Pension Fund Administrative Costs Analysis

Conducted by

DATT January Cohort GROUP 17

Sheffield Hallam University

Background

Pension Fund Accountants Research Group, PFARG has asked us, as a team of business intelligence to investigate the administrative cost efficiency of pension funds for various schemes relating to size, turnover and administrative complexity. This will help pension managers to optimize administrative cost for various pension schemes.

Total cost per active member has been identified as the relevant performance indicator:

Y = 1000(B3 + B4 + B6 + B7 + B8)/A1

NB: B5 has been excluded as this is not in the control of the pension manager.

PFARG has also provided dataset which contain 45 observations for each variables and factors attached below to work with:

C:\Users\JOBA\
Drive - Sheffield Ha

VARIABLE LIST AND DEFINITIONS **Measures of current size & current fund turnover**	
ID	Fund Identification Number
A1	Number of active members
A2	Number of deferred pensioners
A3	Number of pensioners
A4	Number of starters in current year
A5	Number of leavers in current year
A6	Number of new pensioners in current year
B3	Staff cost (gross salaries and wages, £'000)
B4	Staff cost (oncost, £'000)
B5	Premises costs (£'000)
B6	Establishment costs (£'000)
B7	External fees (£'000)
B8	IT costs (£'000)

FACTORS LIST AND DEFINITIONS **Measures of administrative complexity**	
C1	Fund type: 1 = staff only 2 = combined scheme, same scales 3 = separate schemes 4 = combined scheme, different scales
C2	Whether scheme is contracted out (0 = no, 1 = yes)
C3	Whether scheme is contributory (0 = no, 1 = yes)
C4	Whether members can pay AVC's (0 = no, 1 = yes)
C5	Whether all administration is based at one location (0 = no, 1 = yes)
C6	Whether all administrative calculations are performed on one IT platform (0 = no, 1 = yes)
C7	Whether special communications are sent to members at the year end (0 = no, 1 = yes)
C8	Whether rule changes are communicated directly to members (0 = no, 1 = yes)

Exploratory Data Analysis

- some explanatory variables has been added to the provided PFARG06 data set in order to fit a multiple regression model of Y. This variables are described in column 1 & 2 of Fig. 1.

Some observations from a brief explanatory data analysis of these variables with 45 observations represented in Fig 1 and Plot 1 are stated below:

- the average pension fund administrative cost is £28.336 with a standard deviation of £13.206 connotes that the administrative cost is somewhat clustered around the mean.
- Also, the average number of active members which is 19,338.882 is observed to be less than the standard deviation of 22,152.166 which shows that individual records of this variables are highly dispersed.
- The N Miss column of Fig 1 used to check for missing values in the data set shows that there are missing values.
- The resulting histogram as shown in plot 1 shows that the distribution is positively skewed, with few observations having high values. This suggests that there may be some influential observations that are driving the overall relationship between response variable Y and the explanatory variables.
- The histogram also shows that the distribution of Y appears to be approximately normal.

Fig 1. Sample mean and Standard Deviation of Explanatory Variables

Page Break

Sample Mean and Standard Deviation of Variables

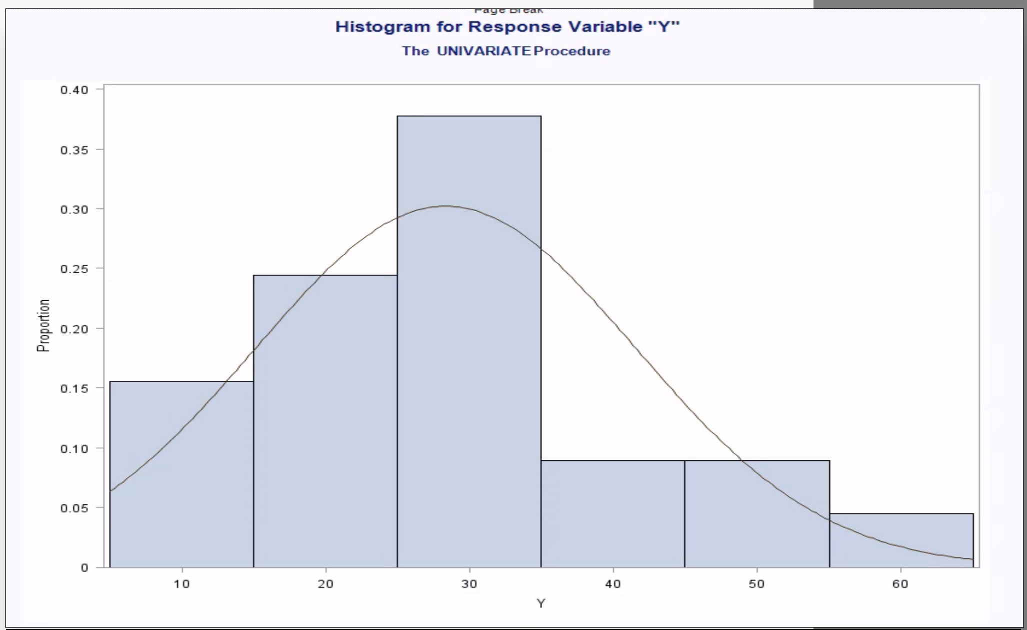
The MEANS Procedure

Variable	Label	N	Mean	Std Dev	N Miss
Y	Admin Cost	45	28.346	13.206	0
A1	Active Member	45	19338.822	22152.166	0
Per2	Deferred Pens/Active Members	45	0.389	0.351	0
Per3	Pens/Active Members	45	0.480	0.247	0
Per4	Starters/Active Members	45	0.096	0.051	0
Per5	leavers/Active Members	45	0.119	0.063	0
Per6	New Pensioner/Active Members	45	0.049	0.023	0
Per7	Cessation/Active Members run	45	0.020	0.015	0

Page Break

Sample Mean and Standard Deviation of Variables

n_missing	p_missing
45	100%



Plot 1. Histogram for Response Variable Y

Frequency Distribution of Explanatory Factors

Fig.1b. shows the frequency distribution for each categorical explanatory factor.

There are 8 categorical explanatory variables. C2-C8 are binary categorical data and C1 (Fund type) has 4 levels.

The percentage contribution of each categorial data is shown in the frequency distribution data. E.g. 57.78% of C1 (fund type) are combined schemes, same scales and 88.89% of schemes are contracted out. 84.4% of schemes are contributory, 91.1% of all administration is based on location, 73.33% of all administration are performed on one IT platform, 64.4% of special communications are sent to member at the end of the year and 55.56% of changes in rule are communicated directly to members.

However, based on the frequency distribution for C4(additional voluntary contribution), we can see that C4 **has very little variability** and should not be included in any regression model for Y because including this additional voluntary contributions can result in **high chances of missing values.**

Sample Frequency Distribution for factors

The FREQ Procedure

Fund Type

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

Is Scheme Contracted Out?

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

Is Scheme Contributory?

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

Can members pay AVCs?

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

Are all administration based at one location?

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

All Administrative calculations performed on one IT platform?

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

Special Communication sent to members at the end of the year?

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

Rule changes communicated directly to members? run

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

Fig 1b. Sample Frequency Distribution for explanatory categorical factors

FITTING THE MODEL

- The observation appears reasonably randomly scattered about the reference line which indicates that the multiple linear regression model adequately describes the systematic variation present in the response Y across the entire range of predicted values.

  Model 1 Equation is shown below:

- **$Y = 37.451 - 0.000317A1 - 3.168Per2 + 17.533Per3 + 57.672Per4 - 80.047Per5 + 3.548Per6 - 175.316Per7 - 1.015C11 + 7.139C12 + 13.255C13 + 0.000C14 + 11.973C2 - 13.945C3 - 7.464C5 - 5.348C6 - 1.260C7 - 2.102C8$**

- **R-Square = 0.582;** Only **58.2%** of the variation in the response variable Y is been explained by the explanatory variables.

Fig.2. Parameter Estimate of the fitted model 1

Page Break

MODEL One (Normal)

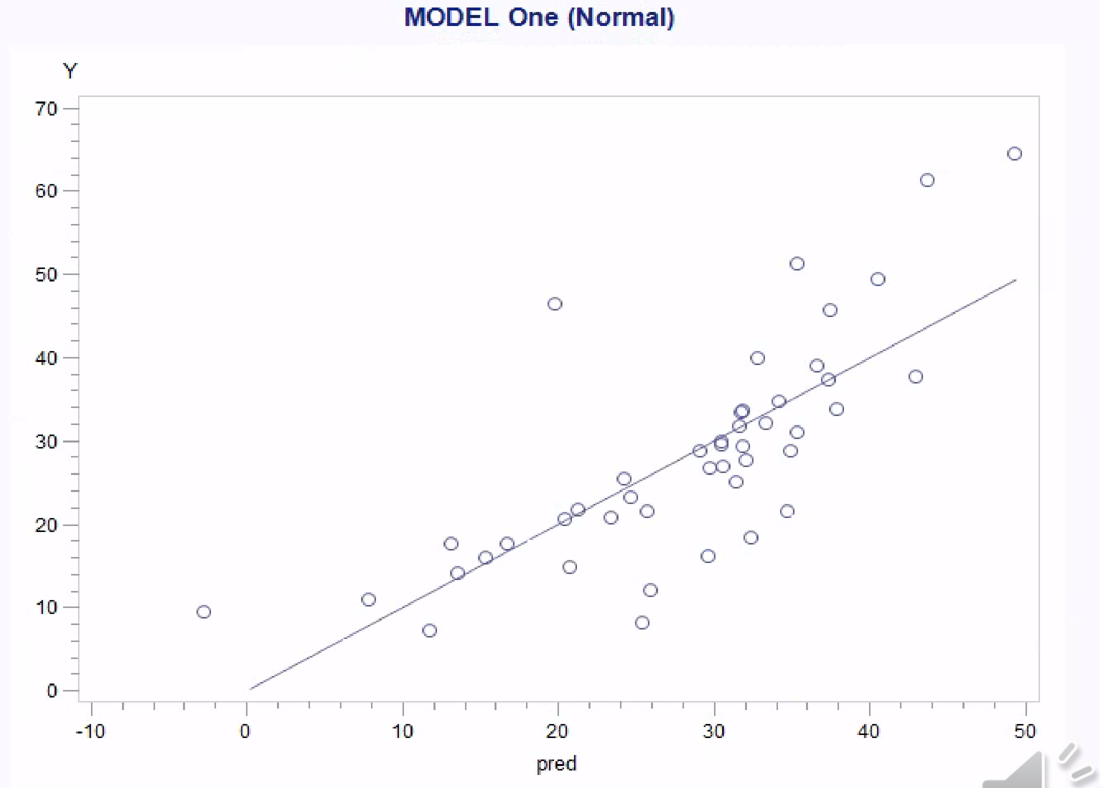
The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	4466.089731	279.130608	2.44	0.0189
Error	28	3206.969744	114.534634		
Corrected Total	44	7673.059475			

