DATT ASSESSMENT II

TENZIN RABYANG – 30017118

SHEFFIELD HALLAM UNIVERSITY

01-12-2021

1a) Adding variables Y and Per2 to Per7 to the data set.

Obs	y	per2	per3	per4	per5	per6	per7
1	39.1459	0.29893	0.09822	0.20925	0.22420	0.011388	0.002847
2	46.5385	0.18269	0.23077	0.12077	0.10615	0.028846	0.005769
3	21.6473	0.13411	0.06600	0.18163	0.11563	0.014784	0.001056
4	49.3992	1.93992	0.29372	0.16956	0.10547	0.009346	0.005340
5	28.9143	0.14184	0.15276	0.11457	0.24386	0.023459	0.002182
6	29.9379	0.40215	0.35579	0.09766	0.07558	0.029938	0.006937

- Undertake a brief exploratory analysis of the variables Y, A1 and Per2 to Per7 by obtaining the sample mean and sample standard deviation for each of these variables

The MEANS Procedure

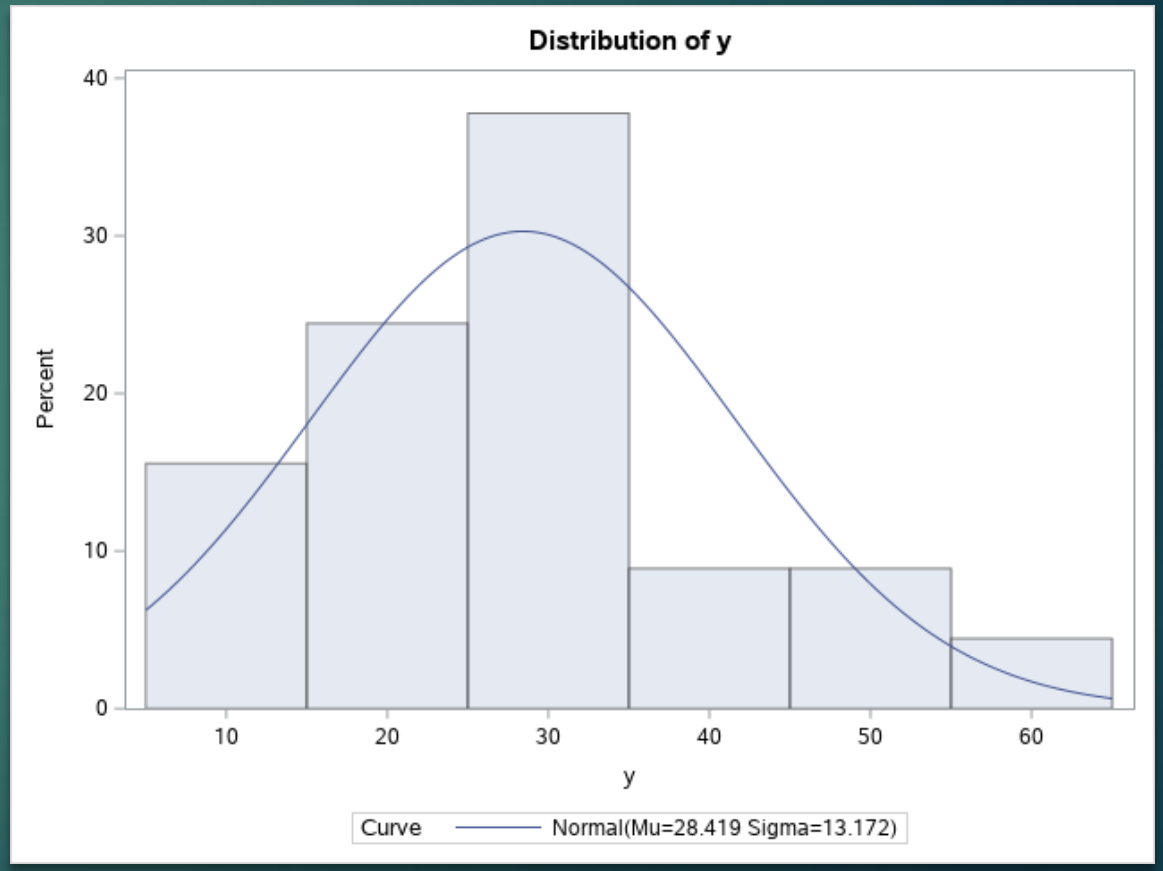
Variable	N Miss	Mean	Std Dev
y	0	28.4187463	13.1720807
per2	0	0.3783857	0.3517812
per3	0	0.4720161	0.2471502
per4	0	0.0980083	0.0496293
per5	0	0.1204053	0.0619437
per6	0	0.0482862	0.0227781
per7	0	0.0194682	0.0152031

NO MISSING VALUE.

Y: The mean of total cost per active member is at 28.4 pound where average spread of data is 13.17
per2: The mean of number of deferred pensioners per active member is 0.378 while average spread of data is 0.3517
per3: The average value of pensioners per active member is 0.472 where spread of data is at 0.247
per4: Starter in current year per active member has average value of 0.098 with standard deviation of 0.0496
per5: Number of leaver has average of 0.120 with 0.0619 spread.
per6: Number of new pensioner in current year has average of 0.0482 with 0.0227 data spread.
per7: Number of cessation in current year has average value of 0.0194 with 0.0152 data spread.

histogram for the response Total cost(y)

- It's Uni-model
- Symmetric/normal distribution
- Bell like shape
- Mean range 25-35
- Little positive skewed



1b) Investigate each of the factors C1 to C8 by obtaining a simple frequency distribution

- C1: Fund type combine scheme of same scales has highest active member with 57.78% while fund type combine scheme of different scales has the lowest with only 4.44%.
- C2: 88.89% of scheme is contracted out.
- C3: 84.44% of scheme is contributory
- C4: 97.78% of member can pay AVC's
- C5: 91.11% of admin base is in one location
- C6: 73.33% of admin calculation are not done in IT platform
- C7: 64.44% of special communication are sent to member at year end, while 35.56% aren't.
- C8: 55.56% are not communicate directly to member when rule changes but 44.44% are communicate directly.

- Since, we only have one frequency for member who can't pay the AVC (additional voluntary contributions), which means that frequency of "0" in C4 is only 1.
- Also AVC is done voluntarily, hence the data won't be consistent with time.

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

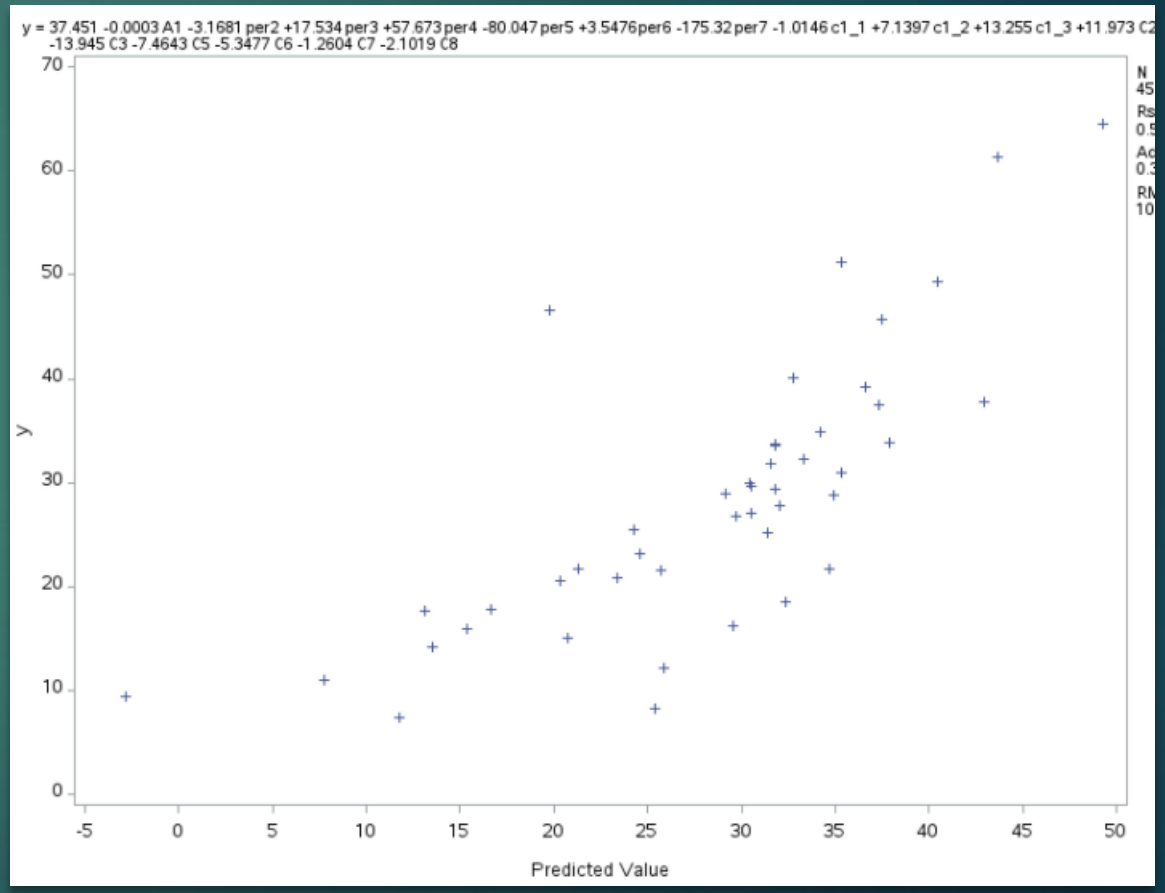
C6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	33	73.33	33	73.33
1	12	26.67	45	100.00

