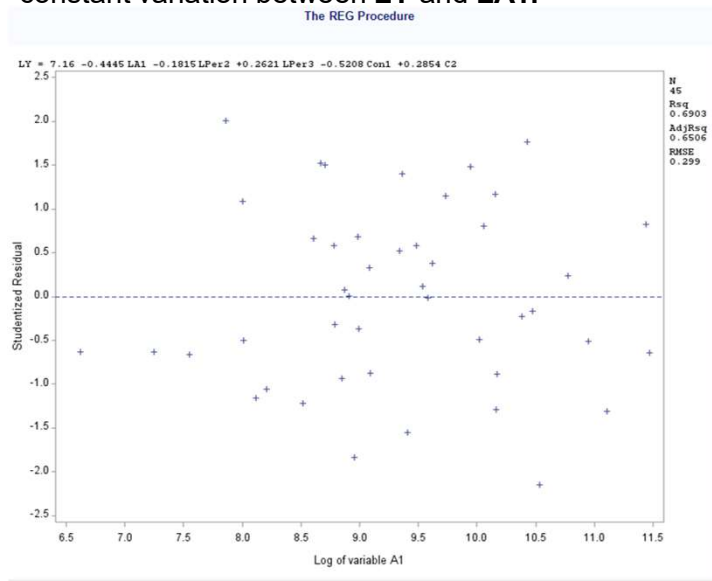


Figure 5.4

5.0 Investigating the Fit of the Model

The plot produced shows that the studentized residuals appear to be randomly scattered around the mean value of zero and with constant variance across the entire range of explanatory variable **LPer2**. The plot therefore satisfies the assumption of linear association and constant variation between **LY** and **LPer2**.

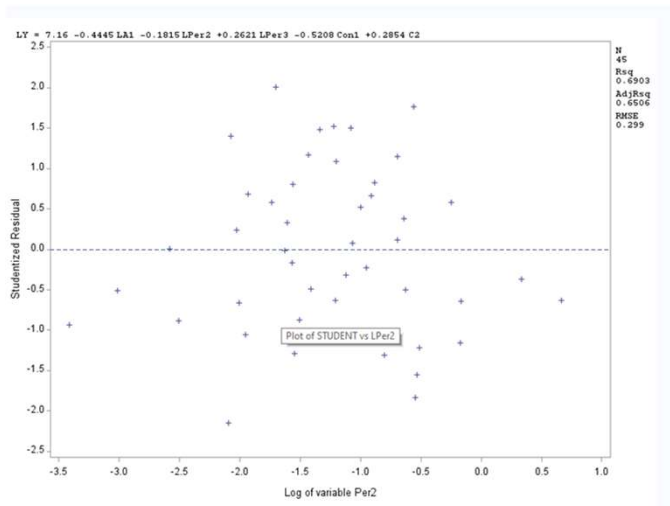


Figure 5.5

The plot produced shows that the studentized residuals appear to be randomly scattered around the mean value of zero but without constant variance across the entire range of explanatory variable **LPer3**. The plot therefore satisfies the assumption of linear association but not constant variation between **LY** and **LPer3**.

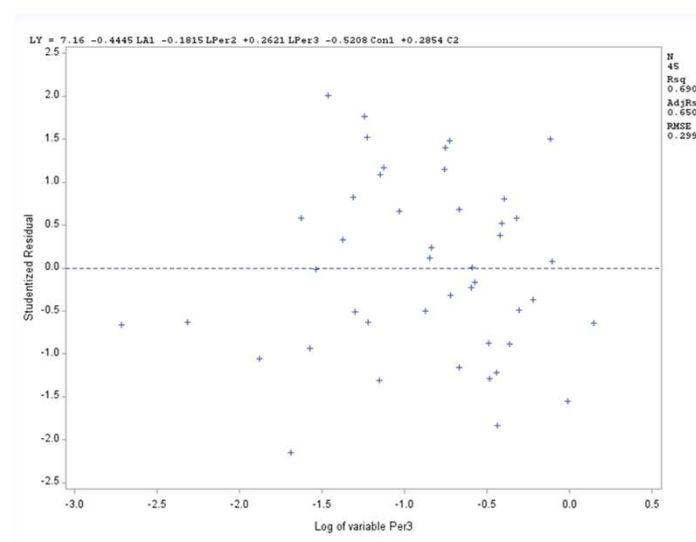


Figure 5.6

6.0 Investigating the Model for Outliers and Influential Points

To further verify the efficiency of our model, we need to check the possibility of potential outliers in our dataset. This helps to check if there are data points or observations that will influence our model. This is checked by plotting the Deleted and studentized residuals against the predicted values as seen below:

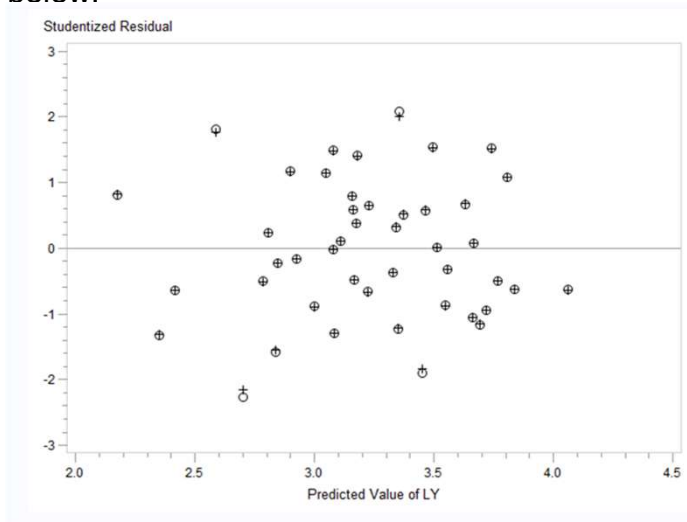


Figure 6.1

For most observations in our plot, we can see that the studentized and deleted residuals are at the same point. Only a few have a small degree of separation with the deleted residual being the most extreme value. From this, we can see that only about 2 observations are outside the range of plus or minus 2 so it is unlikely that there are any outliers in our data set.

To be certain there are no observations that can possibly assert influence on our fitted regression equation, we will carry out other diagnostic procedures. These processes will help to identify influential points in the model. The procedures we will use are:

- Leverage H
- Covariance ratio
- DFFITS

6.0 Investigating the Model for Outliers and Influential Points

The first few lines of code are used to create the Dffits values and sort the data. From our data set, we have $n = 45$ observations, with $k=4$ variables, hence we have $p = k + 1 = 5$ parameters in our regression model. The cut off for DFFITS is calculated as $2(5/45) = 0.730$.

The scatter plot is created from the proc gplot statement. Drawing a line at the cut off, we can see that only the first two leftmost points stand out on the graph but do not exceed the overall cut off of 2.

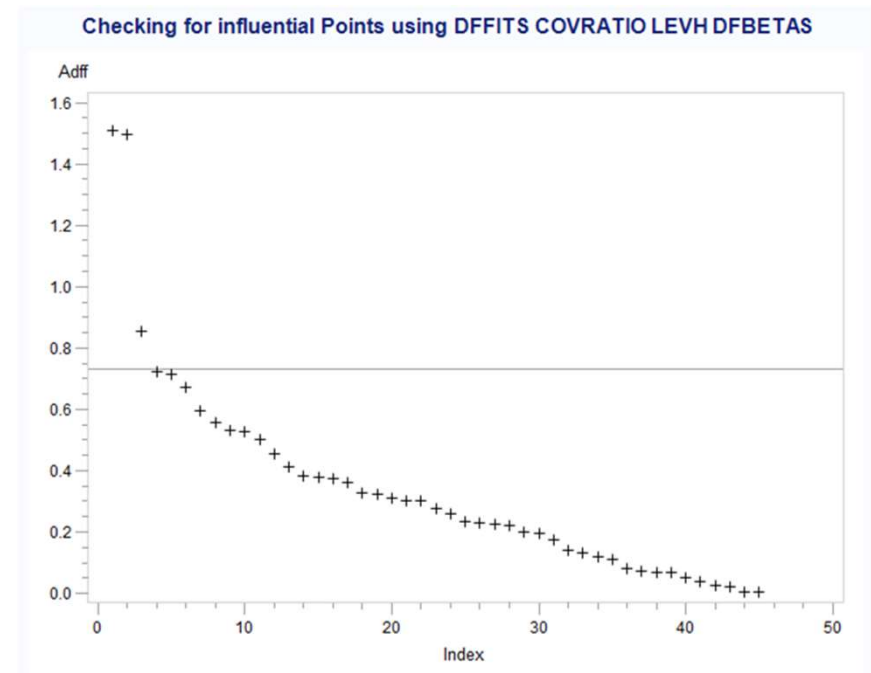


Figure 6.2

6.0 Investigating the Model for Outliers and Influential Points

We will further investigate all 3 observations even though the third point doesn't stand out and is unlikely to be cause for concern.

The proc print statement shows the observations that are potential influential point so we further investigate by producing the table of the Leverage H, Covariance Ratio.

To calculate the cut off of leverage H , $n=45$ and $p=5$

Average value $p/n = 5/45=0.13$

Cut off value $3p/n = 15/45=0.4$

Checking for influential Points using DFFITS COVRATIO LEVH DFBETAS

Obs	ID	LPredr	LY	Dff
1	3	3.35401	3.84028	1.50861
2	31	2.83676	2.49848	-1.49533
3	45	2.70159	2.09924	-0.85457
4	29	3.17739	3.55105	0.72202

Figure 6.3

Checking for influential Points using DFFITS COVRATIO LEVH DFBETAS

Obs	ID	Dff	H	Dresid	C
1	3	1.50861	0.34164	2.09300	1.01240
2	31	-1.49533	0.47081	-1.51018	1.61421
3	45	-0.85457	0.12463	-2.16728	0.73425
4	29	0.72202	0.05649	0.52385	1.16146

Figure 6.4

The limits for COVRatio are

$1 \pm 3p/n = 0.60$ and 1.40 respectively.

