Data Analytics: Tools and Techniques

Assignment – II

Tenzin Rabyang (30017118)
Sheffield Hallam University

DATT assessment-II
02-Dec-2021

Question 1(a):

Add the variables Y and Per2 to Per7 to your data set.

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. Identify any missing values. Obtain a histogram for the response variable Y and comment briefly on its distribution.

Code: 1.1

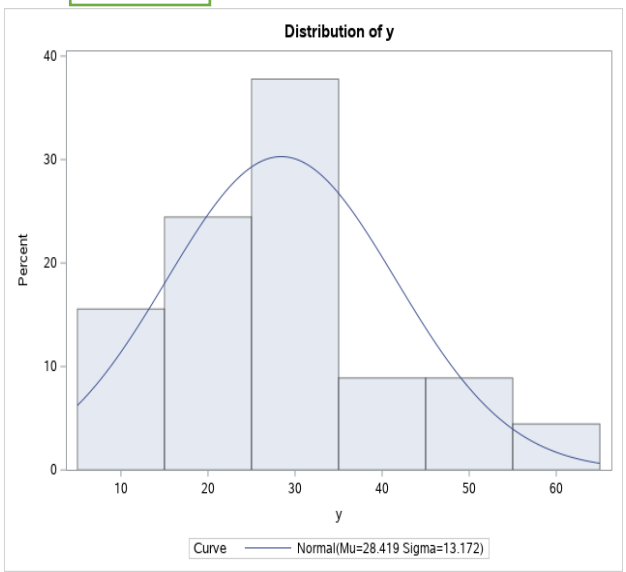
```
* Adding new variables into the data set;
data tcost;
set mydata.pfarg06;
y = 1000*(b3+b4+b6+b7+b8)/a1;
per2 = a2/a1;
per3 = a3/a1;
per4 = a4/a1;
per5 = a5/a1;
per6 = a6/a1;
per7 = a7/a1;
run;

* Running proc means on added new variables;
proc means data=tcost nmiss mean std;
var y per2 per3 per4 per5 per6 per7;
run;

* ploting histogram on variable y;
proc univariate data=tcost;
histogram y / normal;
run;
```

Table: 1.1

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

Fig: 1.1



- In Table 1.1, Not single data is missing for variable y and from per2 to per7.
- In Fig 1.1, Though histogram display near normality, it also shows little skewed towards the positive.
- It unimodal and symmetric distributed, hence it is normally distributed.

1(b):

Investigate each of the factors C1 to C8 by obtaining a simple frequency distribution for each of them. Hence explain why C4 should not be included as an explanatory variable in any regression model for Y.

```
* Running proc freq on each of factor from c1 - c8;
proc freq data=tcost;
table c1 c2 c3 c4 c5 c6 c7 c8;
run;
```

Code: 1.2

Table: 1.2

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

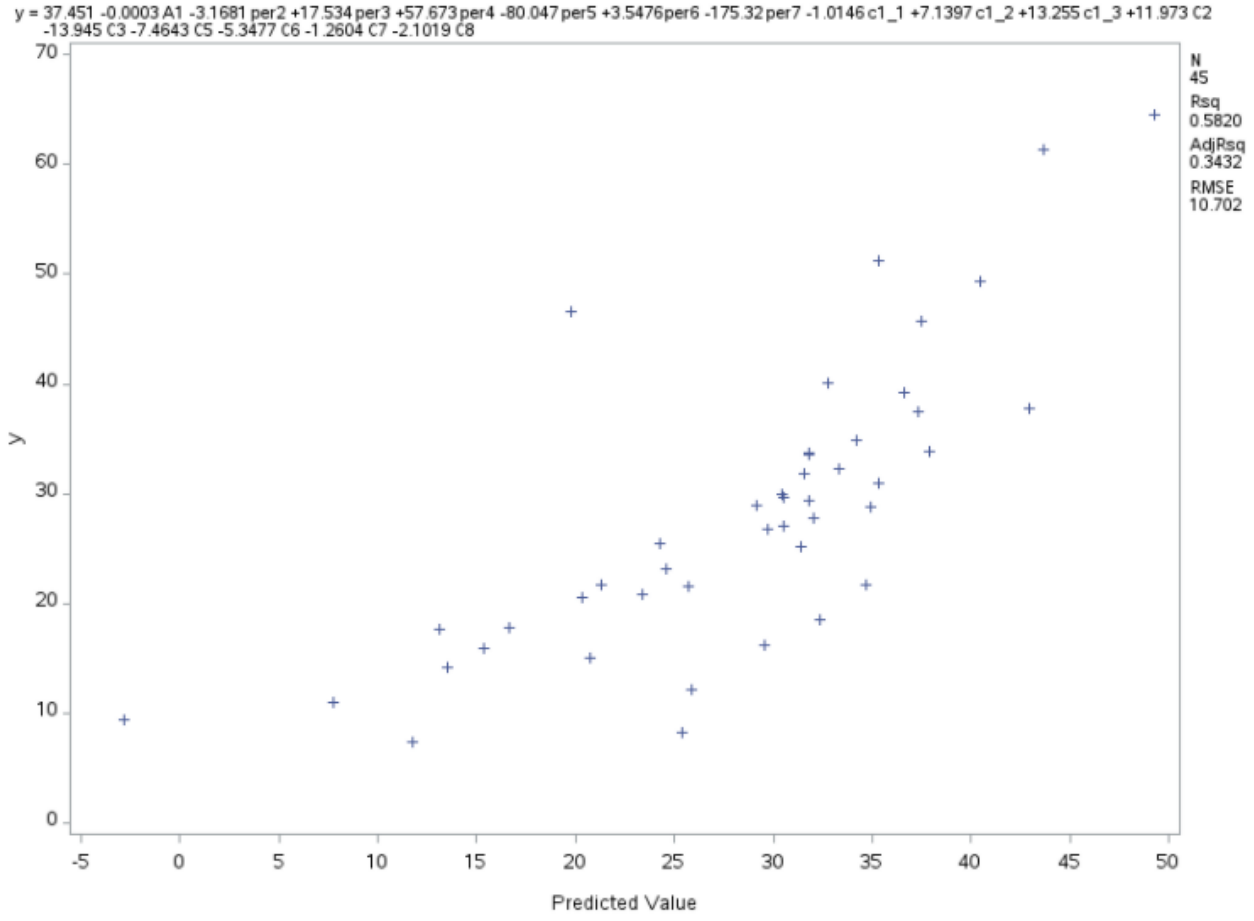
- Since, C1 contain 4 different levels, it is necessary to convert it into dummy variable because regression doesn't have class statement.
- Refer to Code: 4.1
- C1: **57.78%** of people has combine scheme with same Scale of fund type, being the largest percentage but combine scheme with different scale has lowest percentage of **4.44%**.