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. Fit of the systematic component of the model

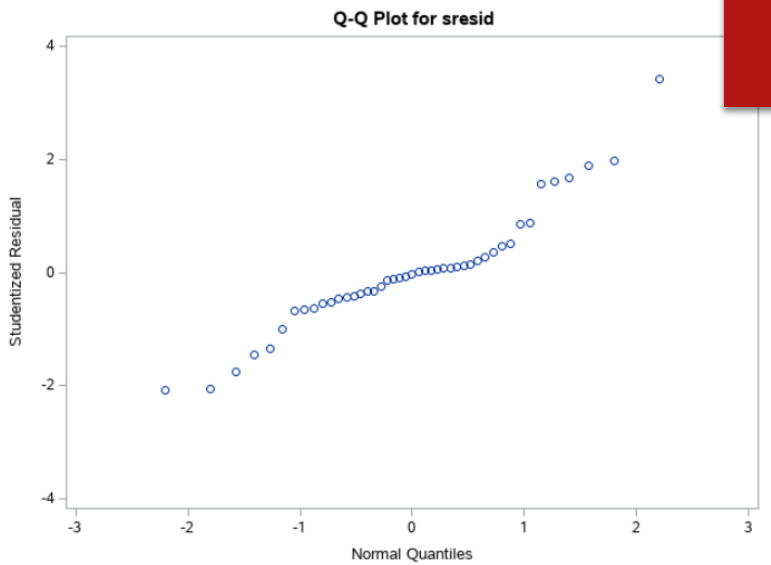
- Scatter plot between the response variable “y” and its predicted value.
- Intercept value is 37.451 and parameter value of all other exploratory variable is non-zero.
- Based on the diagram, I can assume that intercept value is non-zero which **reject** the null hypotheses which states that intercept value is zero.
- Hence the fit of the **SYSTEMATIC COMPONENT IS VALID.**

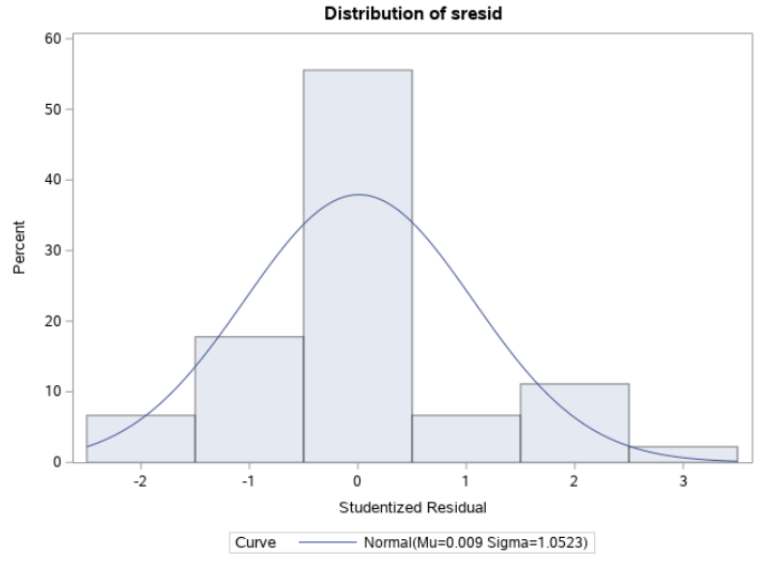


Investigate the tenability of the appropriate underlying statistical assumptions

1. Normality.

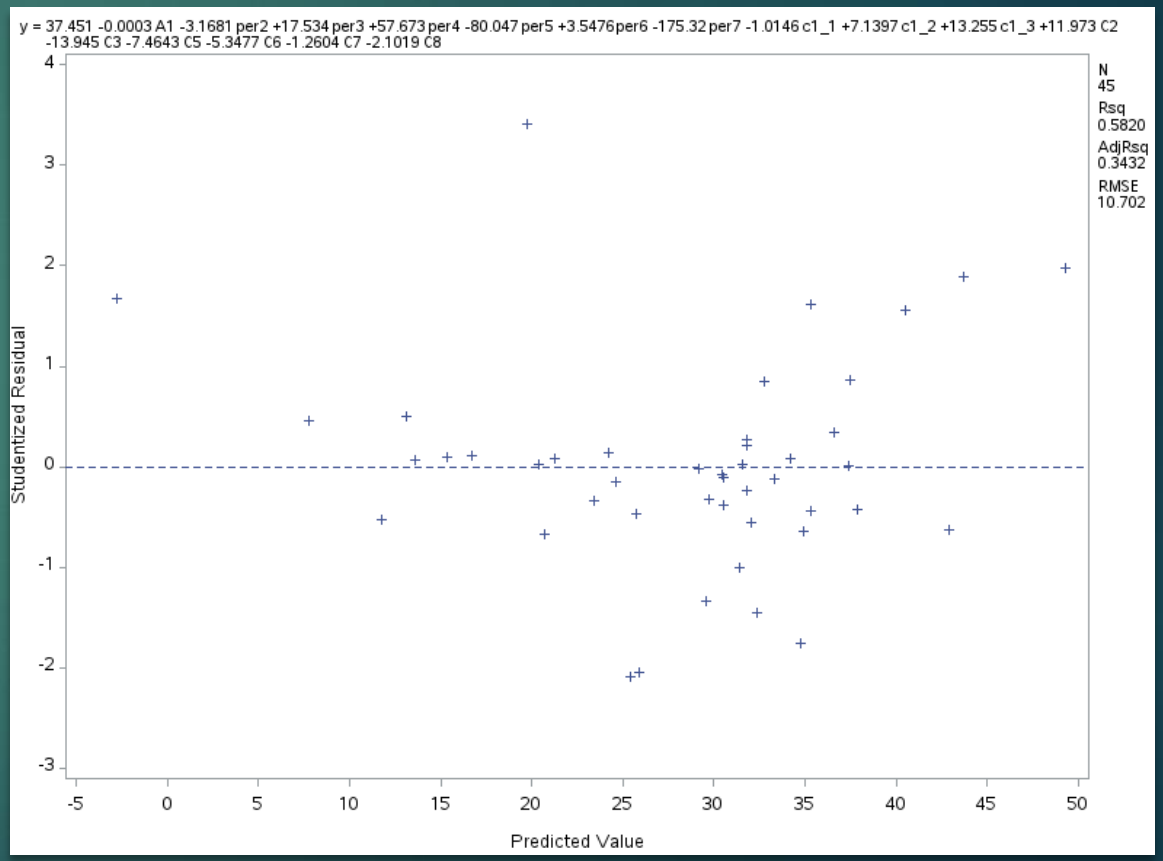
- Uni-Model
- Symmetric/ normal distribution
- Mean at 0
- Q-Q plot shows straight line of studentized residual
- Hence the normality Assumption is ACCEPTED.