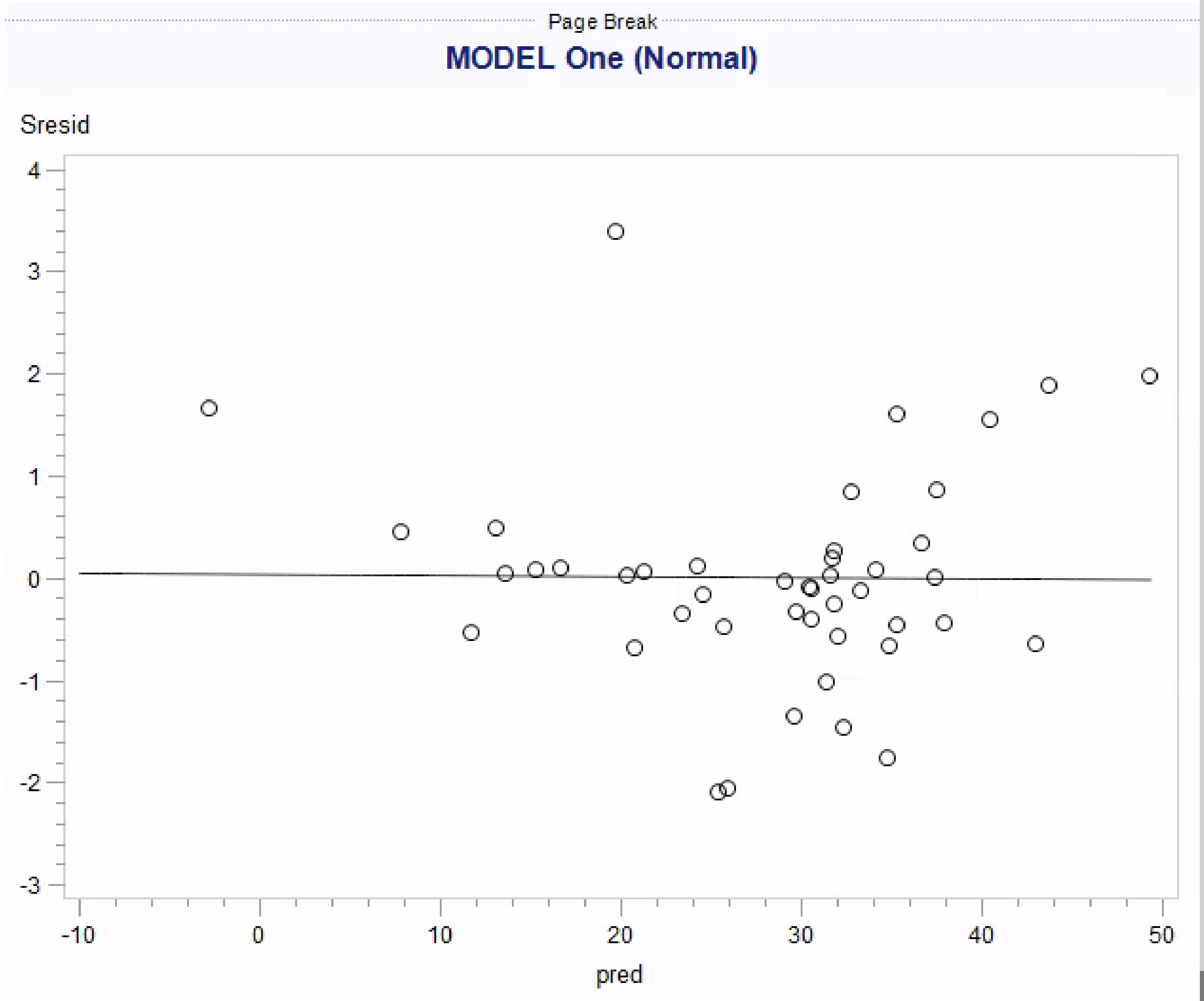
R-Square	Coeff Var	Root MSE	Y Mean
0.582048	37.75508	10.70209	28.34608

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		37.4511486	B	13.3435580	2.81	0.0090
A1		-0.0003173		0.0000946	-3.35	0.0023
Per2		-3.1680502		5.5232263	-0.57	0.5708
Per3		17.5338537		16.2703999	1.08	0.2904
Per4		57.6725597		44.7537225	1.29	0.2081
Per5		-80.0466835		34.3543663	-2.33	0.0272
Per6		3.5476048		150.4139955	0.02	0.9814
Per7		-175.3161165		249.6663333	-0.70	0.4884
C1	1	-1.0146091	B	11.9064281	-0.09	0.9327
C1	2	7.1397006	B	9.8378463	0.73	0.4740
C1	3	13.2553328	B	9.9367337	1.33	0.1930
C1	4	0.0000000	B	.	.	.
C2		11.9731034		6.5903190	1.82	0.0800
C3		-13.9446133		5.8559694	-2.38	0.0243
C5		-7.4643071		7.2621440	-1.03	0.3128
C6		-5.3477058		4.7707533	-1.12	0.2718
C7		-1.2603555		4.2004180	-0.30	0.7664
C8		-2.1019145		3.8899614	-0.54	0.5932



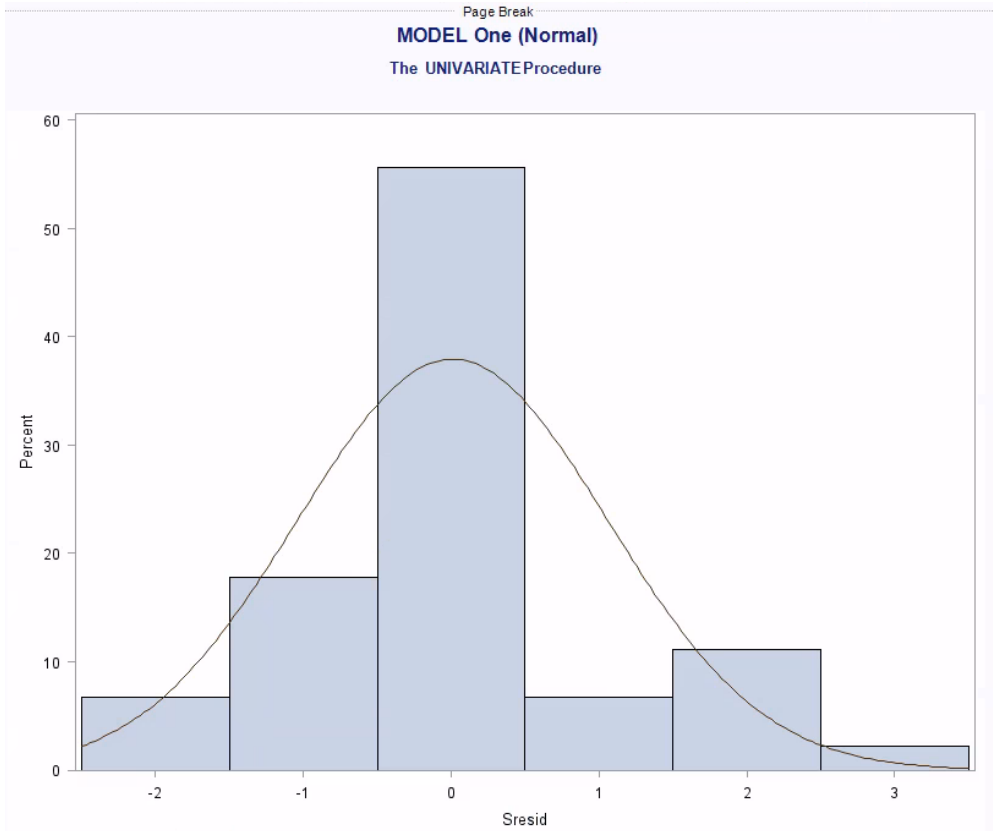
Plot 2. Scattered plot of Response Variable Y versus Predicted