- C2: 88.89% of Scheme is contracted out.
- C3: 84.44% of Scheme is contributory while 15.56% isn't.
- C4: 97.78% of member can pay for AVC's.
- C5: 91.11% of Administration is at one location.
- C6: 73.33% of administrative calculation is not perform 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.

Question 2.

Fit of the systematic component of the model.



- I have scatter plot between the response variable “y” and its predicted value.
- I found 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. Carefully interpret these plots, clearly stating your conclusions concerning the adequacy of the model.

Code: 2.1

```
* performing regression model;
proc reg data=tcost;
model y = a1 per2 per3 per4 per5 per6 per7 c1_1 c1_2 c1_3 c1_4 c2 c3 c5 c6 c7 c8/vif;
plot student. *p.;
output out=plot01 p = Perdicted student=sresid rstudent=dresid;
run;
```

1. Normality.

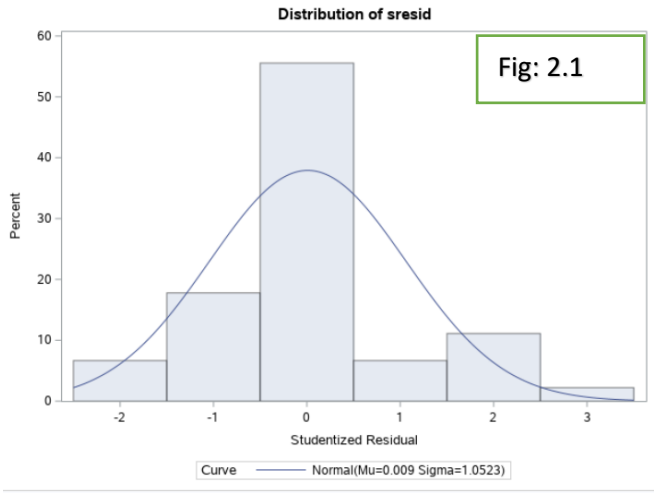


Fig: 2.1

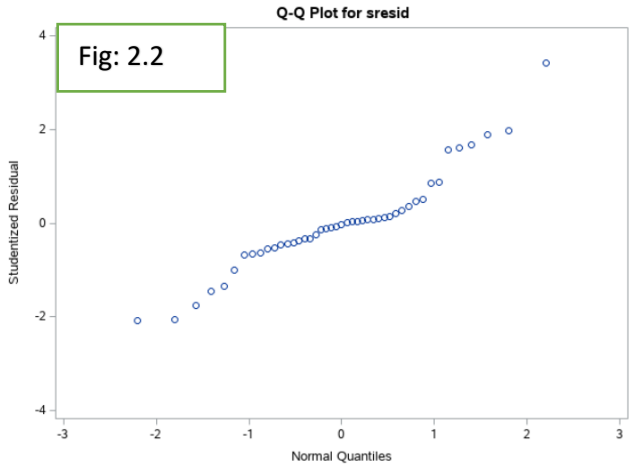


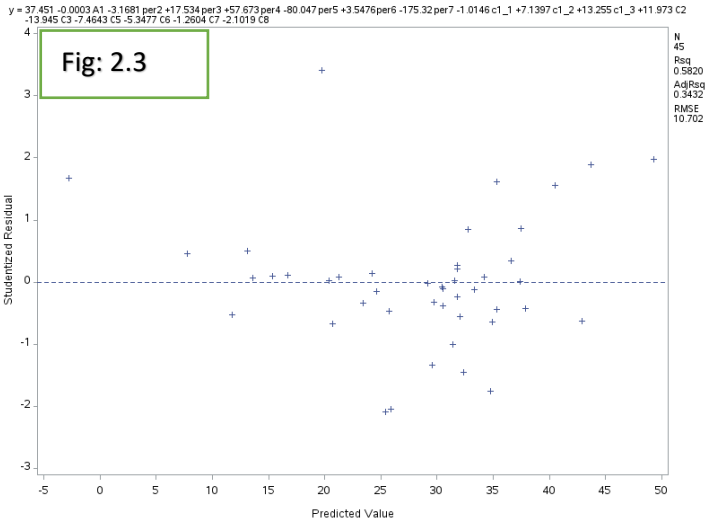
Fig: 2.2

- Considering Fig: 2.1, it is unimodal, and somewhat display symmetrical distribution, hence normality is accepted.
- We can verify using Fig: 2.2, In the Q-Q Plot, Studentized residual indeed conform to an approximate straight line near origin.
- Hence, normality assumption is accepted.

Code: 2.2

```
* Checking normality;
proc univariate data=plot01;
histogram sresid / normal;
qqplot sresid;
run;
```

2. Homoscedasticity.



- Scatter plot follows certain pattern
- Data is not constant around the zero
- One potential outlier detected
- Data is not randomly scatter; hence homoscedasticity assumption is not met.
- For code, Refer Code: 2.1.

3. Mutual Independence.

-Since Fig: 2.3 shows pattern, this means that relations between the observation might present in the model. Hence, **NOT ACCEPTED.**

4. Adequacy of the Systematic Component.

* Considering the Fig: 2.3, Studentized residuals appears not randomly scattered about a mean of value zero, without constant variance across the range. The plot is therefore do not consistent with adequacy of the systematic component of the model.

* Hence, we can't accept the systematic Component assumption.

Question 3.

To find a better model, consider the regression of log(Y) on the respective logs of A1 and Per2 to Per7, and on the (untransformed) factors C1 to C3, and C5 to C8.

a) Briefly explain why it is not necessary to transform the values of factors C1 to C3 and C5 to C8 in this modified regression model.

- C1 to C3 and C5 to C8 is Factor variable rather than interval variable.

- log transformation of data is only applicable for continues variables.

- log(0) transformation will save the data as null, where log(1) will save as 0.

- This transformation of data will disrupt the original data, so transform data is useless.