2. Homoscedasticity.

- Not randomly scatter
- Shows pattern
- No constant variance
- Not Accepted

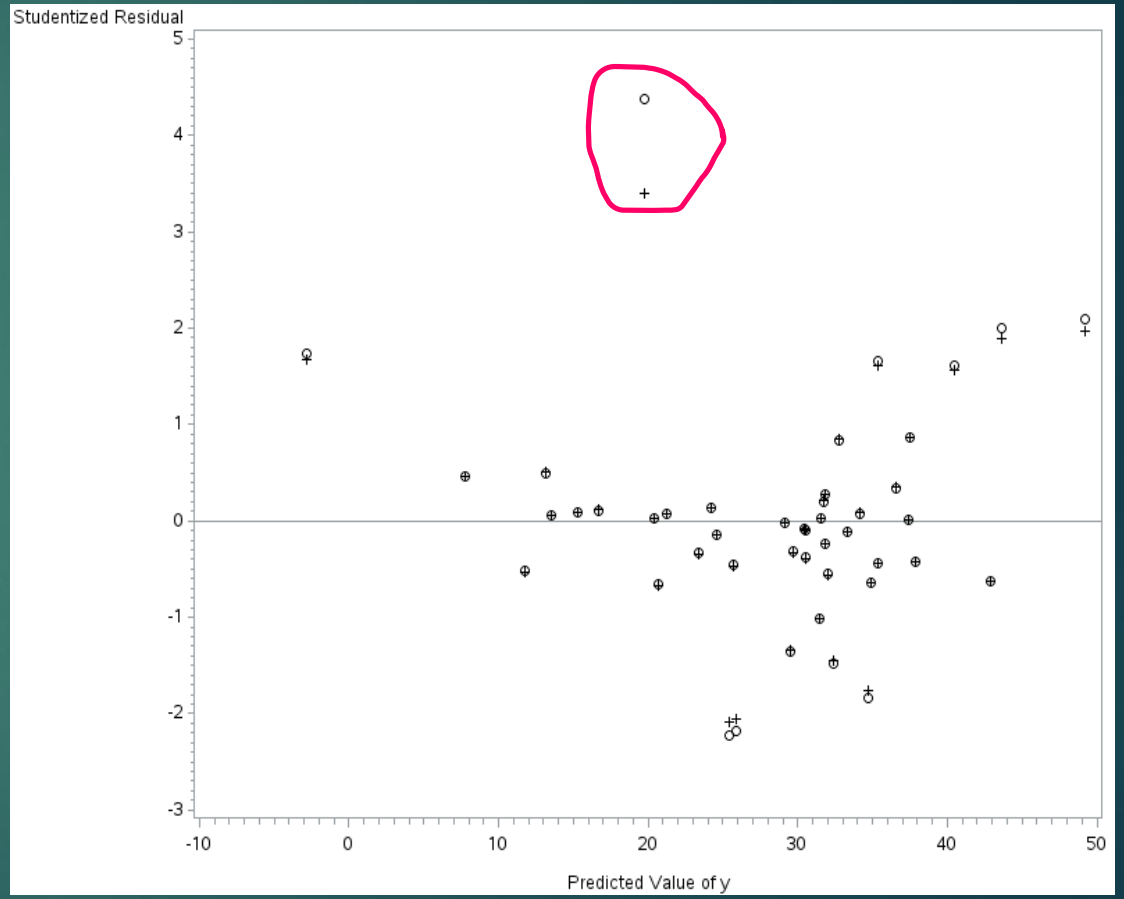


3. Mutual Independence

- Since the plots shows certain pattern, there might be relation between the exploratory variable.
- Not ACCEPTED.

4. Adequacy of the Systematic Component.

- Studentized and deleted residual are super impose with each other.
- Few are not.
- One residual goes beyond +4.
- It has quite long separation.
- Potential outlier
- Plot shows pattern
- No constant variance
- Hence, Not Accepted.



3a) Why not transform C1-C3 and C5-C8

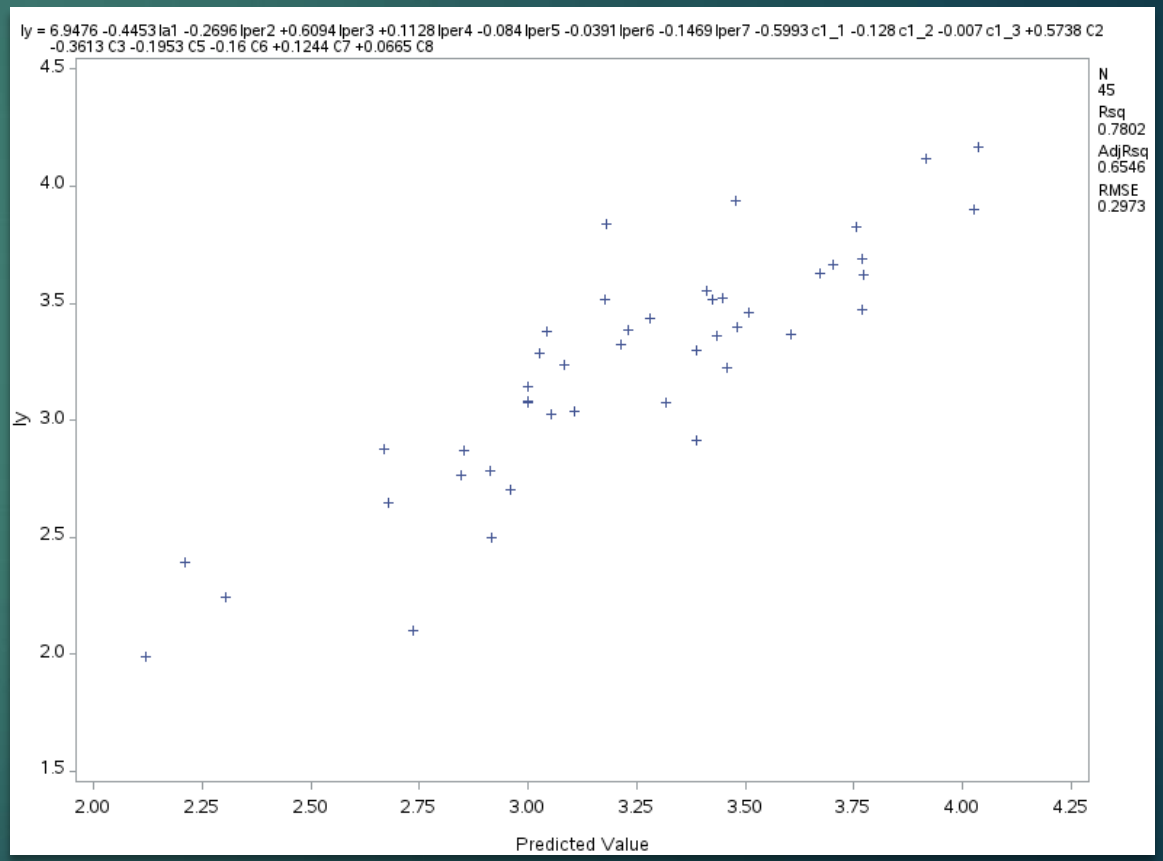
- IT is categorical data
- It contains only either value 1 or zero
- Log(0)= undefine/null
- Log(1) = 0
- Hence data will contain only “0” value.
- Data will get inconsistent and disrupted.

3b) Add the variables LY = log(Y), LA1 = log(A1), LPer2 = log(Per2),...,LPer7 = log(Per7) to your data set.