Plot 2a. A scattered plot of Studentised values versus Predicted (fitted) values

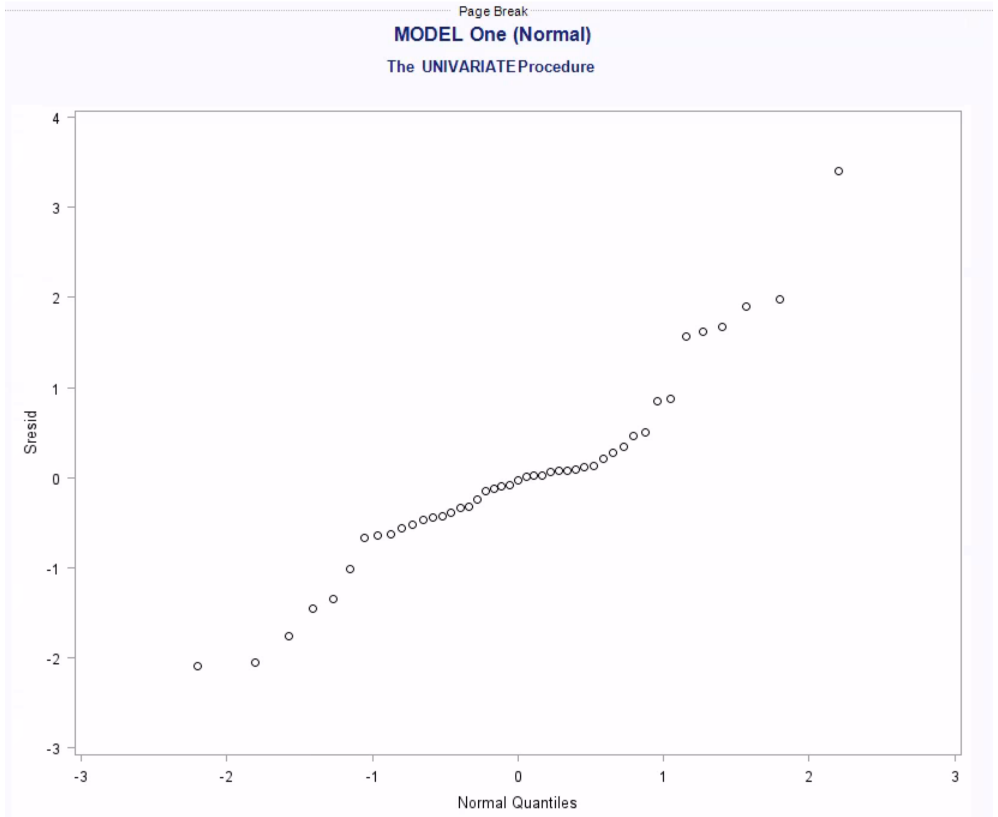


Statistical Assumptions ofo Model 1
-**Homoscedasticity**

- Plot 2a shows that the Studentised residual values appear to be randomly scattered about a mean value of zero.
- This is in line with the assumption of constant variance and the sufficiency of the systematic component of the multiple regression.



Plot 2b. A Histogram of Studentised Residual



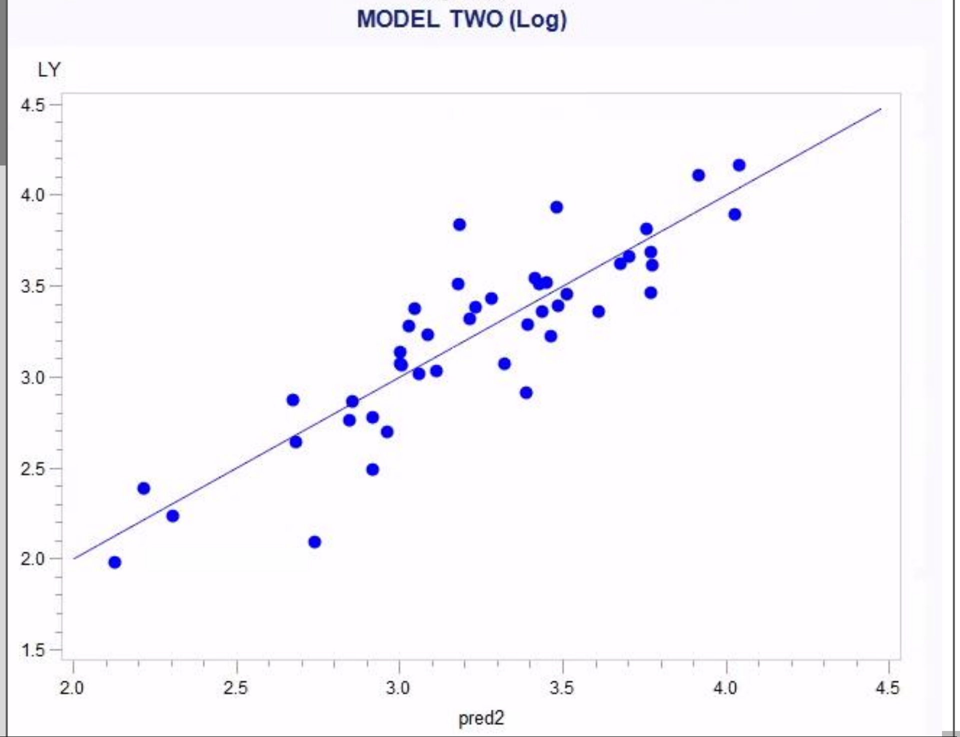
Plot 2c. A normal probability plot (ggplot) of Studentised Residual

Statistical Assumption of Model 1 *contd. – NEAR NOMALITY*

- The histogram plot of the Studentised Residual for Model 1 assumes a near normality distribution with not-so-regular positively skewed data. This might be due to the small sample size, as such, there is no cause for alarm.
- From gg plot 2c, we see that the Studentised residual values conform to an approximate straight line (reference line placed by an eye). However, there exist some irregularities in this plot possibly due relatively small sized sample.
- From both plots, the assumption of near normality is supported.

Transformation

- Transformation of features is done to bring all the features to similar scale.
- The factors C1-C3 & C5-C8 does not have much scale difference hence there is no need to apply techniques for transformation/standardization.
- Taking the log of 1 yield 0, and the log of 0 is undefined as it is not a real number



Plot 3a. A plot of log of actual versus log of predicted

MODEL TWO (Log)

The GLM Procedure

Dependent Variable: LY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	8.78434146	0.54902134	6.21	<.0001
Error	28	2.47453310	0.08837618		
Corrected Total	44	11.25887456			

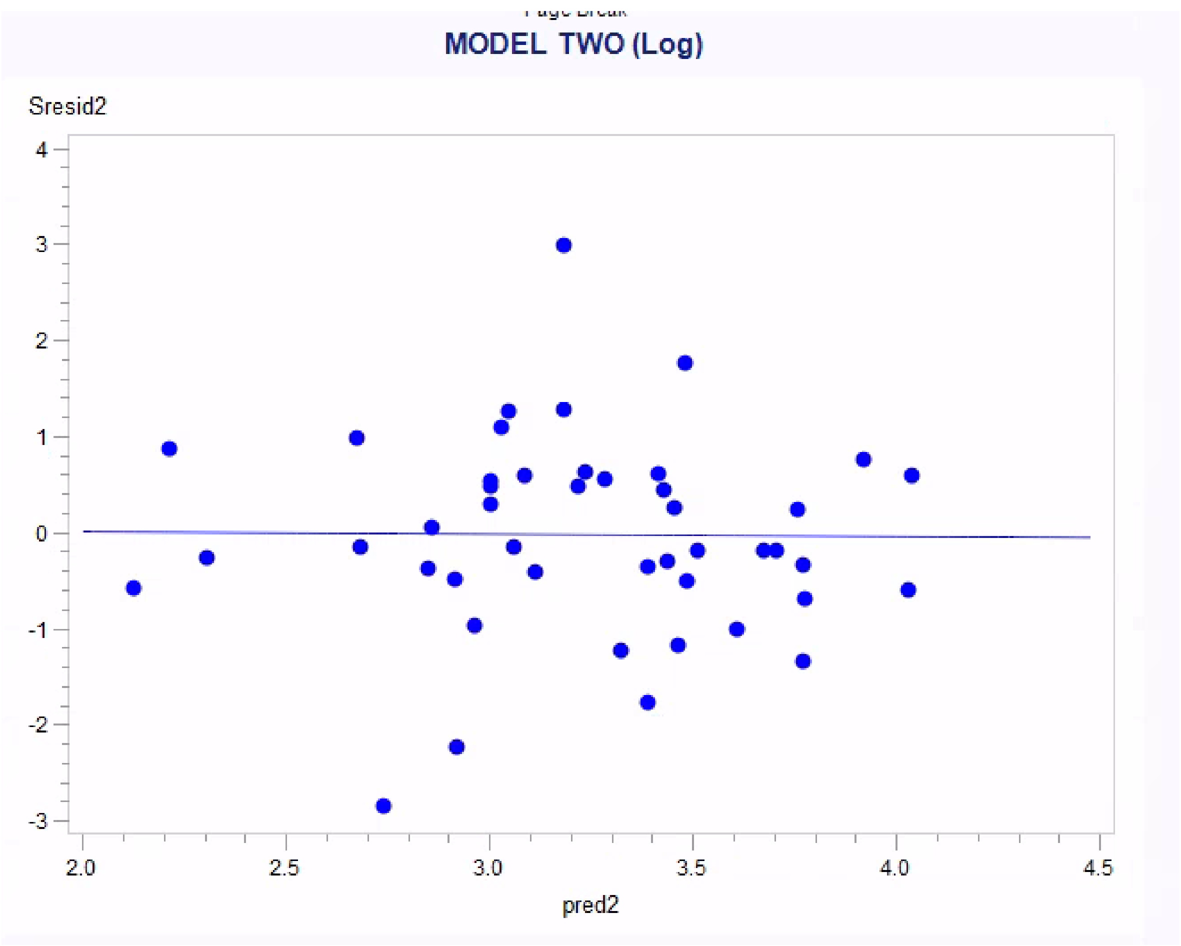
R-Square	Coeff Var	Root MSE	LY Mean
0.780215	9.204068	0.297281	3.229890