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

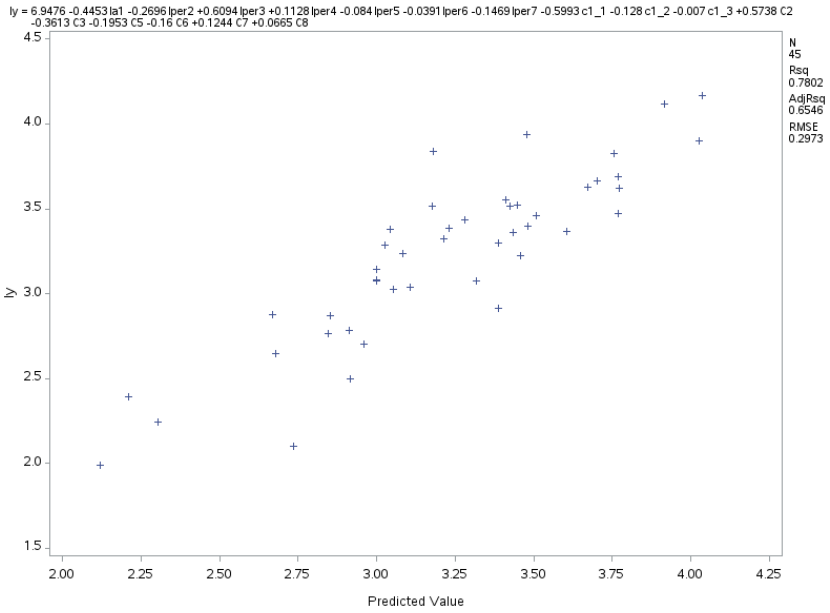
```
* Transforming the data using log();
data tcost;
set work.tcost;
ly = log(y);
la1 = log(a1);
lper2 = log(per2);
lper3 = log(per3);
lper4 = log(per4);
lper5 = log(per5);
lper6 = log(per6);
lper7 = log(per7);
run;
```

Code: 3.1

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

Table: 3.1

Investigate the fit of the systematic component of the model.



-This scatter plot shows better linear than above model.
-Value of both R^2 and adj R^2 is much higher than the above model.
-Diagram clearly shows that null hypotheses is rejected because the intercept value is higher than 0 as well as the value of exploratory parameters.
- Hence this model displays Fit Systematic Component.

Investigate the tenability of the appropriate underlying statistical assumptions. Carefully interpret these plots, clearly stating your conclusions concerning the adequacy of the model.

Code: 3.2

```
* Applying regression on transform variables;
proc reg data=tcost;
model ly = la1 lper2 lper3 per4 lper5 lper6 lper7 c1_1 c1_2 c1_3 c1_4 c2 c3 c5 c6 c7 c8/vif;
plot student. *p.;
output out=plot02 p = Predicted student = sresid rstudent=dresid;
run;
```

1. Normality.

Fig: 3.1

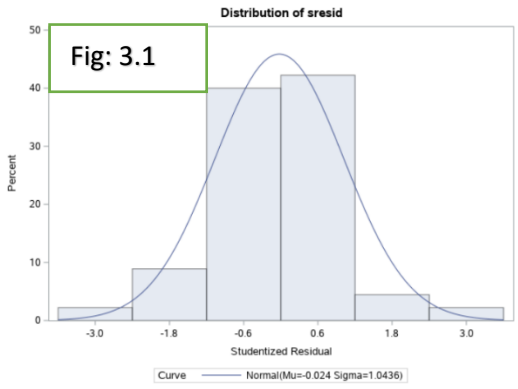
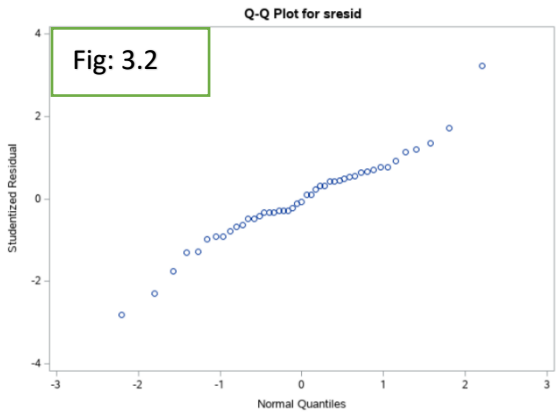


Fig: 3.2

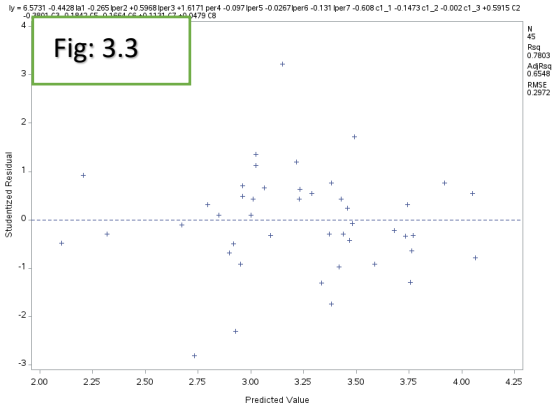


- Considering the Fig: 3.1, its Unimodal, And symmetrical distribution of data, displaying bell like structure, hence normality assumption is accepted.
- We can verify using Q-Q plot on Fig: 3.2.
- Since the plot shows linear slop of studentized residual confirm normal distribution.

Code: 3.3

```
* Checking normality;
proc univariate data=plot02;
histogram sresid / normal;
qqplot sresid;
run;
```

2. Homoscedasticity.

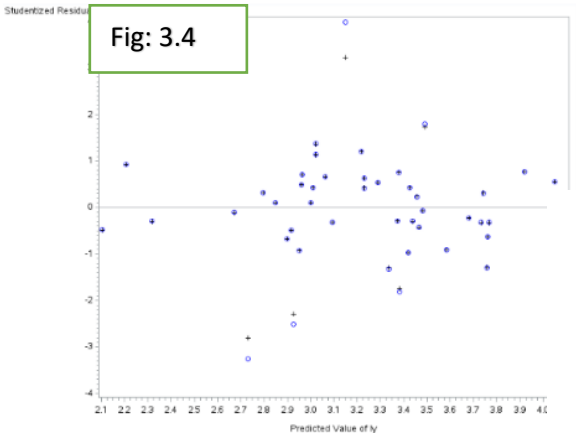


- Scatter plot on Fig: 3.3 shows, randomly scatter and constant variance across the range.
- Doesn't show any kind of pattern, hence homoscedasticity assumption is accepted.
- One potential +3 outlier is detected.
- Refer Code: 3.2.

3. Mutual Independence.

- Since Fig 3.3 show randomly scatter, hence observation is **mutually independent.**

4. Adequacy of the Systematic Component.



Code: 3.4

```
* plot of Sresid and Dresid;
symbol1 v = plus c = black;
symbol2 v = circle c = blue;
proc gplot data=plot02;
plot (sresid dresid)* Predicted / overlay vref=0;
run;
quit;
```

- Figure shows similar scatter plot with Fig: 3.3
- It is randomly scatter with constant variance across the zero range for both deleted and studentized residual.
- We accept this assumption
- One potential outlier detected near +3.