Obs	ly	la1	lper2	lper3	lper4	lper5	lper6	lper7
1	3.66730	7.2478	-1.20754	-2.32054	-1.56421	-1.49522	-4.47520	-5.86150
2	3.84028	7.8633	-1.69995	-1.46634	-2.11387	-2.24287	-3.54578	-5.15522
3	3.07488	7.5464	-2.00911	-2.71813	-1.70580	-2.15737	-4.21424	-6.85330
4	3.89993	6.6187	0.66265	-1.22511	-1.77455	-2.24929	-4.67283	-5.23244
5	3.36434	8.2069	-1.95303	-1.87892	-2.16660	-1.41115	-3.75251	-6.12741
6	3.39913	8.6085	-0.91092	-1.03342	-2.32623	-2.58263	-3.50863	-4.97091
7	4.11632	8.0064	-1.20397	-1.14991	-2.12026	-2.81341	-3.21888	-4.82831

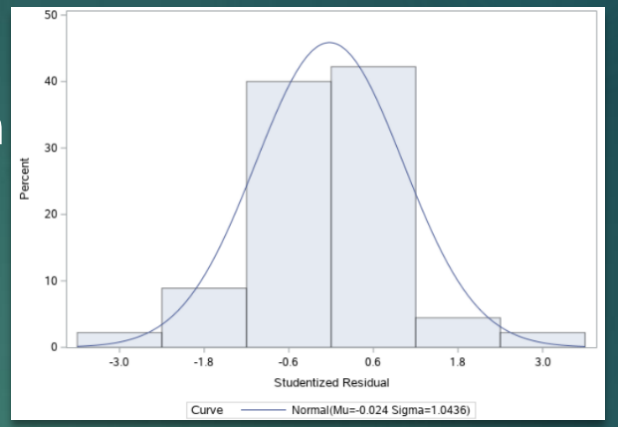
Investigate the fit of the systematic component of the model

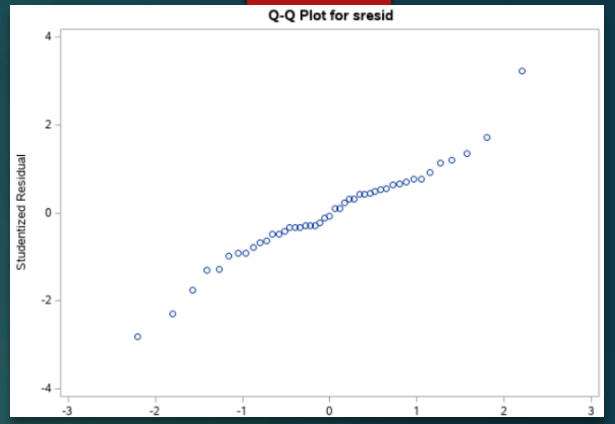
- Much better than earlier model.
- Shows better linear/straight line
- Intercept value of 6.94
- Hence Reject Null Hypotheses
- Fit of model is Valid



Normality

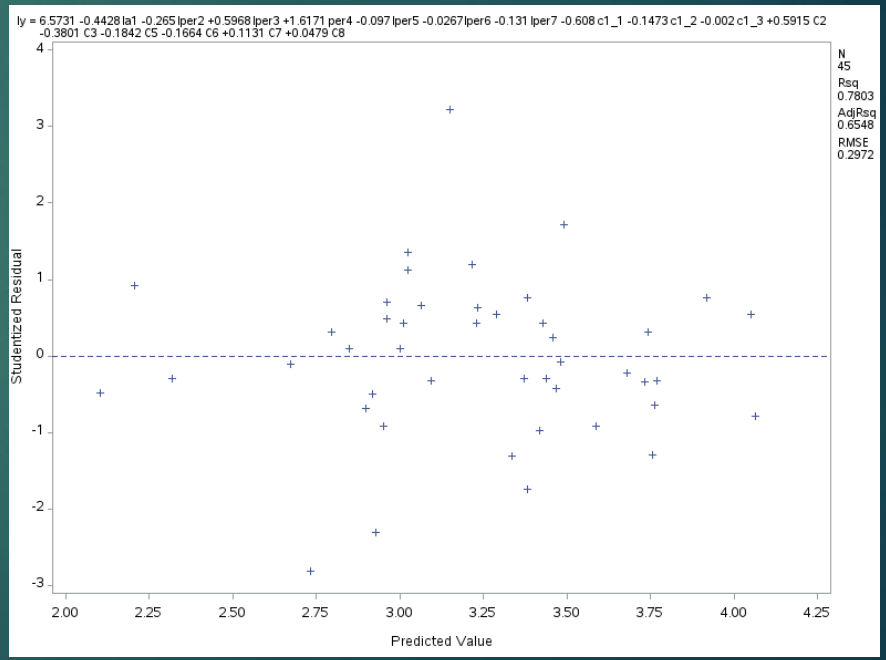
- Its is Uni-model with Symmetric distribution
- Mean around zero
- Bell like shape
- Can’t notice skewedness
- Q-Q plot shows straight line
- Accpeted





Homoscedasticity

- Much better randomly scatter
- No pattern detected
- But less variance
- Accepted

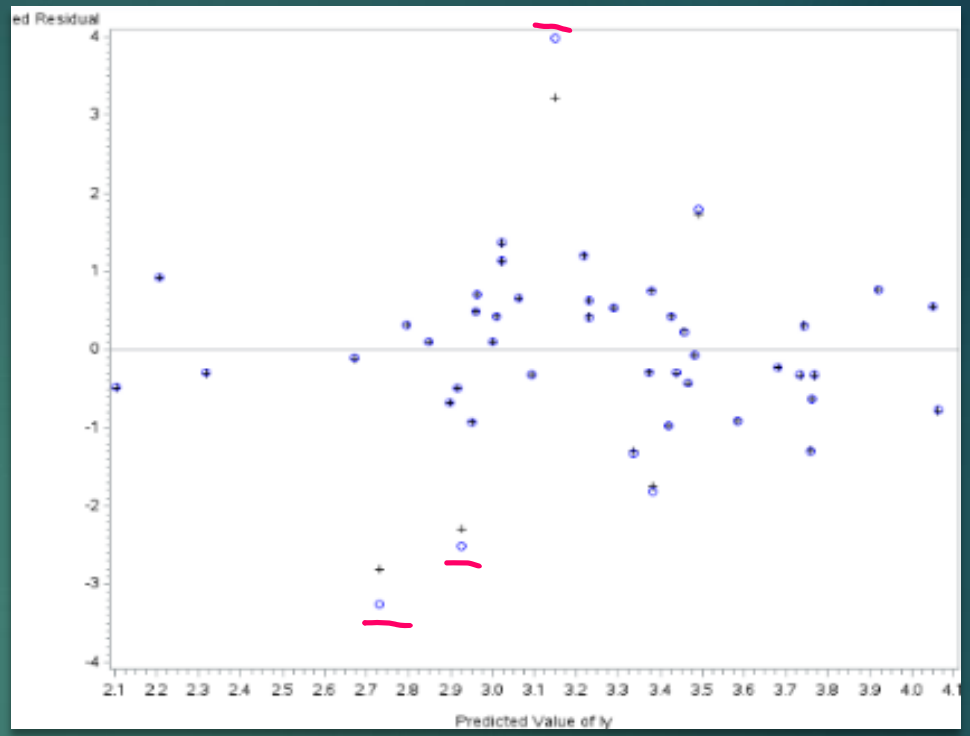


Mutual Independence

- Since no pattern recognize, less likely to have mutual relation between the observation.
- Hence, Accepted

Adequacy of Systematic Component

- Almost perfectly super impose
- Randomly scatter
- Few potential outlier detected
- Extreme value are within ±4.
- Since no recognition of pattern, adequacy is accepted.