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		6.947569094	B	1.13084321	6.14	<.0001
LA1		-0.445334307		0.06480951	-6.87	<.0001
LPer2		-0.269607349		0.07937917	-3.40	0.0021
LPer3		0.609406759		0.25535655	2.39	0.0240
LPer4		0.112835099		0.08815263	1.28	0.2110
LPer5		-0.083974731		0.09538272	-0.88	0.3861
LPer6		-0.039056865		0.15590209	-0.25	0.8040
LPer7		-0.146857294		0.16705275	-0.88	0.3868
C1	1	-0.599255902	B	0.40177753	-1.49	0.1470
C1	2	-0.128028432	B	0.35599160	-0.36	0.7218
C1	3	-0.006970062	B	0.33841627	-0.02	0.9837
C1	4	0.000000000	B	.	.	.
C2		0.573799687		0.19660967	2.92	0.0069
C3		-0.361334095		0.17614338	-2.05	0.0497
C5		-0.195290582		0.20105724	-0.97	0.3397
C6		-0.160023269		0.13729253	-1.17	0.2536
C7		0.124358296		0.11865186	1.05	0.3036
C8		0.066528847		0.11065740	0.60	0.5525

Fig 3a. Parameter Estimate of the fitted model 2 (Log transformation)

MODEL FITTING
-Logarithm Data Transformation

- Plot 3a also shows that the multiple linear regression model adequately describes the systematic variation present in the response log Y (LY) across the entire range of its predicted values.
- **Model 2,** LY = 6.948 – 0.445LA1 – 0.269LPer2 + 0.609LPer3 + 0.113LPer4 – 0.084LPer5 – 0.039Lper6 – 0.147Lper7 – 0.599C11 – 0.128C12 – 0.007C13 +0.00C14 + 0.574C2 – 0.361C3 – 0.195C5 – 0.160C6 + 0.124C7 + 0.067C8
- **R-Square**= 0.7802; **Only 78.02%** of the variation is been explained by the explanatory variables. ***This model performs 20% better than model 1*** (whose variation is 58%). This shows that model performs better after the transformation.

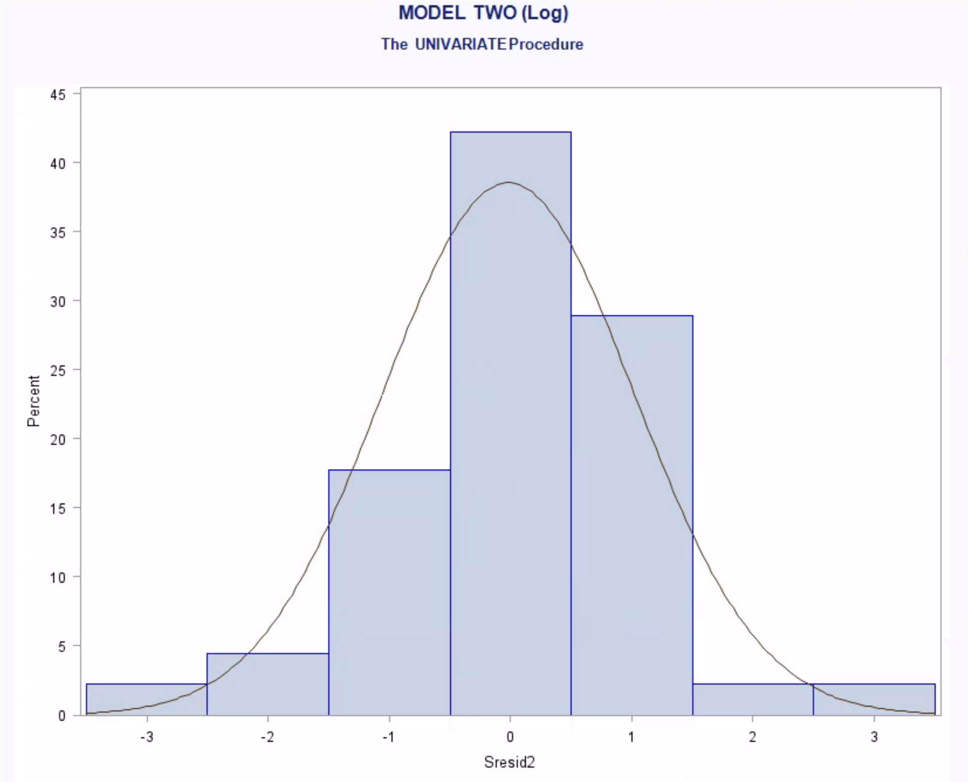


Plot 3a. A scattered plot of Studentised residual values versus predicted transformed values

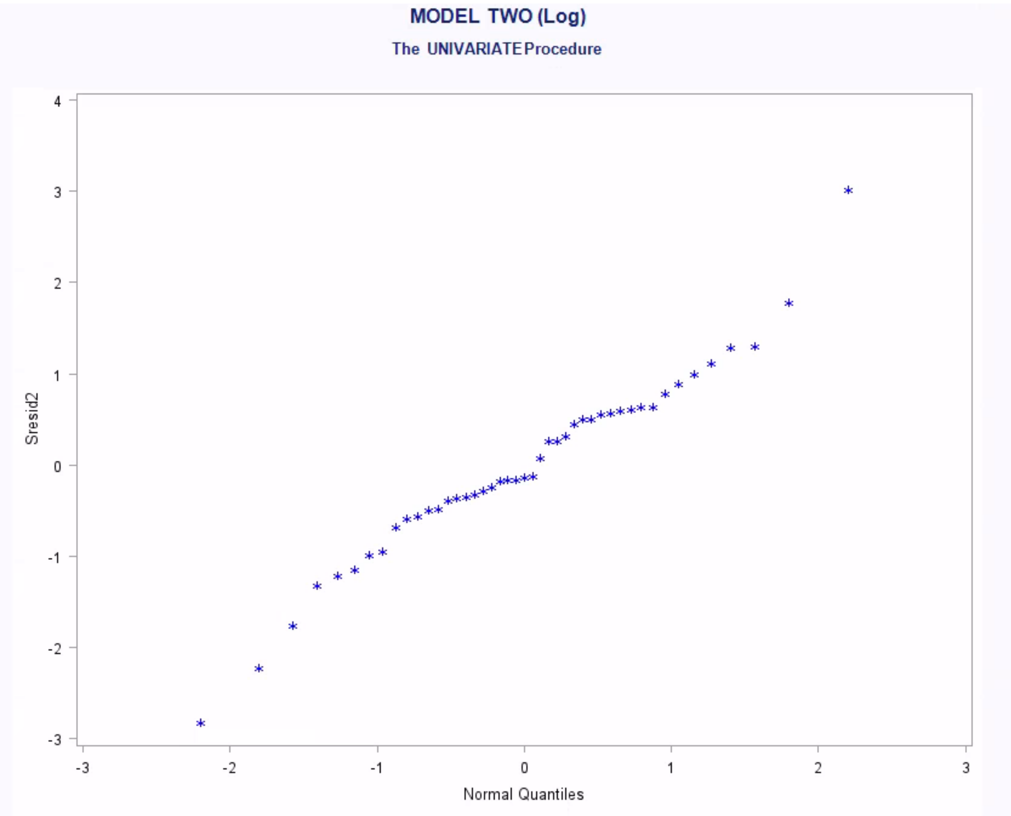
Statistical Assumptions 2– **Homoscedasticity**

Studentised residual values from MODEL-2 appears to be randomly scattered around the mean value of zero, with an approximately constant range across the entire range of its fitted predicted values.

This fulfils the assumption of constant variance/homoscedasticity



Plot 3b. A scattered plot of Studentised residual values versus predicted transformed values



Plot 3c. A normal probability plot (ggplot) of Studentised Residual of transformed variables

Systematic component and tenability Investigation of Log transformation *contd. –*

NEAR NOMALITY

- The histogram of the Studentised residuals of the transformed values in plot 3b shows that this Sresids are symmetrically distributed and unimodal as expected after transformation.
- Also, the Studentised residuals conform to an approximate straight line (this is place with an eye) of the unit slope passing near the origin on this normal. However, some irregularities were still observed which might be due to the relatively small sample size.
- This two plots have satisfied the assumption of near normality of random errors for model 2.

Model 1 and Model 2 Comparison

- The effect of the transformation carried out on the explanatory variables can be seen in the much-improved performance of the model. **The explanatory variables are better predictions of the response variable.**
- **Model 2 with 78.02% variation performs 20% better than model 1 (whose variation is 58%).** This shows that model performs better after the transformation.

Page Break

MODEL TWO (Log)

The GLM Procedure

Dependent Variable: LY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	8.78434146	0.54902134	6.21	<.0001
Error	28	2.47453310	0.08837618		
Corrected Total	44	11.25887456			

R-Square	Coeff Var	Root MSE	LY Mean
0.780215	9.204068	0.297281	3.229890

Fig 4a. A table of Model 2 Dependent Variable Summary

Model Selection
Possible Models

- All possible models = $\mathbf{2^K - 1} = 2^{16} - 1 =$ 65,536 - 1 = 65,535 where k is the total number of possible explanatory variables.

- **It is not practicable to generate 65,535 possible models,** because this number will roughly double for each additional potential explanatory variables.