Now comparing current model with earlier model, consider earlier model as M1 and current model as M2. I found,

1. **Normality.**

- In both cases, the normality assumption is accepted.
  M1: Accepted
  M2: Accepted

2. **Homoscedasticity.**

- In M1, plot shows certain kind of pattern instead of random, hence rejected.
  M1: Not Accepted
  M2: Accepted

3. **Mutual Independence.**

- Since, I didn't notice any sort of relations between the observation in both models.
  M1: Not Accepted
  M2: Accepted

4. **Adequacy of the Systematic Component.**
   M1: Not Accepted
   M2: Accepted

Table: 3.1

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

Question 4.

(a) You should now reduce the model introduced in Question 3 by removing "unnecessary" variables and factors, and hence identify an overall "best" model for the prediction of log(Y). Explain why approaches based upon a consideration of all possible models are impracticable in this instance.

- Since Categorical variable C1 contain 4 different levels, I have created dummy variable because **proc reg** procedure doesn't have **class** statement.

```
* Creating dummy variable;
data tcost replace;
set work.tcost;
if c1 = 1 then c1_1 = 1;
else c1_1 = 0;
if c1 = 2 then c1_2  = 1;
else c1_2 = 0;
if c1 = 3 then c1_3 = 1;
else c1_3 = 0;
run;
```

Code: 4.1

Using R^2 method

```
*Implementing R2 method;
proc reg data=tcost;
model ly = la1 lper2 lper3 lper4 lper5 lper6 lper7 c1_1 c1_2 c1_3 c2 c3 c5 c6 c7 c8 /
selection = rsquare mse;
run;
quit;
```

Code: 4.2

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

Table: 4.1

- I have gathered best of model form every group as we can see in Table: 4.1.
- Now, considering the table I found that model with **k = 10** is the best model because it has the lowest value of MSE i.e., 0.07926.
- Variables contain in this model is **la1, lper2, lper3, lper4, lper7, c1_1, c1_2, C2, C3, and C7.**
- Because MSE is error mean square which determine the value of error present in the model, hence naturally we want model with less error.
- Even the R^2 value is not small nor large but somewhere in the middle.
- Hence, we can say using R^2 and MSE method, we found that above model is the best model.

Using Adjusted R^2 Method

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

Table: 4.2

Code: 4.3

```
* Implementing Adjusted R2 method;
proc reg data=tcost;
model ly = la1 lper2 lper3 lper4 lper5 lper6 lper7 c1_1 c1_2 c1_3 c2 c3 c5 c6 c7 c8 /
selection = rsquare adjrsq;
run;
quit;
```

- Now considering the Adjusted R^2 method, I found that model with **k = 10** has highest value of 0.6903.
- This model contains variables of **la1, lper2, lper3, lper4, lper7, c1_1, c1_2, C2, C3, and C7.**
- Both models chosen by the MSE and adjusted R^2 is exactly the same variables with same number of k, which is 10.
- As we can notice from the Table: 4.2, value of adjusted R^2 steadily increase until k = 10, then again reduces after 10.
- Since **k = 10** has the highest value of adjusted R^2, this method also displays the same best model with MSE method.

Using Cp method

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

Table: 4.3

Code: 4.4

```
* implementing Cp method;
proc reg data=tcost;
model ly = la1 lper2 lper3 lper4 lper5 lper6 lper7 c1_1 c1_2 c1_3 c2 c3 c5 c6 c7 c8 /
selection = rsquare cp;
run;
quit;
```

- Since Cp measure the discrepancies of sum of squares of current model, Cp needs to be low as possible.
- Hence form Table: 4.3, I notice that Cp value decreases till k = 7 (being the lowest), and then again increases from k = 8.
- Since **k = 7** model has the lowest Cp value, I conclude that this model is best model using Cp method.
- This model contains variables of **la1, lper2, lper3, c1_1, c1_2, C2, and C3.**

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.

(b) Hence employ a backward elimination procedure, justifying your choice of final model.

Code: 4.5

```
* Regression Backward elimination model;
proc reg data=tcost;
model ly = la1 lper2 lper3 lper4 lper5 lper6 lper7 c1_1 c1_2 c1_3 c1_4 c2 c3 c5 c6 c7 c8 /
selection = backward slstay = 0.05;
run;
quit;
```

Table: 4.4

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

- The final model produces using BACKWARD method is shown in Table: 4.4.
- The model contains total of 4 exploratory variables.
- All of them are significant.
- la1 has highest significant level while lper2 has the lowest.
- R^2 value, **0.6626**, is relative low comparing with another model produce by MSE, adjusted R^2 and Cp. This is because of lower amount of variable present in the model.
- Cp value, **7.9841**, is quite higher than model produce in Cp method.
- Also, Value of MSE, **0.09497**, is large as compared to model in R^2 method.
- Hence, I am choosing different model which is better model than above model.

Table: 4.5

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

- Now the Table: 4.5 is **my best model.**
- Comparing with final model produce my backward elimination, it has lower value of both Cp and MSE. Meaning less error in model.
- Secondly, R^2 is relatively large around 7.
- Though, this model contains one low non-significant variable, C2, but this model will produce better prediction of response variable with less error.
- **Hence, I choose this model as my best model.**

c)

```
* Final best model parameter estimation;
proc reg data=tcost;
model ly = la1 lper2 lper3 c1_1 c2;
run;
quit;
```

Code: 4.6

Table: 4.6 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

- Total of **6 parameter** including Intercept and exploratory variable.
- Parameter Estimate value of intercept is **7.16004**. This is the default value of response variable (y-axis) if exploratory variable is **zero.**

Parameter Estimate:

* I found 3 variable which has negative relation with the response variables are la1, lper2, and c1_1. Because (**-ve)** sign in the parameter estimate shows that mean of the response variable will have opposite reaction to the exploratory variable. If the value of exploratory increases, value of response variable will decrease and vice versa.

* However, the degree of the reaction is determined by the value of corresponding parameter and all of parameter estimate value seems in the range of -0.4811 to 0.2175.

P value:

* Now, among the variables, 4 of them are significant with respect to their t-value, where **la1** is the most significant one.

* only C2 variable is not significant in the model.

Standard Error:

* Now, Standard error are like the standard deviation, which measures the spread to the data.

* Large standard error means, more spread of the data, while low standard error means less spread of the data and data will be near mean value of the data sample.

* **c1_1** has the highest Standard error of **0.18878** while **la1** has lowest of **0.04795.**

t-value:

* Negative value is **t-value** shows negative relation of response variable with corresponding exploratory variable and positive values shows, positive relationship.