Assumptions	M1	M2
Normality	Accepted	Accepted
Homoscedasticity	Not Accepted	Accepted
Mutual Independence	Not Accepted	Accepted
Adequacy of Systematic Component	Not Accepted	Accepted

4a) identify an overall "best" model for the prediction of log(Y)

NUMBER IN MODEL	R^2	MSE
1	0.5291	0.12328
2	0.5899	0.10995
3	0.6168	0.10523
4	0.6626	0.09497
5	0.6903	0.08942
6	0.7113	0.08553
7	0.7335	0.08109
8	0.7441	0.08003
9	0.7527	0.07956
10	0.7607	0.07926
11	0.7660	0.07983
12	0.7710	0.08055
13	0.7770	0.08098
14	0.7797	0.08267
15	0.7802	0.08533
16	0.7802	0.08838

R-square and MSE

Number in Model	R^2	Adjusted R^2
1	0.5291	0.5182
2	0.5899	0.5703
3	0.6168	0.5888
4	0.6626	0.6289
5	0.6903	0.6506
6	0.7113	0.6657
7	0.7335	0.6831
8	0.7441	0.6872
9	0.7527	0.6891
10	0.7607	0.6903
11	0.7660	0.6880
12	0.7710	0.6852
13	0.7770	0.6835
14	0.7797	0.6769
15	0.7802	0.6665
16	0.7802	0.6546

Adjusted R-square

Number in Model	R^2	Cp
1	0.5291	18.9851
2	0.5899	13.2510
3	0.6168	11.8176
4	0.6626	7.9841
5	0.6903	6.4591
6	0.7113	5.7772
7	0.7335	4.9480
8	0.7441	5.6013
9	0.7527	6.5093
10	0.7607	7.4912
11	0.7660	8.8096
12	0.7710	10.1679
13	0.7770	11.4059
14	0.7797	13.0634
15	0.7802	15.0004
16	0.7802	17.0000

Cp method

Now it is impracticable to consider all the model because,

1. Not all models are significant.
2. Not all models accept the assumptions of regression.
3. More model means more computation power is required.
4. Not all models produce accuracy results.

4b) Employ a backward elimination procedure, justifying your choice of final model.

Backward Elimination: Step 14

Variable C2 Removed: R-Square = 0.6626 and C(p) = 7.9841

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7.46010	1.86503	19.64	<.0001
Error	40	3.79877	0.09497		
Corrected Total	44	11.25887			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7.22187	0.46560	22.84861	240.59	<.0001
la1	-0.42340	0.04802	7.38188	77.73	<.0001
lper2	-0.14433	0.06195	0.51554	5.43	0.0249
lper3	0.21755	0.08741	0.58831	6.19	0.0171
c1_1	-0.48115	0.19331	0.58834	6.20	0.0171

Model	Step14	Spte13
R-square	0.6626	0.6903
MSE	0.09497	0.0894
Cp	7.9841	6.4591

Backward Elimination: Step 13

Variable C3 Removed: R-Square = 0.6903 and C(p) = 6.4591

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			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7.16004	0.45299	22.33896	249.83	<.0001
la1	-0.44446	0.04795	7.68404	85.94	<.0001
lper2	-0.18152	0.06332	0.73473	8.22	0.0067
lper3	0.26208	0.08811	0.79119	8.85	0.0050
c1_1	-0.52082	0.18878	0.68061	7.61	0.0088
C2	0.28542	0.15291	0.31153	3.48	0.0695

- R-square value not much high
- High Cp value
- High MSE value
- High error sum of square (residual)

- R-square value bit better
- Low Cp value
- Low MSE value
- Lower error sum of square
- Final “Best” Model

4b) Obtain, discuss, and interpret the parameter estimates for your final model.

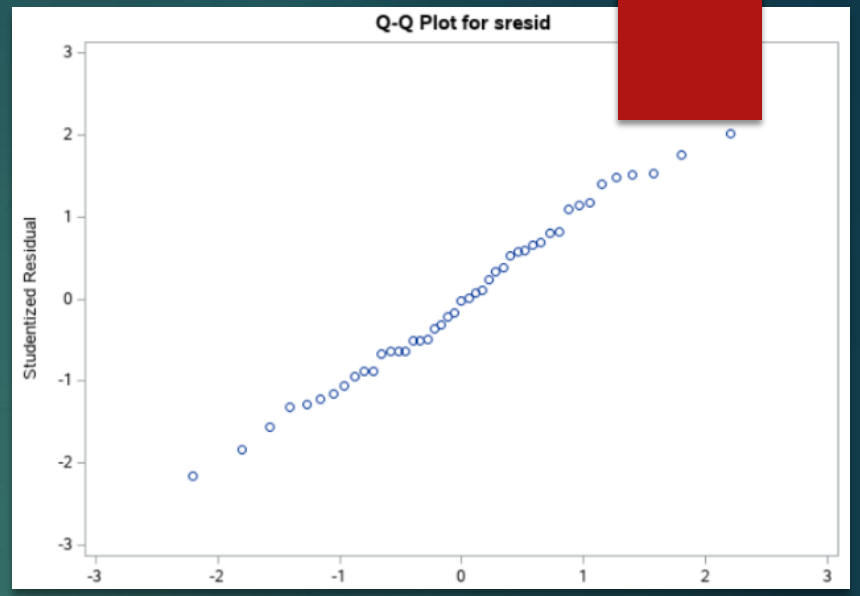
- -ve shows, negative relationship
- Intercept value is 7.16004
- Standard error shows spread of the mean and tell the accuracy.
- t-value can tell the similarity between the target variable and independent variable.
- More, t-value means more similarity or vice versa.
- All are significant except for C2.

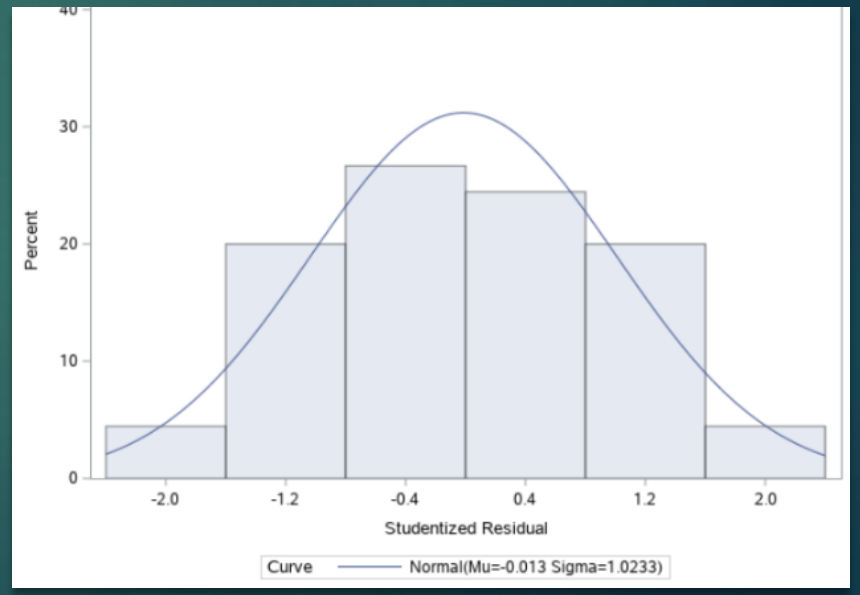
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.16004	0.45299	15.81	<.0001
la1	1	-0.44446	0.04795	-9.27	<.0001
lper2	1	-0.18152	0.06332	-2.87	0.0067
lper3	1	0.26208	0.08811	2.97	0.0050
c1_1	1	-0.52082	0.18878	-2.76	0.0088
C2	1	0.28542	0.15291	1.87	0.0695

5) Use appropriate plots to investigate the fit of your final model.

Normality

- Uni-model with symmetric distribution.
- Mean at almost zero
- Average data spread is 1.0233(sigma)
- Bell like shape
- No skewedness notice
- Q-Q plot shows straight line passing through the origin.
- Hence, accepted.