Fig 4b1. A table of manual backward elimination _Final Step

MODEL Three (log)

The GLM Procedure

Class Level Information		
Class	Levels	Values
C1	4	1 2 3 4

Number of Observations Read	45
Number of Observations Used	45

Page Break

MODEL Three (log)

The GLM Procedure

Dependent Variable: LY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8.18244458	1.16892065	14.06	<.0001
Error	37	3.07642998	0.08314676		
Corrected Total	44	11.25887456			

R-Square	Coeff Var	Root MSE	LY Mean
0.726755	8.927603	0.288352	3.229890

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LA1	1	8.05838523	8.05838523	96.92	<.0001
LPer2	1	1.01807021	1.01807021	12.24	0.0012
LPer3	1	1.10468048	1.10468048	13.29	0.0008
C1	3	1.09142093	0.36380698	4.38	0.0098
C2	1	0.43291014	0.43291014	5.21	0.0283

Backward Elimination Procedure					
Elimination Step	Model DF	Model R-squared	Root MSE	Variable Eliminitated	P-value (Pr > F)
1	16	0.780215	0.297281	LPer6	0.804
2	15	0.779722	0.292438	C8	0.5663
3	14	0.777166	0.289186	C7	0.3715
4	13	0.771051	0.288361	LPer5	0.3737
5	12	0.765035	0.287524	C5	0.4566
6	11	0.760865	0.285635	C6	0.388
7	10	0.75532	0.284647	LPer7	0.5039
8	9	0.752036	0.282428	LPer4	0.2971
9	8	0.744097	0.2829	C3	0.127

Fig 4b2. A table of manual backward elimination _Removed Variables

FINAL MODEL SELECTION -Backward Elimination

- Backward elimination procedures has been applied manually because proc reg and proc glm does not have inbuilt selection model procedures.
- The manual backward elimination is done by dropping the model with the largest associated non-significant p-value.
- And this process step is repeated until we are left with only models with significant p-values at 95% confidence level, ($P_0 < 0.05$)
- Tables 4b2 shows, the eliminated variables and resulting model dropped. **9 variables** with p-value higher than 0.05 have been dropped as they are not effective predictors of the response variable LY.
- Only five(5) explanatory variables with significant p-value < 0.05 are remaining for the final model selected.
- The final model selected contain explanatory variables, **LA1, LPer2, LPer3, C1 and C2**

The GLM Procedure

Dependent Variable: LY

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8.18244458	1.16892065	14.06	<.0001
Error	37	3.07642998	0.08314676		
Corrected Total	44	11.25887456			

R-Square	Coeff Var	Root MSE	LY Mean
0.726755	8.927603	0.288352	3.229890

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LA1	1	8.05838523	8.05838523	96.92	<.0001
LPer2	1	1.01807021	1.01807021	12.24	0.0012
LPer3	1	1.10468048	1.10468048	13.29	0.0008
C1	3	1.09142093	0.36380698	4.38	0.0098
C2	1	0.43291014	0.43291014	5.21	0.0283

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		7.706981023	B	0.51087422	15.09	<.0001
LA1		-0.466919662		0.04742865	-9.84	<.0001
LPer2		-0.226301422		0.06467271	-3.50	0.0012
LPer3		0.328144288		0.09002625	3.64	0.0008
C1	1	-0.935674700	B	0.29292546	-3.19	0.0029
C1	2	-0.471212605	B	0.23458264	-2.01	0.0519
C1	3	-0.349491667	B	0.24173518	-1.45	0.1567
C1	4	0.000000000	B	.	.	.
C2		0.368270973		0.16139544	2.28	0.0283

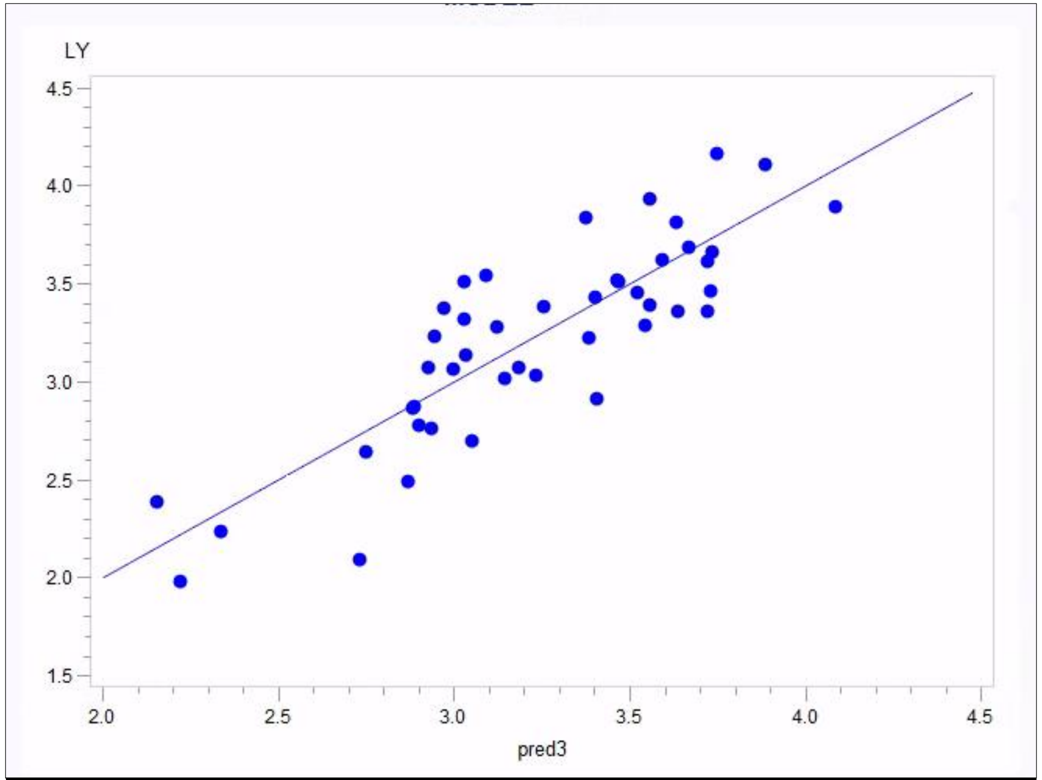
Fig 4c Parameter Estimate of the fitted selected model

Parameter Estimate for Final Selected Model

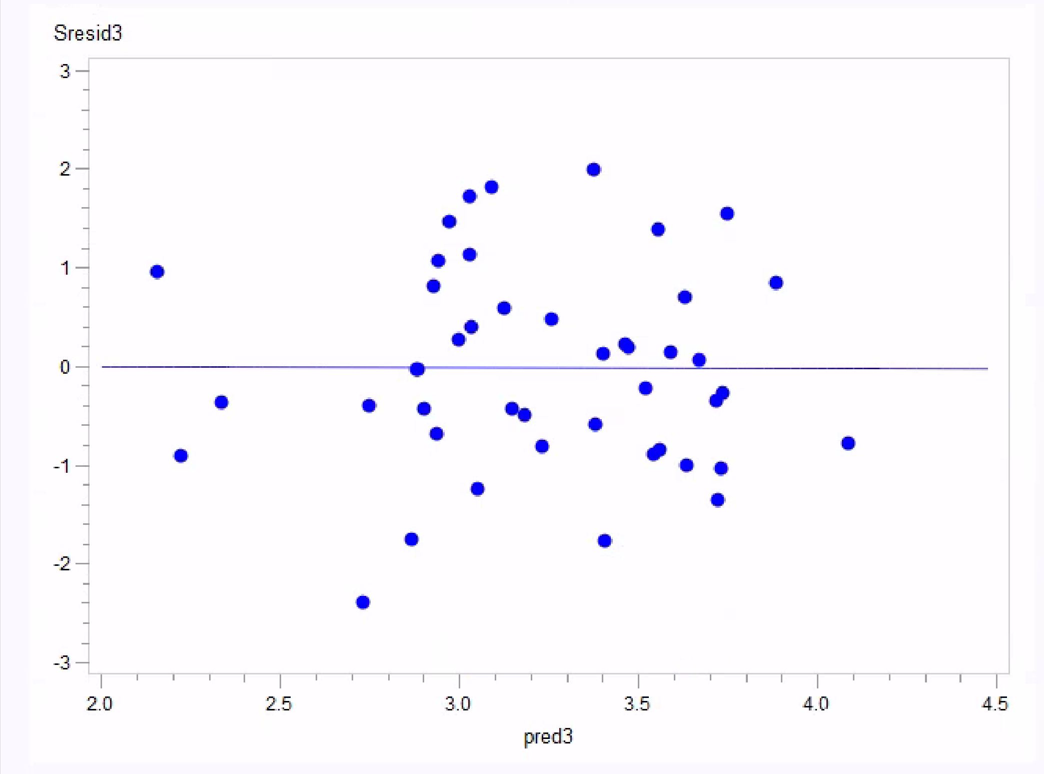
- The fitted regression model equation for the final model:

Model 3, LY =7.706 – 0.467LA1 – 0.226LPer2 + 0.328LPer3 – 0.935C11 – 0.471C12 – 0.349C13 + 0.00C14 + 0.368C2

- With an **R-Square of 0.726,** which explains that 73% **of variation is accounted for in the response variable.**
- With this **R-Square of 73%,** model 3 will predict the response variable LY well using the selected five(5) explanatory model.



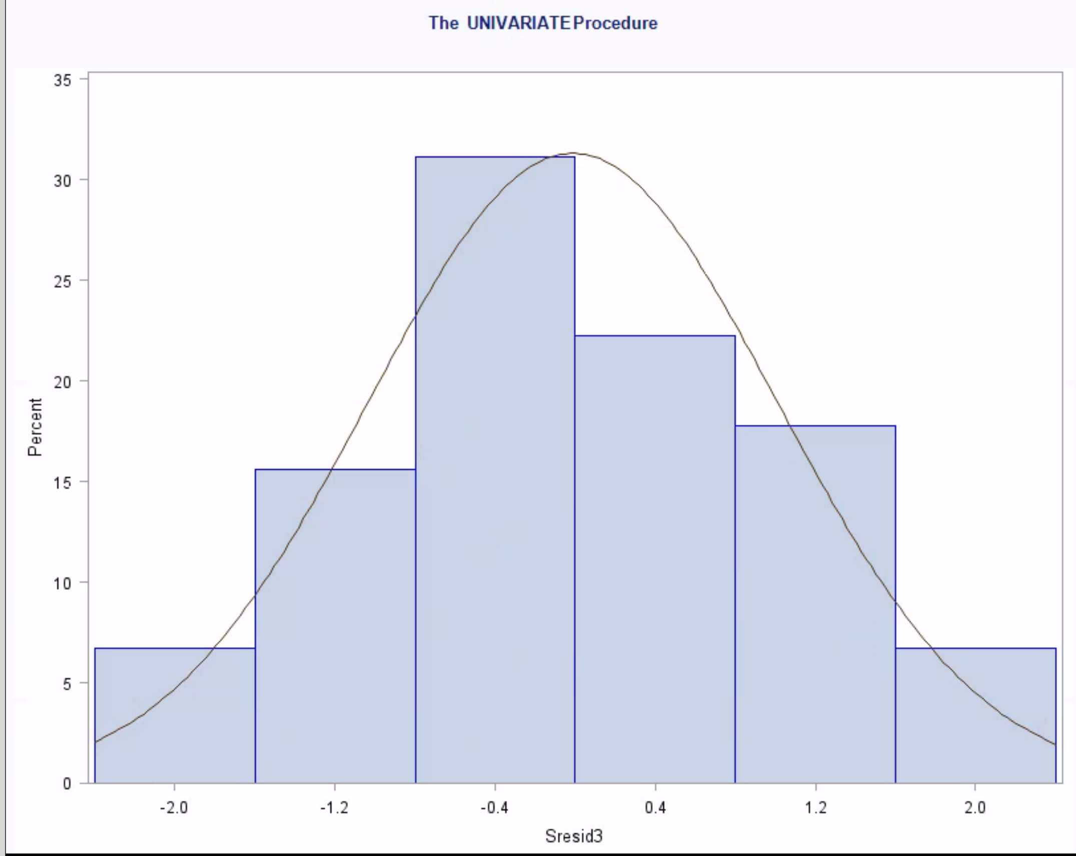
Plot 5a. A normal probability plot (ggplot) actual LY vs Predicted of final Model