* t-value usually use to determine the hypotheses testing but it can also determine the similarity between the response variable and exploratory variable. More the t-value, lesser the similarity or lesser the t-value, more similarity between the variables.

* **la1** has the largest t-value while **C2** has the lowest t-value.

QUESTION 5

Use appropriate plots to investigate the fit of your final model. In addition to the overall fit, investigate the fit with respect to each of the explanatory variables (but not the classifications or factors) present in this model.

Exploring fit of the model.

```
* Checking fit of the FINAL model;
proc reg data=tcost;
model ly = la1 lper2 lper3 c1_1 c2;
plot student. *p.;
output out=plot03 p=predicted student=sresid rstudent=dresid;
run;
quit;
```

Code: 5.1

1. **Normality.**

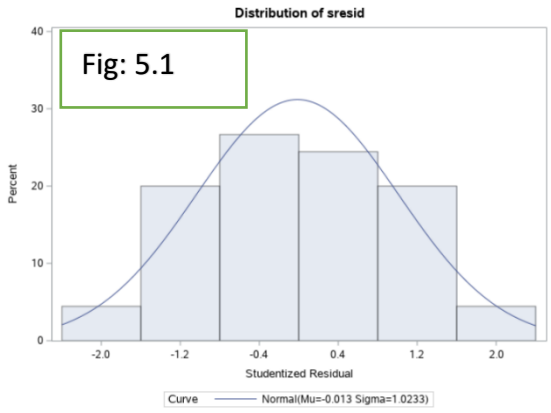


Fig: 5.1

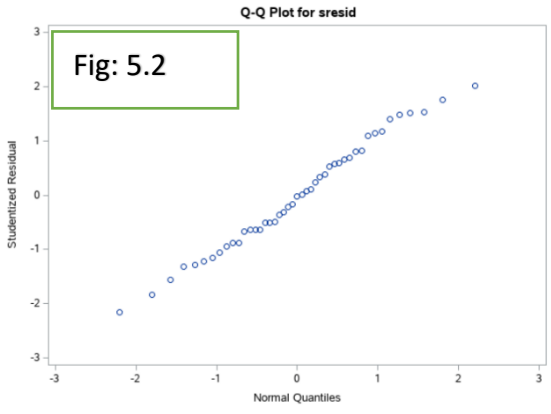


Fig: 5.2

```
* Checking Normality of Final model;
proc univariate data=plot03 noprint;
histogram sresid / normal;
qqplot sresid;
run;
```

Code: 5.2

- I have use histogram to check the normality of final model.
- Model is Uni-model, and symmetrically distributed hence **Normality is Accepted.**
- We can use Q-Q plot to verify this.
- The studentized residual indeed conform to an approximate straight line of unit slope passing near to the origin.

2. **Homoscedasticity.**

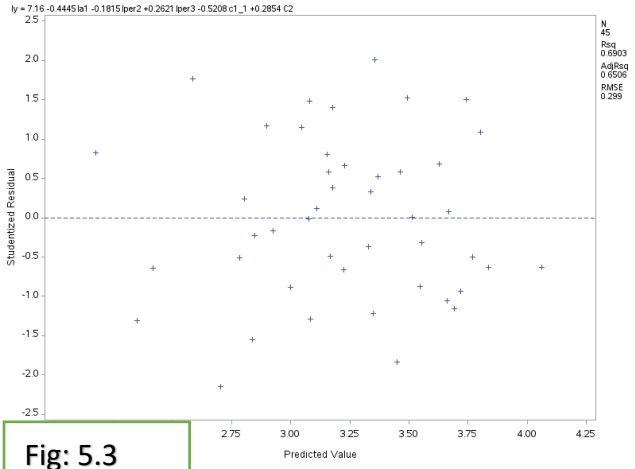


Fig: 5.3

- The studentized residuals is randomly scatter about a mean value of zero.
- It shows constant variance across the range of fitted values.
- It doesn't show any kind of pattern.
- **Hence, Homoscedasticity is ACCEPTED.**
- For code, Refer Code: 5.1.

3. **Mutual Independence.**

- Since I didn't notice any sort of relations between the observations, hence Mutual Independence is ACCEPTED.

4. **Adequacy of Systematic Component.**

```
* Checking Systematic Component of final model;
symbol1 v = plus c = black;
symbol2 v = circle c = black;
proc gplot data=plot03;
plot (sresid dresid)*predicted / overlay vref=0;
run;
quit;
```

Code: 5.3

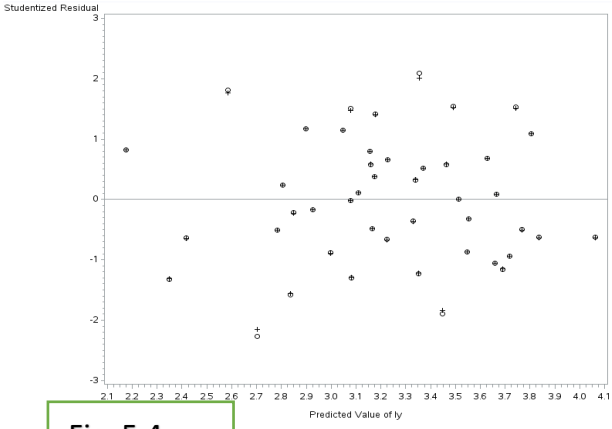


Fig: 5.4

- Both studentized and deleted residuals are almost perfectly superimposed on the plot. Fig: 5.4.
- Though we notice few separations between them but distance is marginally small. And couple of points are -2 to +2, but very near its range.
- Since, most extreme residual seems to be within the range of +3 to -3, so it is unlikely to have any outlier in the model.
- Hence Systematic Component is **Accepted.**

- Now investigating fit with respect to each of the model for interval variables.

la1

- **Normality**

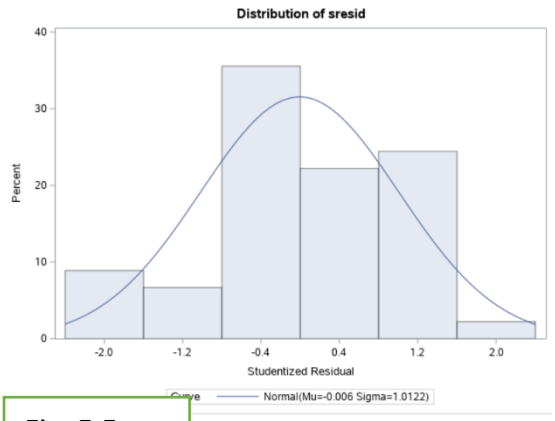


Fig: 5.5

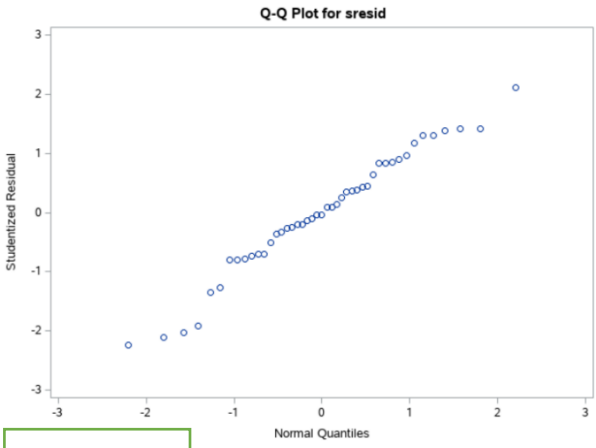


Fig: 5.6

- Its Uni-model with symmetric.
- Bell like shape
- Hence normality accepted.