Homoscedasticity

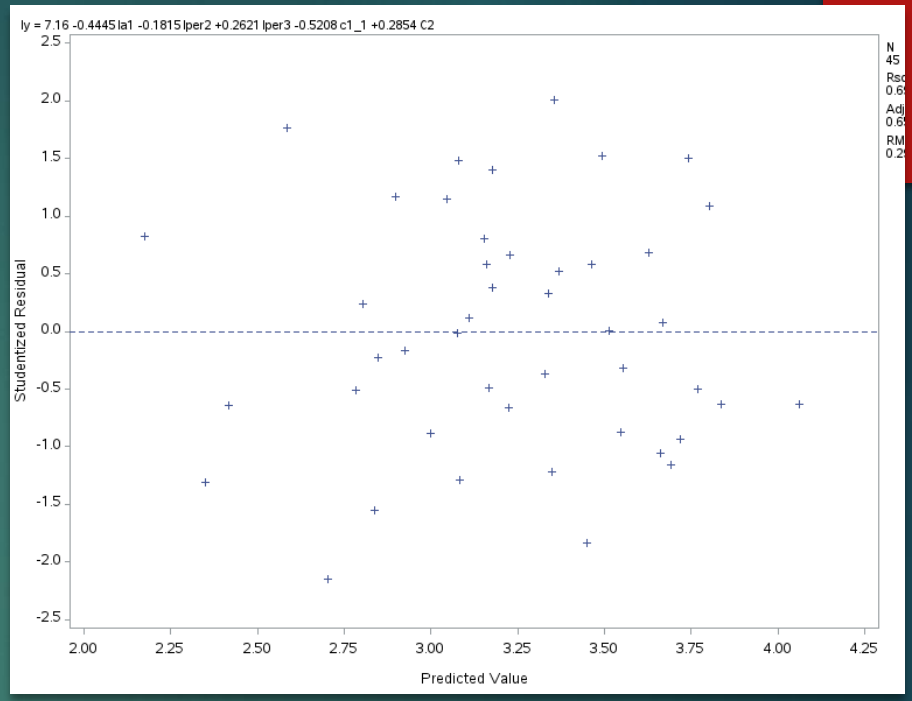
- Randomly scatter
- Much better constant variance
- No pattern detected
- R-square – 0.6903, Adj R – 0.6506
- RMSE- 0.299

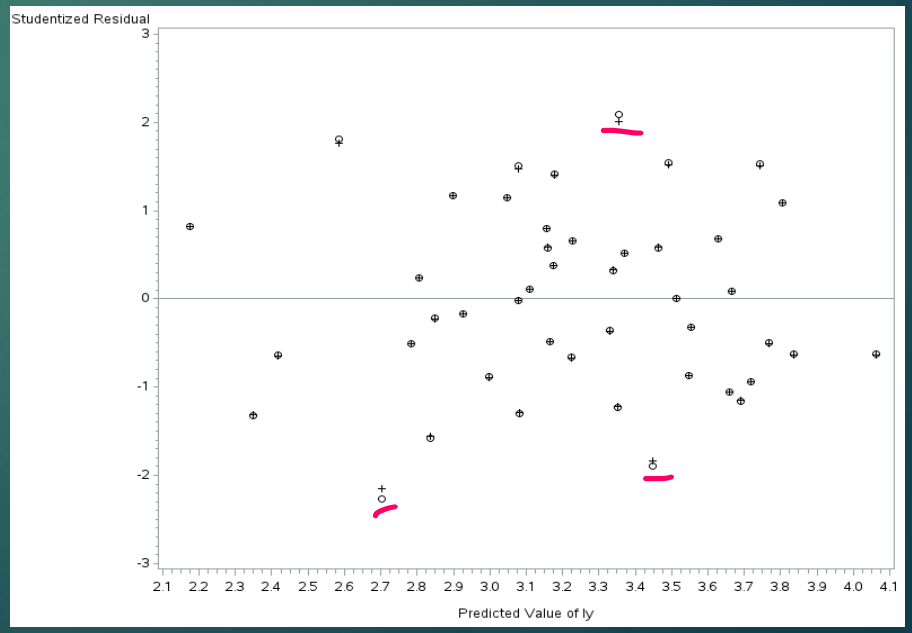
Mutual Independence

- Since no sign of pattern are recognize, its safe to say that to relationship are not present between the observation.

Adequacy of Systematic Component

- Almost perfectly imposed
- Few of them are not impose
- Residual are well within ±3

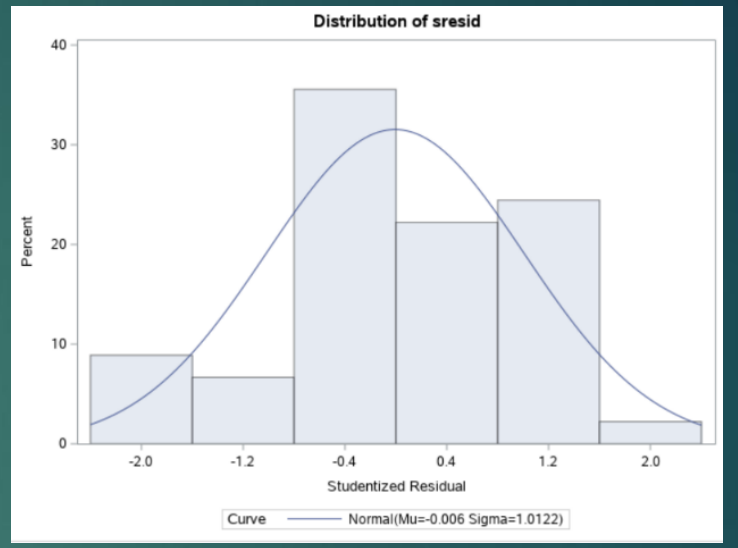


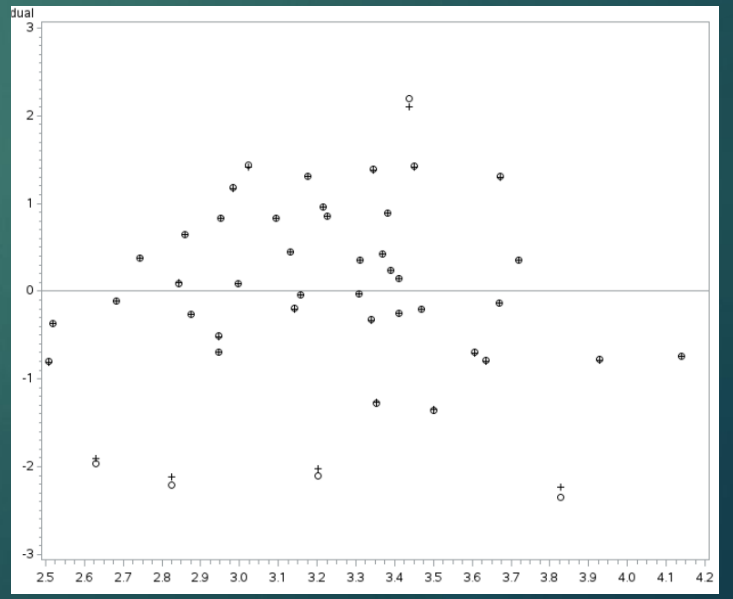


In addition to the overall fit, investigate the fit with respect to each of the explanatory variables

Ia1- First variable.

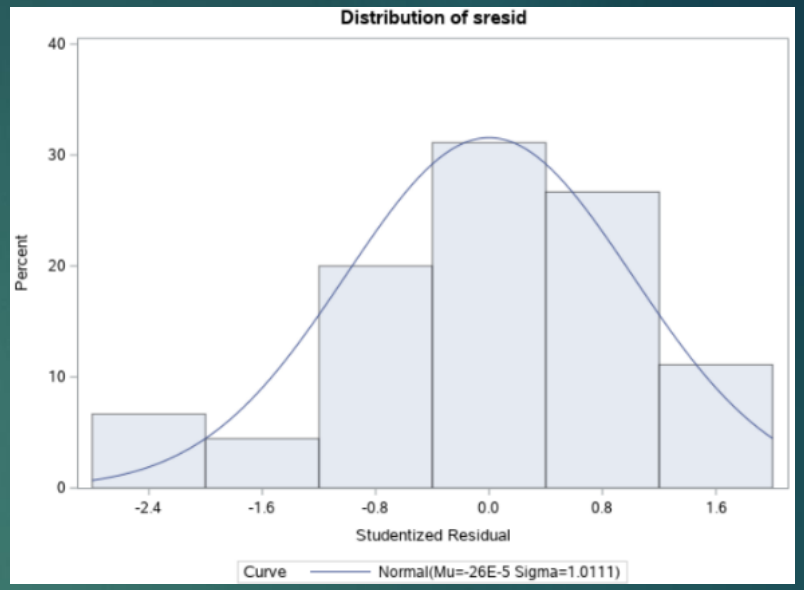
- Uni-Model with normal distribution
- Bell like shape
- Mean at 0.006
- Not skewedness notice
- Random scatter
- No pattern
- Better constant variance
- Few potential outlier