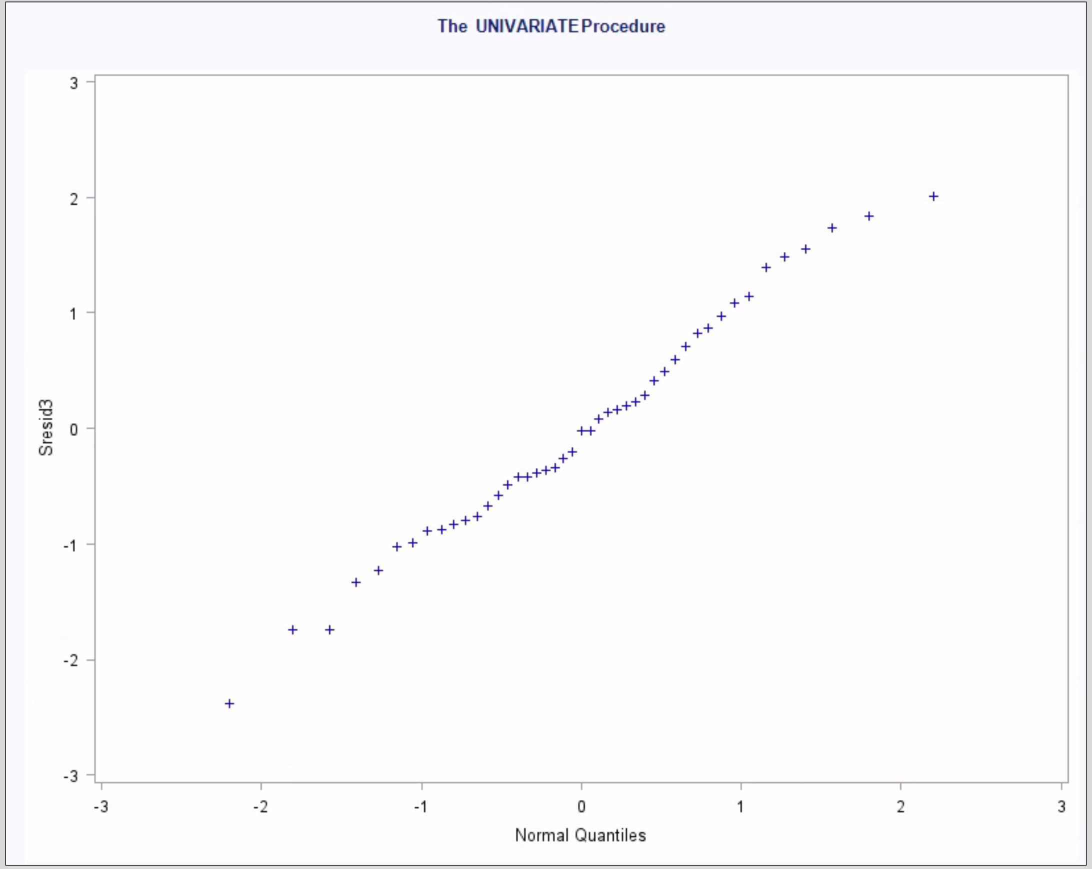
Plot 5b. A scatter plot of Studentised variable and predicted of final Model

Final Model Fit Investigation

- The normal probability scatter plot of actual LY versus predicted values of LY shows that the multiple linear regression model adequately describes the systematic variation present in the response LY across the entire range of predicted values.
- The Studentised residuals appear to be reasonably randomly scattered about a mean value of zero. The assumption of near normality is supported.



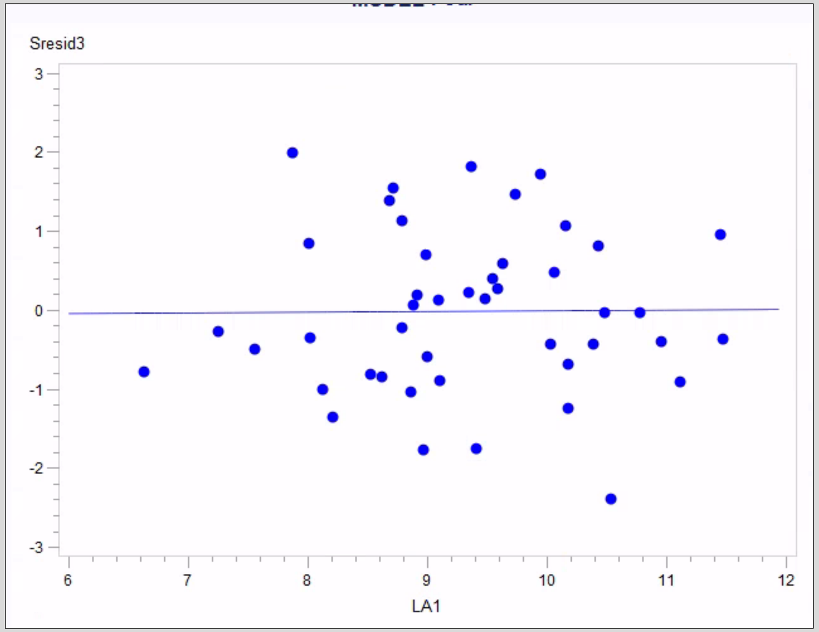
Plot 5c. A normal probability plot (ggplot) actual LY vs Predicted of final Model



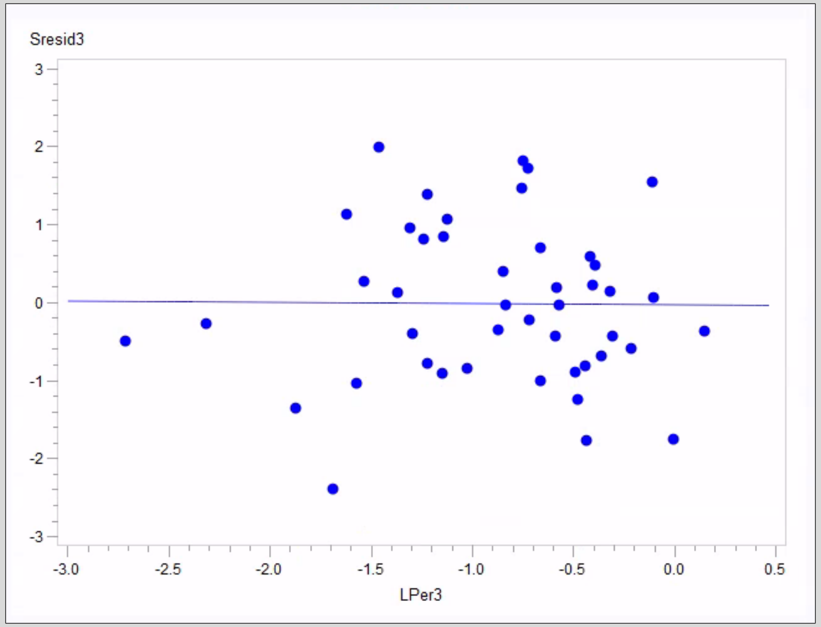
Plot 5d. A normal probability plot (ggplot) actual LY vs Predicted of final Model

Final Model Investigation
-Studentised residuals

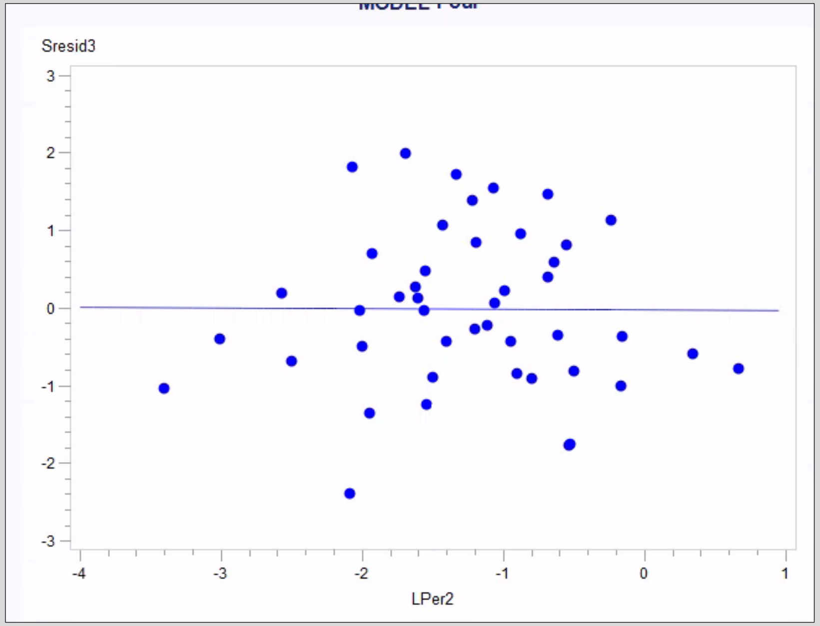
- The histogram of Studentised residuals of **MODEL-3** is also based on 45 observations. The Studentised residuals conform to a bell shape, indicating that it is normally distributed.
- The studentized residuals indeed conform to an approximately straight line (added by eye) of unit slope passing near the origin on this normal.
- The assumption of near-normality of random errors is satisfied.