- Q-Q plot shows near linear straight line passing through the origin.

- **Homoscedasticity**

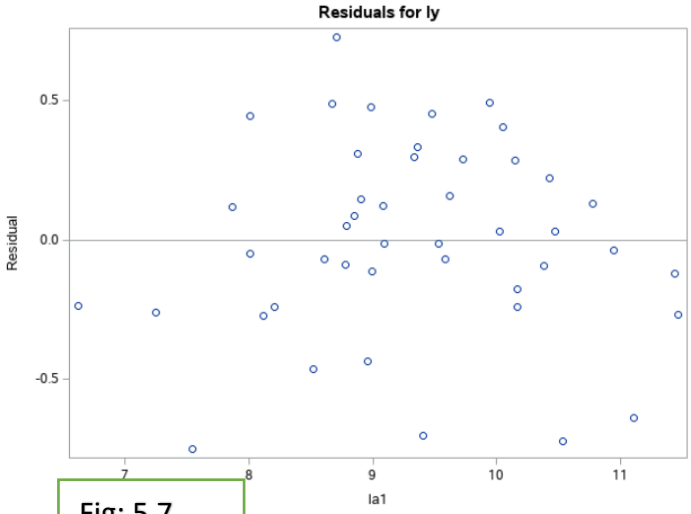


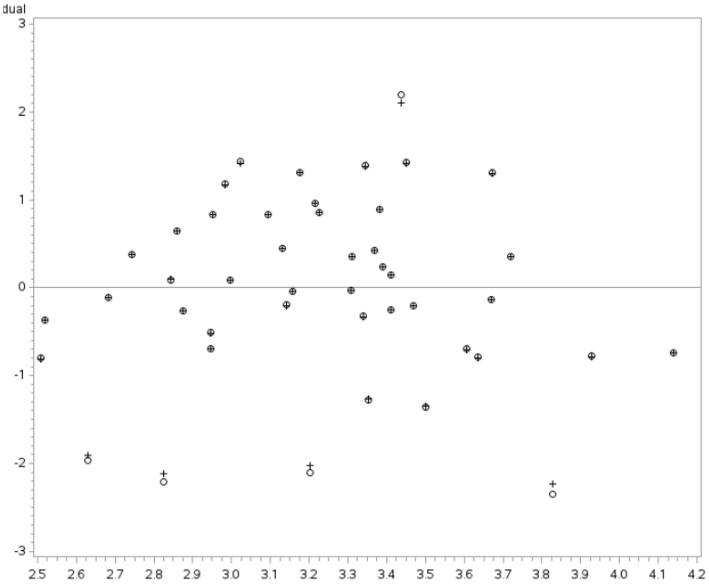
Fig: 5.7

- Residual are randomly scatter around zero.
- With constant variance.
- No pattern is detected.

Mutual Independence.

- Since fig: 5.7 doesn't show any sort of relation present between the observation.

Systematic Component.



- Most of the residual are imposer with each other.
- Only few are not by marginally difference.
- All residual are seems well within the range of ±3.

Fig: 5.8

lper2

Normality

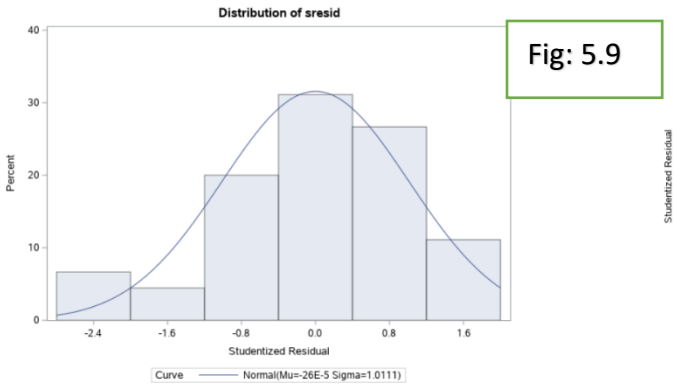


Fig: 5.9

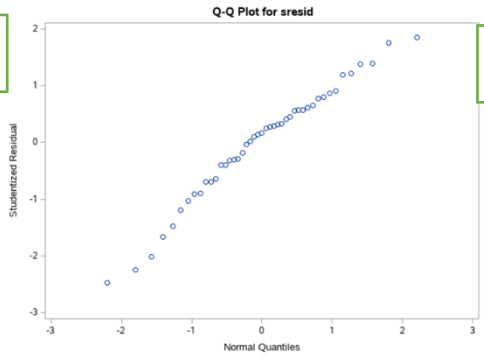


Fig: 5.10

- It Uni-model with symmetric.
- Display bell like shape
- Little negative skewed

- Q-Q plot show straight line passing through origin.

Homoscedasticity

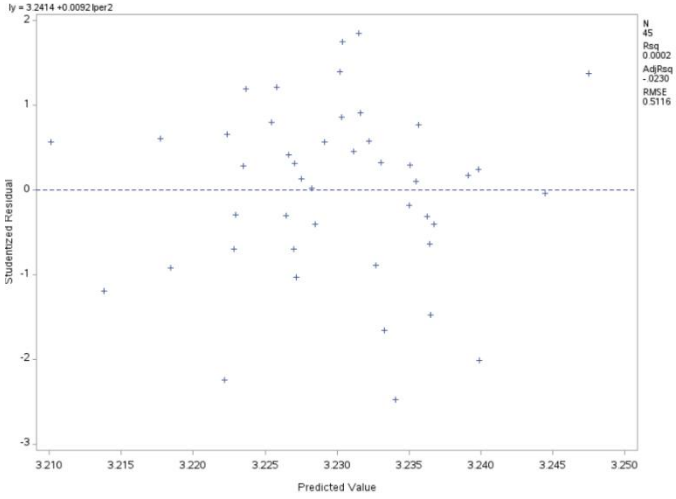


Fig: 5.11

- Studentized residual is randomly scatter
- No show of pattern
- But not so sure about the Constant variance

Mutual Independence

- Since there is no pattern, So relation between the observation.