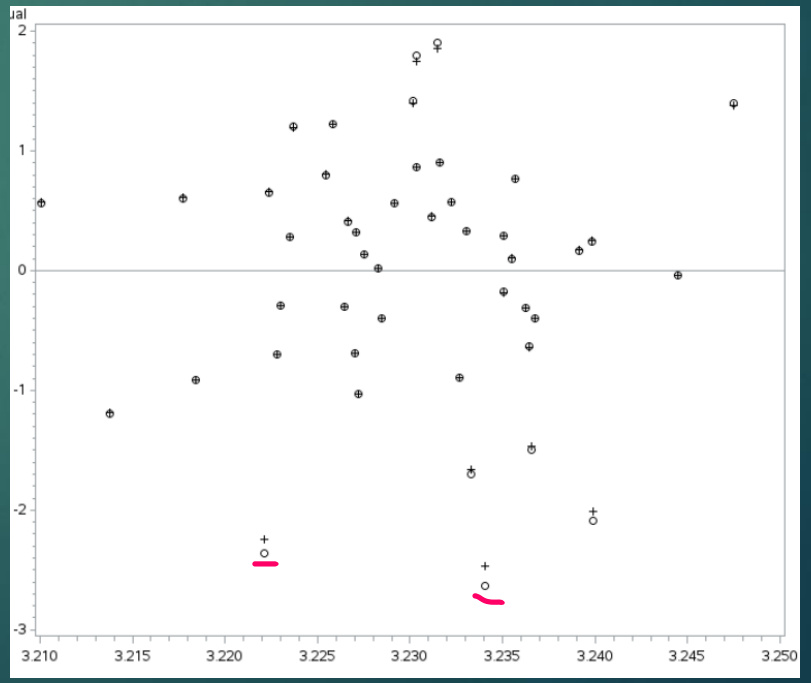




Iper2 – Second Variable

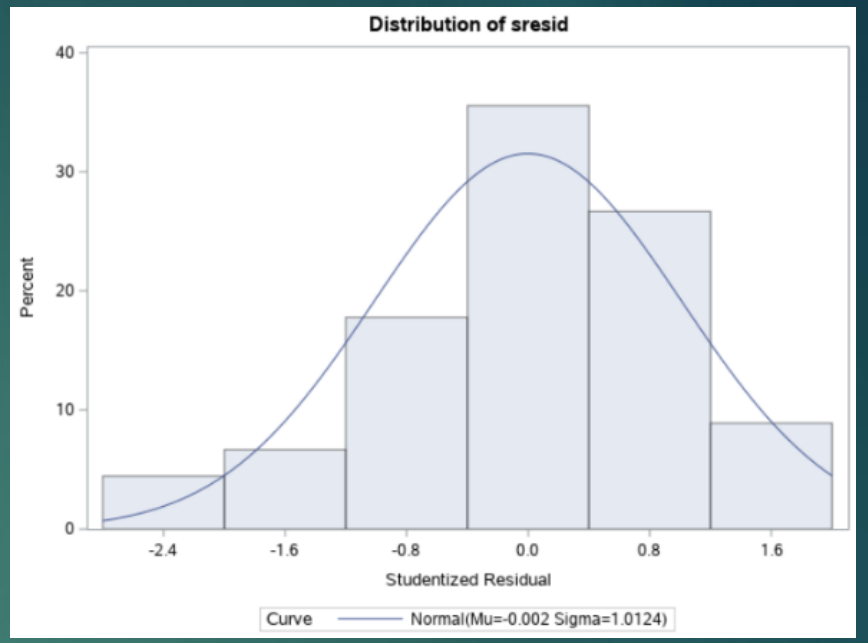
- Near normality distributed
- Bell like structure
- Little negative skewed
- Uni-Model.
- Random scatter
- Not much variance
- No pattern
- Few potential outlier

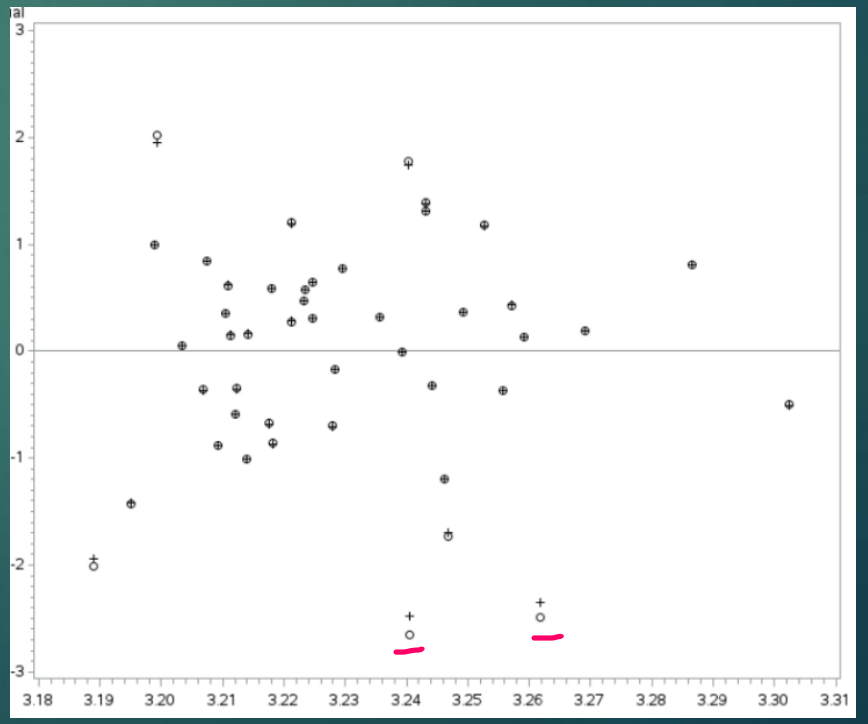




lper3 – Third variable

- Near normality
- Bit negative skewed
- Uni-model
- Shape like bell
- Mean is near zero
- Average data spread is 1.0124
- Randomly scatter
- No pattern
- But less constant variance
- Few potential outlier





6a) Using your final model, investigate and briefly discuss any issues relating to outliers or influential points