From the table, the observation for ID=3 has a value that is just slightly above the leverage H and a deleted residual of approximately 2. Its covariance ratio is approximately 1, indicating that the inclusion of this point has little effect on the precision of the fitted multiple regression. On further investigation, this point therefore causes little concern.

The observation for ID=31 has a value that is just slightly above the leverage H cut off and a moderately large deleted residual (but with an absolute value below 2). Its covariance ratio is above 1.33, indicating that the inclusion of this point has an effect on the precision of the fitted multiple regression. On further investigation, this point might be a cause for concern.

The observation for ID=45 has a value that is just slightly above the leverage H average and a moderately large deleted residual (with an absolute value above 2). Its covariance ratio is within range although below 1, indicating that the inclusion of this point slightly reduces precision of the fitted multiple regression. On further investigation, this point causes little concern.

The observation for ID=29 has a value that is just slightly above the leverage H cut off and a moderately large deleted residual (but with an absolute value below 2). Its covariance ratio is above 1.33, indicating that the inclusion of this point has an effect on the precision of the fitted multiple regression. On further investigation, this point might be a cause for concern.

6.0 Investigating the Model for Multicollinearity

Because of the large number of explanatory variables, there is an increased chance of inter relationships existing between the variables. We will use multicollinearity to determine the relationships between the explanatory variables.

This will produce tables to check the following:

Correlation

Variance Inflation factor

Condition Index

From figure 6.5, the Correlation matrix does not show any of the explanatory variables to have a high/noteworthy relationship with each other. It also shows that variable LA1

Correlation							
Variable	Label	LA1	LPer2	LPer3	Con1	C2	LY
LA1	Log of variable A1	1.0000	-0.1785	0.3261	-0.2595	0.0685	-0.7274
LPer2	Log of variable Per2	-0.1785	1.0000	0.2761	-0.0521	0.2019	0.0148
LPer3	Log of variable Per3	0.3261	0.2761	1.0000	-0.2267	-0.1477	-0.0476
Con1	C1 level 1	-0.2595	-0.0521	-0.2267	1.0000	0.0945	-0.0492
C2	Whether scheme is contracted out (0 = no, 1 = yes)	0.0685	0.2019	-0.1477	0.0945	1.0000	-0.0168
LY	Log of variable Y	-0.7274	0.0148	-0.0476	-0.0492	-0.0168	1.0000

Figure 6.5

Figure 6.6 below shows that the VIF values above all are below 10 so there is no evidence of a severe dependency of any variable on the other variables.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	Intercept	1	7.16004	0.45299	15.81	<.0001	0
LA1	Log of variable A1	1	-0.44446	0.04795	-9.27	<.0001	1.35630
LPer2	Log of variable Per2	1	-0.18152	0.06332	-2.87	0.0067	1.31999
LPer3	Log of variable Per3	1	0.26208	0.08811	2.97	0.0050	1.40957
Con1	C1 level 1	1	-0.52082	0.18878	-2.76	0.0088	1.11592
C2	Whether scheme is contracted out (0 = no, 1 = yes)	1	0.28542	0.15291	1.87	0.0695	1.16222

Figure 6.6

The only row with a condition index 4 times more than the previous one is row 5 but its value is below 30. The variables with highest loadings are LA1 and the Intercept. However, given the low VIF's obtained in the above table and the fact that row 5 only meets one

Collinearity Diagnostics							
Number	Eigenvalue	Condition Index	Proportion of Variation				
			Intercept	LA1	LPer2	LPer3	Con1
1	4.52344	1.00000	0.00045195	0.00045105	0.00866	0.00879	0.00517
2	0.91290	2.22599	0.00012709	0.00022337	0.00171	0.00063810	0.87210
3	0.27717	4.03982	0.00209	0.00230	0.38519	0.12220	0.01223
4	0.22337	4.50012	0.00060370	0.00202	0.26132	0.64765	0.05691
5	0.05810	8.82340	0.03402	0.02797	0.27402	0.04129	0.00914
6	0.00502	30.02076	0.96270	0.96704	0.06911	0.17943	0.04445

Figure 6.7

7.0 Conclusion

Now that we have been able to fit a final model and investigate it through diagnostic procedures, the model can now be confidently applied to predicting the administrative costs of pension funds, in relation to the cost, measures of the size, turnover and administrative complexity of the funds.

We will be using prediction limits so we can be able to take into account the variations in the schemes about the underlying mean.

Predictions made from this model will help a pension fund manager predict future values of the administrative efficiency from provided details of no of active members, number of deferred pensioners per active member ,number of pensioners per active member, fund type and whether the scheme is contracted out.

The confidence limits will give a range of values the predicted variable should fall in.