Systematic Component

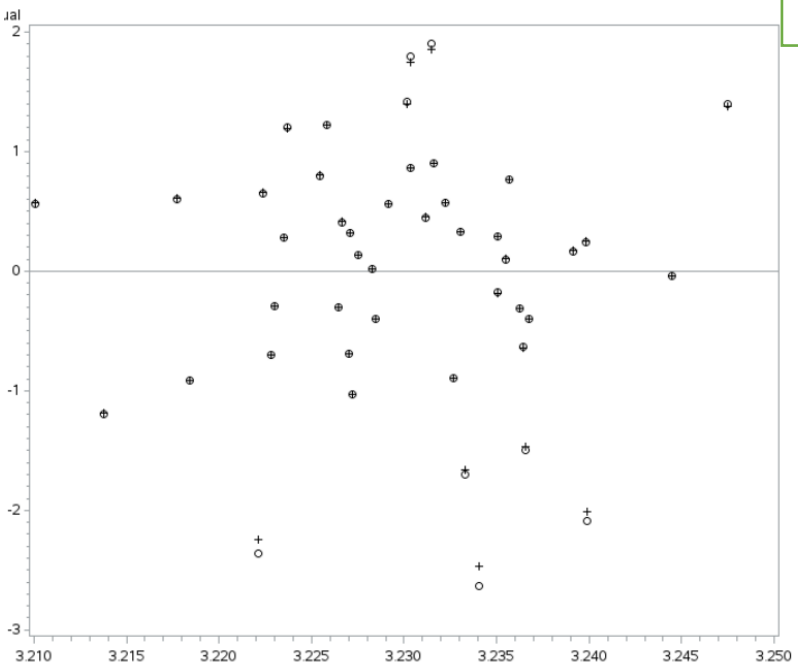


Fig: 5.12

- Both residuals are imposed with each other.
- Few of them are not perfectly imposer but with little distance.
- It is again randomly scatter
- All residual are well with the range of ± 3.

lper3

Normality

Fig: 5.13

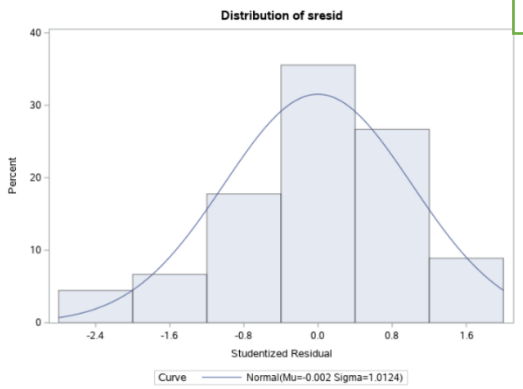
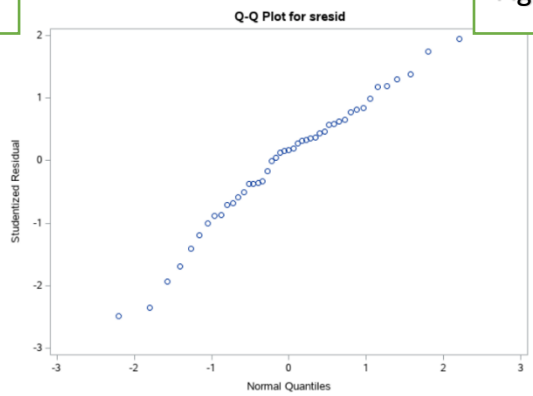


Fig: 5.14



- Its Uni-model with symmetric
- Bell like shape
- Little negatively skewed

- plot shows straight line passing through origin.

Homoscedasticity

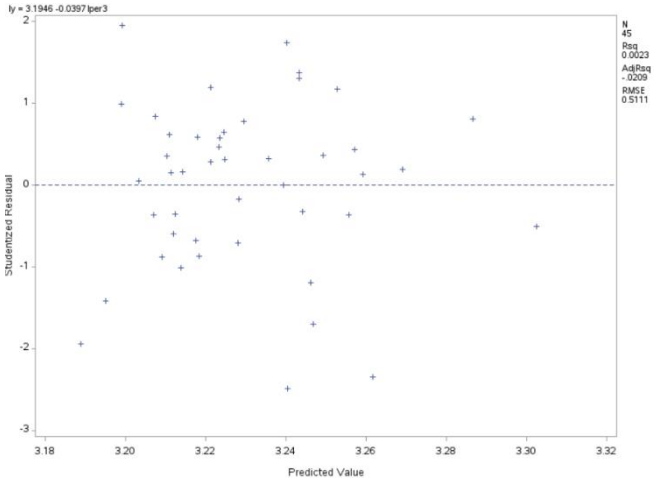


Fig: 5.15

- It is randomly scatter
- Not so sure about the constant variance
- No sign of pattern

Mutual independence

- since no sign of pattern is notice, observation are mutually independent with each other.

Systematic Component

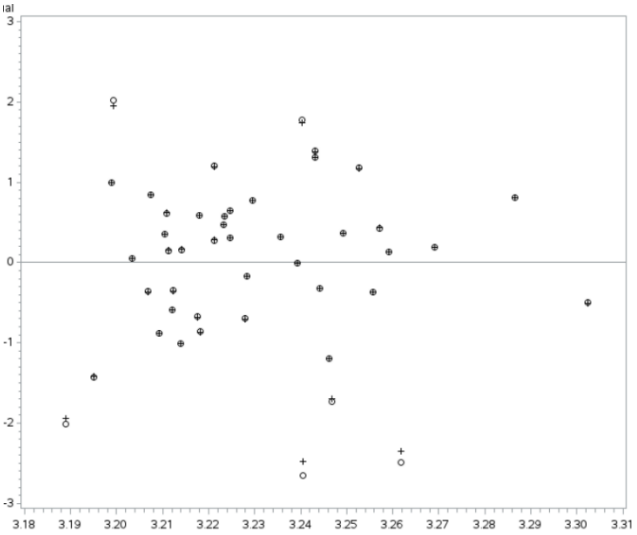


Fig: 5.16

- Both residuals are imposer with each other
- Few of them are not impose with each other but with little distance
- It randomly scatter
- No pattern recognizes
- Residuals are well within the range of ±3.

QUESTION 6

a) Using your final model, investigate and briefly discuss any issues relating to outliers or influential points, considering implications only for the overall fit of the model.