Plot 5e. Plot of the studentised residuals against LA1



Plot 5f. Plot of the studentised residuals against LPer2



Plot 5g. Plot of the studentised residuals against LPer2

Final Model Investigation
-Explanatory Variables

- The studentized residuals appear to be reasonably randomly scattered about a mean value of zero, with variance across the entire range of values of the explanatory variables LA1, LPer2, and LPer3. The plots are therefore consistent with the adequacy of the assumption of a linear relationship between the response LY and the explanatory variables LA1, LPer2, and LPer3.

Influential Point Investigation
DFFITS

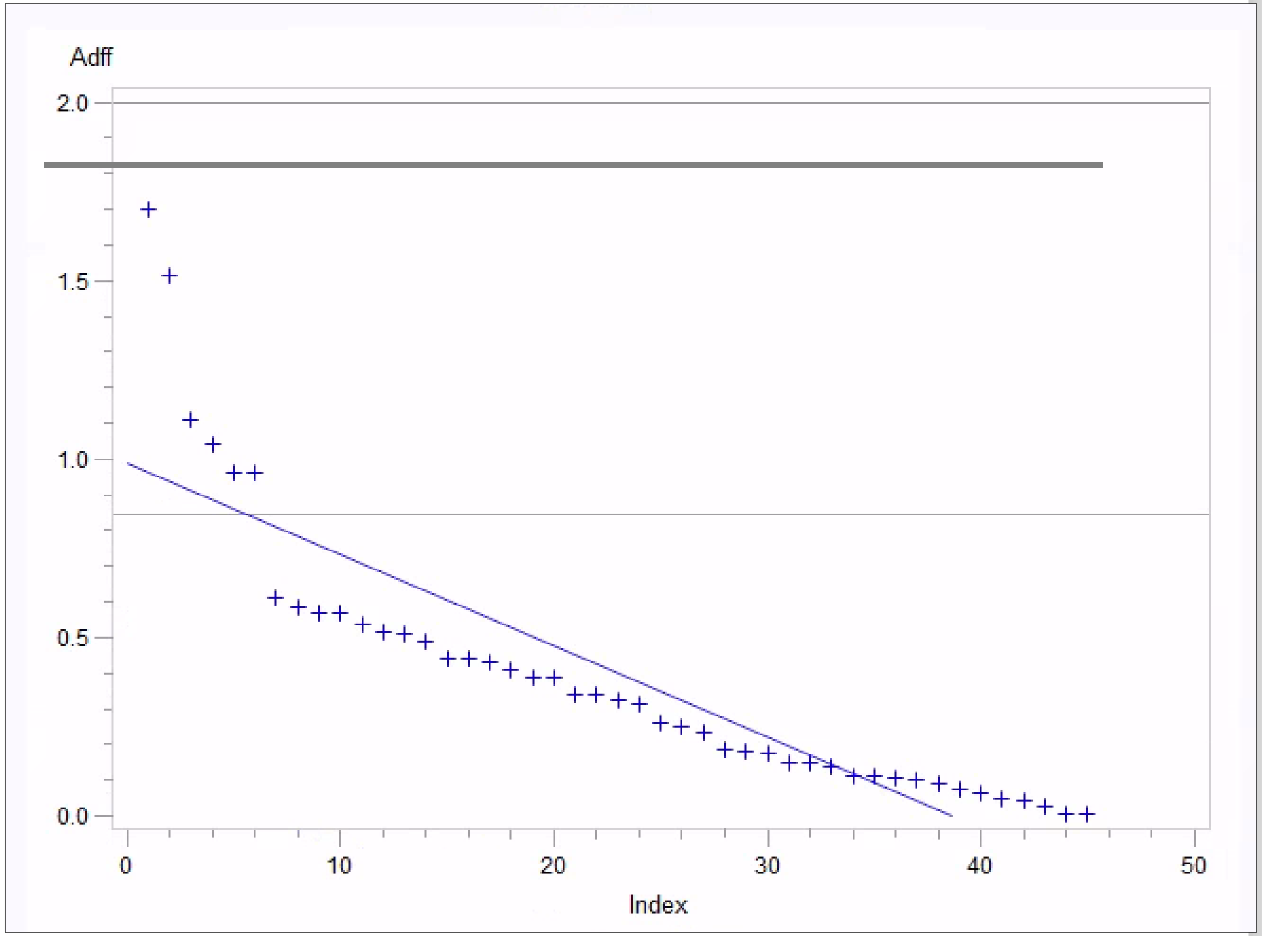
Fig. 6a Potential influential observations

Obs	ID	Pred	LY	Dff
1	31	2.86333	2.49848	-1.70128
2	3	3.37159	3.84028	1.51548
3	45	2.72747	2.09924	-1.11185
4	29	3.08588	3.55105	1.03958
5	7	3.55454	3.39913	-0.96006
6	40	2.92447	3.07988	0.96006

K	7
P	K+1 = 8
N	45
DFFIT	2 * sqrt(p/n) = 0.843

Calculation for DFFITS

- Even though all six of the potential influence points have absolute DFFITS values over the **size-related cutoff (0.843**), none of them is more than the overall cutoff of 2. **This implies that there is just weak evidence that the points may be influential.**



Plot 6a Absolute DFFITS values in descending order for PFARG data

Influential Point Investigation

-Additional Influence Statistics
Leverage H and COVARATIO c

- Observations 1 have above-average leverage, but a moderately large deleted residual. Its covariance ratio C is well within the acceptable limits and above 1, indicating that the inclusion of this observation will not reduce the precision of the fitted regression equation. On further investigation, this point causes little concern.
- Observations 2 & 4 have slightly above-average leverage coupled with large deleted residuals. The covRatio C are well within the acceptable limits although below 1, indicating that the inclusion of the observations will slightly reduce the precision of the fitted equation. On further investigation, this point causes little concern.
- Observation 3 has below-average leverage but with a large deleted residual. The covRation C is below the lower limit, indicating that the inclusion of the observation reduces the precision of the fitted equation. On further investigation, this point certainly causes concern.
- Observations 5 & 6 have above leverage cut-off but with a large deleted residual. Its covRatio C is well above 1, indicating that the inclusion of these observations will not reduce the precision of the fitted equation. On further investigation, this point causes no concern.
- **In conclusion, observation 3 appears unduly influential. It may be better to exclude it from the regression.**

Table 6b Potential influential observations

Obs	ID	Dff	H	Dresid	c
1	31	-1.70128	0.47324	-1.79491	1.19066
2	3	1.51548	0.34336	2.09576	0.75493
3	45	-1.11185	0.16002	-2.54741	0.39365
4	29	1.03958	0.22997	1.90230	0.75145
5	7	-0.96006	0.57585	-0.82395	2.52792
6	40	0.96006	0.57585	0.82395	2.52792

K	6
P	K+1 = 8
N	45
DFFIT	2 * sqrt(p/n) = 0.843
H	Average H=p/n, cut off is H > 3p/n = (0.178, 0.533)
C	Lies between 1 $\pm$ 3p/n = (0.467 and 1.533)

Investigation of **Multicollinearity** of Final Model -Correlation Matrix,

Dependent Variable: LY

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	7	8.18244	1.16892	14.06	<.0001
Error	37	3.07643	0.08315		
Corrected Total	44	11.25887			

Root MSE	0.28835	R-Square	0.7268
Dependent Mean	3.22989	Adj R-Sq	0.6751
Coeff Var	8.92760		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	7.70698	0.51087	15.09	<.0001	0
LA1	1	-0.46692	0.04743	-9.84	<.0001	1.42730
LPer2	1	-0.22630	0.06467	-3.50	0.0012	1.48062
LPer3	1	0.32814	0.09003	3.64	0.0008	1.58269
C11	1	-0.93567	0.29293	-3.19	0.0029	2.88953
C12	1	-0.47121	0.23458	-2.01	0.0519	7.26542
C13	1	-0.34949	0.24174	-1.45	0.1567	6.77816
C2	1	0.36827	0.16140	2.28	0.0283	1.39237

Table 6c. Variance Inflation Factor

Table 6d. Sample Correlation

The REG Procedure

Number of Observations Read	45
Number of Observations Used	45

Correlation

Variable	LA1	LPer2	LPer3	C11	C12	C13	C2	LY
LA1	1.0000	-0.1785	0.3261	-0.2595	0.0647	0.0535	0.0685	-0.7274
LPer2	-0.1785	1.0000	0.2761	-0.0521	-0.0806	0.0525	0.2019	0.0148
LPer3	0.3261	0.2761	1.0000	-0.2267	0.2516	-0.1062	-0.1477	-0.0476
C11	-0.2595	-0.0521	-0.2267	1.0000	-0.3126	-0.1796	0.0945	-0.0492
C12	0.0647	-0.0806	0.2516	-0.3126	1.0000	-0.7861	-0.1591	-0.0387
C13	0.0535	0.0525	-0.1062	-0.1796	-0.7861	1.0000	0.2376	0.0659
C2	0.0685	0.2019	-0.1477	0.0945	-0.1591	0.2376	1.0000	-0.0168
LY	-0.7274	0.0148	-0.0476	-0.0492	-0.0387	0.0659	-0.0168	1.0000

- Most of the correlations in the sample correlation table(12) are too low to be of much interest.
- No **VIF** comes close to the cut-off value of 10, which means that none of the prospective explanatory variables can be very well predicted by its fellow explanatory variables. There is currently minimal evidence to support the existence of significant collinearities among the potential explanatory variables for this regression issue.