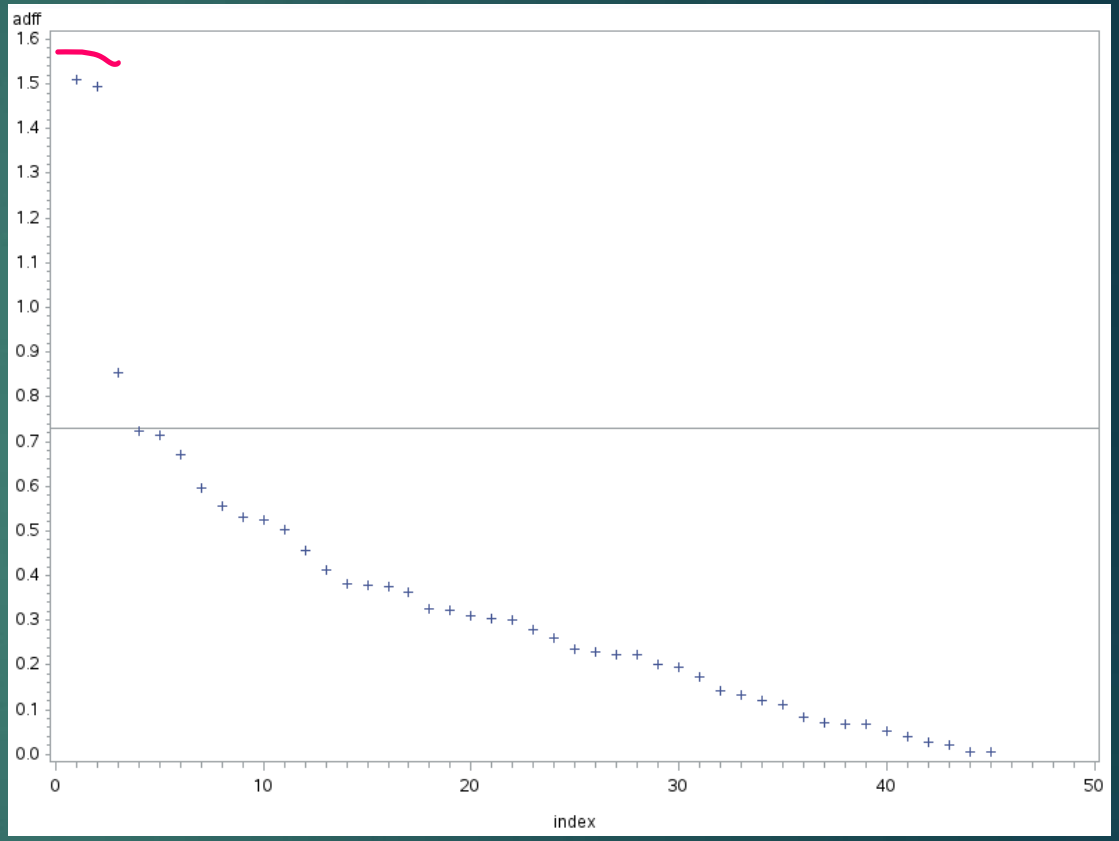
Outlier

- Based on scatter diagram, few potential outlier might present.

Influential points.

* Reference line(vref) is 0.730 (2*sqrt(p/n))

- Three are observed
- Lower one point can be ignore since it is near to vref
- But Two points near 1.5 might be potential outlier
- Not to be concern since all the points are less than 2.



DIFFITS

- Dff in the table is Difference of fits.
- Since both of observation has lower than ±2
- So, should not be concern about.

Obs	ly	la1	lper2	lper3	c1_1	C2	Dff
1	3.84028	7.86327	-1.69995	-1.46634	1	1	1.50861
2	2.49848	9.40648	-0.53078	-0.01307	1	1	-1.49533

Need more details

- First observation has slightly higher deleted residual of 2.09
- Cut-off value for H is 0.4 (3*(p/n))
- Range of C is 1±3(p/n) = 0.6 to 1.4

Obs	ID	ly	predicted	Dff	H	dresid	C
1	3	3.84028	3.35401	1.50861	0.34249	2.09029	0.92507
2	31	2.49848	2.83676	-1.49533	0.47089	-1.58506	1.50423

6b) Investigate and discuss any issues of multicollinearity in your final model.

Correlation

- Not much of correlation is found
- Maximum correlation has value of lper3 and la1 with 0.3261.
- Since most of the value is very low, this means weak correlation.
- Don't use ly (target variable)

Correlation						
Variable	la1	lper2	lper3	c1_1	C2	ly
la1	1.0000	-0.1785	0.3261	-0.2595	0.0685	-0.7274
lper2	-0.1785	1.0000	0.2761	-0.0521	0.2019	0.0148
lper3	0.3261	0.2761	1.0000	-0.2267	-0.1477	-0.0476
c1_1	-0.2595	-0.0521	-0.2267	1.0000	0.0945	-0.0492
C2	0.0685	0.2019	-0.1477	0.0945	1.0000	-0.0168
ly	-0.7274	0.0148	-0.0476	-0.0492	-0.0168	1.0000

Variance Inflation Factors

- Usually, VIF value of more than 10 is considered as high correlation.
- But all of the variables has less than 2 VIF value.
- This also shows weak correlation between exploratory variable.
- However, we can still use advance method call condition indices.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	7.16004	0.45299	15.81	<.0001	0
la1	1	-0.44446	0.04795	-9.27	<.0001	1.35630
lper2	1	-0.18152	0.06332	-2.87	0.0067	1.31999
lper3	1	0.26208	0.08811	2.97	0.0050	1.40957
c1_1	1	-0.52082	0.18878	-2.76	0.0088	1.11592
C2	1	0.28542	0.15291	1.87	0.0695	1.16222

Condition Indices

- Better method to find correlation
- Two rule of thumb,
  1. Condition index value is more roughly 3-4 times than preceding.
  2. Should be more than 30.
- 6th observation checks both rule.
- Now look for corresponding value of variables.
- Only la1 has high value (OK)
- No other variables has high value as la1.
- Might have very little relation between **lper3** and **la1**.(no concern)
- It shows weak correlation.

Collinearity Diagnostics								
			Proportion of Variation					
Number	Eigenvalue	Condition Index	Intercept	la1	lper2	lper3	c1_1	C2
1	4.52344	1.00000	0.00045195	0.00045105	0.00866	0.00879	0.00517	0.00418
2	0.91290	2.22599	0.00012709	0.00022337	0.00171	0.00063810	0.87210	0.00059494
3	0.27717	4.03982	0.00209	0.00230	0.38519	0.12220	0.01223	0.08211
4	0.22337	4.50012	0.00060370	0.00202	0.26132	0.64765	0.05691	0.00469
5	0.05810	8.82340	0.03402	0.02797	0.27402	0.04129	0.00914	0.90023
6	0.00502	30.02076	0.96270	0.96704	0.06911	0.17943	0.04445	0.00820

7) Determine the form of confidence interval that best serves this purpose - the confidence interval for the fitted mean, or the confidence interval for a predicted observation.

- The range of confidence intervals of predict is more than that of mean.
- It means that there is more confidence in predict than mean.
- This is mainly because, confident interval of mean considered whole observation making it harder to make decision.
- But in predict confident interval, it only takes individual corresponding value rather than whole observation.
- Which make model more confident while making decision
- Hence the pension fund manager should use prediction confident interval rather than mean confident interval

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	3.67	3.8351	0.1372	3.5576	4.1127	3.1697	4.5006	-0.1678
2	3.84	3.3540	0.1750	3.0000	3.7080	2.6532	4.0548	0.4863
3	3.07	3.2229	0.2002	2.8179	3.6279	2.4950	3.9508	-0.1480
4	3.90	4.0623	0.1541	3.7506	4.3741	3.3819	4.7428	-0.1624
5	3.36	3.6599	0.1020	3.4535	3.8663	3.0208	4.2990	-0.2956
6	3.40	3.2284	0.1490	2.9271	3.5297	2.5527	3.9041	0.1707
7	4.12	3.8041	0.0802	3.6419	3.9663	3.1779	4.4303	0.3122
8	3.62	3.7673	0.0815	3.6025	3.9322	3.1404	4.3942	-0.1449
9	3.36	3.6916	0.0894	3.5108	3.8724	3.0603	4.3229	-0.3294
10	3.23	3.3293	0.1024	3.1221	3.5365	2.6900	3.9686	-0.1026

Obtain the relevant predictions and confidence intervals. Explain how a participating pension fund manager would use this information. Illustrate you answer by considering the results for two different schemes.

- "y" is the actual total cost of active member and "expoP" is prediction from the model.
- In observation 2, actual cost of a member is 39.1459 but model has predicted that total cost is 46.299. With 95% assurance, total cost would be in the range of 23.79 to 90.07.
- Similarly in observation 3, actual cost of a member is 46.53, but model has predicted 28.6173. And with 95% confident, total cost would lies in between 14199 to 57.67.

ID	y	expoP	expoL	expoU
2	39.1459	46.2996	23.7994	90.072
3	46.5385	28.6173	14.1996	57.674
4	21.6473	25.1002	12.1214	51.976
5	49.3992	58.1098	29.4260	114.754
6	28.9143	38.8586	20.5087	73.627
7	29.9379	25.2397	12.8417	49.607
8	61.3333	44.8861	23.9967	83.960
9	37.4296	43.2643	23.1137	80.982
10	28.8519	40.1095	21.3347	75.406

THANK YOU