Outlier

* Considering the Fig: 5.3, studentized residual and deleted residual are perfectly superimposed with each other in the plot.

* Couple of them are not perfectly superimpose with very small marginally distance and most extreme residual seems to be inside the range of +3 to -3. Hence it is unlikely to have any sort of outlier in the model.

Influential Points

```
 * Checking for influential points;
 proc reg data=tcost noprint;
 model ly = la1 lper2 lper3 c1_1 c2;
 output out=plot04 dffits= Dff;
 run;
 quit;

 * Adding absolute value of Diffits;
 data plot04;
 set plot04;
 adff = abs(Dff);
 run;

*Sorting the data;
proc sort data=plot04;
by descending adff;
run;

* Adding ranks of Diffits;
data plot04;
set plot04;
index = _N_;
run;

* plotting the Adff;
proc gplot data=plot04;
plot adff*index / vref = 0.730 vref = 2;
run;
quit;
```

Code: 6.1

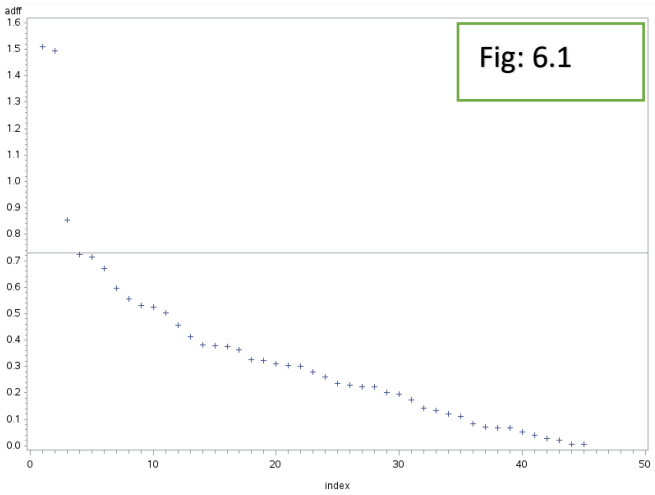


Fig: 6.1

- Cutoff value of vref is calculated as 2*sqrt(p/n), where "p" is number of parameters i.e., 6 including the intercept and "n" is number of observations, i.e., 45. Hence I got verf = 0.730
- Since there is three points above out reference line, however, one is near to the line and two them are observed around 1.5.
- Since no Absolute value of DIFFITS observed no way near the 2, second reference line, it unlikely that those points are influential.
- How to make it sure, I have run test on these points as follow.

Table: 6.1

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

- Table: 6.1 shows two potential influential points with respect to variables observations.

- We can also notice these points in the Fig: 6.1.
- However, this table is not enough to identify the exact influential points, hence we need to test or diagnose these two points further using leverage, deleted residual and covariance ratio as follow.

```
* printing relevent statistics;
proc print data=plot04;
var id ly predicted dff H dresid C;
where index <= 2;
run;
```

Code: 6.2

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

Table: 6.2

- Considering the Table: 6.2, Dff is difference in fitting, H is leverage, dresid is deleted residual and C is Covariance ratio.

Dff (Difference of fitting)

- Since cut-off for Dff is 0.730, where both observation is above the cut-off value, which means these two observations might influence on model prediction. However, Sine both value of Dff is less the 2, I can say that these two observations will have very weak influential and not something we should worry about.

H(leverage)

- Average value of H has p/n = 6/45 = 0.13333 and cut-off value is 3*0.133333 = 0.4.
- I found that second observation has more H value than its cut-off value, where first observation is not.
- So first observation will have lower influence compare with observation, however, since it is only marginally difference from the cut-off, this isn't something we should worry about.

Dresid (deleted residual)

- Observation 1 has high deleted residual which is above our reference line 2. But since there is very small marginally difference or slightly higher than 2 by 0.09, this observation will have none or very weak influence in model prediction.
- However, observation 2 is well within the range of 2, so don't have to concern much with that.

C (Covariance Ratio)

- Corresponding range of C can be calculated as 1±3(p/n) = 0.6 and 1.4.
- Observation 1 has its value very close to 1 means, it will have very weak influence in model. So no need to concern about.
- However, observation 2 is outside of the range by 0.1, very small margin, hence this isn't something we should concern about.

Conclusion.

- Though I found 2 potential influential points, where observation 2 shows more influence than observation 1.
- However, none of them is sufficiently abnormal to give much of concern, hence I can include these two observation in multiple regression analysis.

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

```
* Checking for multi-collinearity in final model;
proc reg data=tcost corr;
model ly = la1 lper2 lper3 c1_1 c2 / vif collin;
run;
quit;
```

Code: 6.2

Table: 6.3

Correlati

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

Correlation.

- The maximum correlation I found is between the **lper3** and **la1** with 0.326.
- Since all the correlation value are less than 0.5 or more than -0.5, we can consider that all the exploratory variable has very weak correlation.
- Since **ly** is the response variable, we can't use that variable for correlation analysis.

Variance Inflation Factors.

Table: 6.4

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

- $1/1-R^2$
- Considering the value of Variance Inflation, Value of VIF seems to much lower than 10 for every variable in the model.
- Every variable has value around 1, which means low VIF.
- Hence, no correlation is detected using VIF method.

Condition Indices

Table: 6.5

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			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

- Firstly, I found column 6 has the highest value of condition index of 30.020, which is more than 30.
- Secondly, that condition index has 3-4 times higher value than its preceding one.
- Since column 6 applied both rules, now we can consider this column.

- **la1** has the highest value of 0.967 followed by **lper3** of 0.17943.
- Since huge gap is present between these two values, I can't assume both variables are correlated with each other.
- Also, la1 is the only variable with high value, so it is save to say that no correlation between the exploratory variable is present in the model.

QUESTION 7

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.

Code: 7.1

```
* getting confident intervals of model;
proc reg data=tcost;
model ly = la1 lper2 lper3 c1_1 c2/cli clm;
run;
```

Table: 7.1

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

- Considering the table, 7.1, I notice that 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.

Code: 7.1

```
proc reg data=tcost;
model ly = la1 lper2 lper3 c1_1 c2;
output out=plot05 p=predicted lcl=lower ucl=upper;
run;
quit;

data Ftcost;
set tcost;
run;

* Data are reverting the log transformation;
data Ftcost;
set plot05;
expoP = exp(predicted);
expoL = exp(lower);
expoU = exp(upper);
run;

*Printing the data;
proc print data=Ftcost noobs;
var id y expoP expoL expoU;
run;
quit;
```

Table: 7.2

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

- Considering the table 7.2, “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.