Table 6e. Condition Index table

Collinearity Diagnostics										
		Condition	Proportion of Variation							
Number	Eigenvalue	Index	Intercept	LA1	LPer2	LPer3	C11	C12	C13	C2
1	5.36645	1.00000	0.00023904	0.00031122	0.00544	0.00536	0.00092374	0.00109	0.00117	0.00251
2	1.02998	2.28260	0.00001536	0.00004781	0.00041515	0.00498	0.16122	0.01825	0.01245	0.00005637
3	1.00523	2.31052	1.620098E-8	0.00000420	0.00076909	0.00007012	0.11944	0.00590	0.05249	0.00018421
4	0.29115	4.29322	0.00098067	0.00144	0.29114	0.17310	0.03147	0.00572	0.00454	0.04433
5	0.21309	5.01837	0.00005381	0.00079104	0.32008	0.51595	0.03269	3.486057E-7	0.01268	0.01201
6	0.05768	9.64530	0.01964	0.01817	0.23970	0.03674	0.02669	0.01070	0.02781	0.85196
7	0.03230	12.88938	0.01125	0.06842	0.00798	0.01292	0.41427	0.75535	0.72754	0.03879
8	0.00412	36.09026	0.96782	0.91081	0.13448	0.25088	0.21330	0.20298	0.16133	0.05015

Investigation of Multicollinearity of Final Model - **Condition Indices**

- Only Rows 8 have a condition index higher than 30 and almost exceeds the condition index in the preceding row by a factor of 3.
- In row 8, only the explanatory variable LA1 and possibly the intercept have high loadings.
- Weak evidence suggests that LA1 and the intercept may be correlated.
- This evidence is not particularly convincing, though, given the low VIFs found in table 6c and the fact that the condition index for row 8 above barely satisfies the requirement for further investigation. We have no real concerns regarding potential collinearities amongst the explanatory variables for this regression problem.

Prediction Statistics

Applying the Statistical model

We now have a final model (**Model 3**) that has been successfully fitted and has passed through the appropriate diagnostic analysis and influence investigation satisfactorily.

The model can be confidently adopted by the pension fund managers to be applied to predicting the pension fund administrative cost from the explanatory variables below.

- Number of active members
- Number of deferred pensioners per active member
- Number of pensioners per active member
- Staff-only fund type
- Combined scheme, same scales fund type
- Separate schemes fund type
- Combined scheme, different scales fund type

The form of confidence interval that best serves this purpose is the confidence interval for the fitted mean, this is because the confidence interval for predicted values of the response is always far wider than the corresponding confidence limits for the underlying mean of the response.

Predictions and confidence interval of fitted mean

- We have fitted the model and obtained the relevant predictions and 95% confidence interval for all observations in our pension fund dataset.
- The values shown have been transformed back to normal values and here we can clearly see the predicted administrative cost of the pension fund for all observations

Predictions and confidence interval of fitted mean

- We have fitted the model and obtained the relevant predictions and 95% confidence interval for all observations in our pension fund dataset.
- The values shown have been transformed back to normal values and here we can clearly see the predicted administrative cost of the pension fund for all observations

Table 7. Condition Index table

Page Break

Obs	EXA1	EXPer2	EXPer3	C11	C12	C13	C2	EXY	EXPred	EXLclM	EXUclM
1	1405	0.29893	0.09822	0	1	0	1	39.1459	41.7500	31.1755	55.9112
2	2600	0.18269	0.23077	1	0	0	1	46.5385	29.1249	20.6815	41.0155
3	1894	0.13411	0.06600	1	0	0	1	21.6473	24.0160	16.2057	35.5903
4	749	1.93992	0.29372	0	0	1	1	49.3992	59.3451	42.8982	82.0977
5	3666	0.14184	0.15276	0	0	1	1	28.9143	41.2328	32.8449	51.7627
6	5478	0.40215	0.35579	0	0	0	0	29.9379	34.9716	22.4474	54.4835
7	3000	0.30000	0.31667	0	0	1	1	61.3333	48.5481	39.7449	59.3013
8	3019	0.53528	0.41537	0	1	0	1	37.4296	41.0951	34.4505	49.0213
9	3362	0.84325	0.51160	0	1	0	1	28.8519	37.7571	31.2019	45.6895
10	8057	1.39518	0.80117	0	0	1	1	25.1955	29.3132	23.0409	37.2930
11	8891	0.22090	0.60994	0	1	0	1	26.9936	34.3957	29.2973	40.3813
12	5000	0.60000	0.64000	0	1	0	0	20.8000	25.2311	18.3433	34.7052
13	6518	0.78091	0.19638	0	1	0	1	27.7693	20.5989	16.0948	26.3635
14	6031	0.33974	0.88940	0	1	0	1	64.5001	42.3304	34.5838	51.8122
15	6536	0.32635	0.48531	0	1	0	1	31.8237	33.7266	29.3237	38.7907
16	6981	0.03295	0.20685	0	1	0	1	32.2303	41.5382	30.7824	56.0522
17	7750	0.58065	0.64516	0	1	0	1	18.4516	30.0173	25.7541	34.9863
18	7400	0.07568	0.55405	0	1	0	0	33.6486	32.0166	23.9182	42.8571
19	8800	0.20000	0.25227	0	0	1	1	31.0227	29.8816	25.1858	35.4529
20	7145	0.34416	0.89755	0	1	0	1	40.0280	39.1122	32.2259	47.4700
21	7956	0.14480	0.51144	0	1	0	1	45.7516	37.6223	31.3689	45.1222
22	5831	0.29395	0.29343	0	0	1	1	51.2777	34.8781	29.4428	41.3168
23	14545	0.19587	0.21451	0	1	0	1	21.5882	19.9334	16.7274	23.7539
24	22500	0.24444	0.73333	0	1	0	1	20.5778	23.1488	19.8502	26.9954
25	23303	0.20950	0.67150	0	0	1	1	29.6099	25.8754	21.2573	31.4968
26	11650	0.12532	0.46953	0	1	0	0	34.8498	21.8868	16.5387	28.9643
27	13100	0.17557	0.72519	0	0	1	1	37.7099	36.1417	28.8249	45.3159
28	12167	0.58815	0.98701	1	0	0	1	12.1641	17.5198	11.7213	26.1868
29	15064	0.52556	0.65593	0	1	0	1	26.7525	22.6338	19.5837	26.1590
30	11345	0.36730	0.66346	0	0	1	1	33.8475	31.7654	26.4095	38.2075
31	25678	0.23818	0.32358	0	0	1	1	25.4693	18.9041	15.8858	22.4959
32	13851	0.49939	0.42726	0	1	0	1	23.1752	20.6877	17.8940	23.9175
33	16820	0.50054	0.46772	0	1	0	1	29.4293	19.4534	16.8029	22.5221
34	20763	0.26263	0.48259	0	1	0	1	33.5693	20.6127	18.0946	23.4813
35	26133	0.08147	0.69292	0	1	0	0	15.9186	18.7996	13.9794	25.2818
36	33784	0.56965	0.28735	0	0	0	1	21.7559	18.6244	11.9546	29.0156
37	95639	0.84790	1.15442	0	1	0	1	9.4104	10.3161	7.9003	13.4708
38	47800	0.13180	0.43096	0	0	1	1	17.6778	17.7642	14.4544	21.8319
39	93152	0.41212	0.26855	0	0	1	1	10.9713	8.6060	6.4127	11.5493
40	37500	0.12267	0.18400	0	0	1	1	8.1600	15.2942	12.1067	19.3209
41	66850	0.44877	0.31473	0	1	0	1	7.2999	9.1930	7.0948	11.9118
42	57000	0.04912	0.27193	0	1	0	1	14.0877	15.5728	12.0700	20.0920
43	25998	0.21221	0.61539	0	1	0	1	14.9627	21.0929	18.2000	24.4456
44	32252	0.38683	0.55073	0	0	1	1	16.1540	18.1324	15.0965	21.7789
45	35284	0.20808	0.56156	0	1	0	1	17.7418	17.8274	15.2968	20.7767

Fund Manager Application

A participating pension fund manager can now predict the administrative cost as illustrated below:

Obs	EXA1	EXPer2	EXPer3	C11	C12	C13	C2	EXY	EXPred	EXLclM	EXUclM
11	8891	0.22090	0.60994	0	1	0	1	26.9936	34.3957	29.2973	40.3813

Example 1: *where ID =12,* EA1 = 8891, EPer2 =0.221, EPer3 = 0.61, C11 = 0, C12 = 1, C13 = 0, C2 = 1

Predicted EXY = £34.39

With 95% assurance, the administrative cost (EXY) lies between £29.297 & £40.38

Obs	EXA1	EXPer2	EXPer3	C11	C12	C13	C2	EXY	EXPred	EXLclM	EXUclM
28	12167	0.58815	0.98701	1	0	0	1	12.1641	17.5198	11.7213	26.1868

Example 2: *where ID =31,* EA1 = 12167, EPer2 =0.588, EPer3 = 0.987 C11 = 1, C12 = 0, C13 = 0, C2 = 1

Predicted EY = £29.12 against **Actual Cost of £12.16**

With 95% assurance, the administrative cost (EY) lies between £11.72 & £26.18

EXA1- No of active members
EXPer2 – No of deferred pensioner per active member
EXPer3 – Number of pensioners per active member
C11- Staff only fund type
C12 – Combined scheme, different scales fund type
C13 – Separate schemes
C2- Whether scheme is contracted out
EXY - Cost per Active Member
EXPred - Predicted Cost per active member
EXLclM =

Thank You

Investigation conducted by

DATT January Cohort GROUP 17