Applied Regression Analysis

STAT 512

Statistical Analysis of Cheddar Cheese Dataset

BY

AMEYA THOMBRE

JANANEE PARTHASARATHY

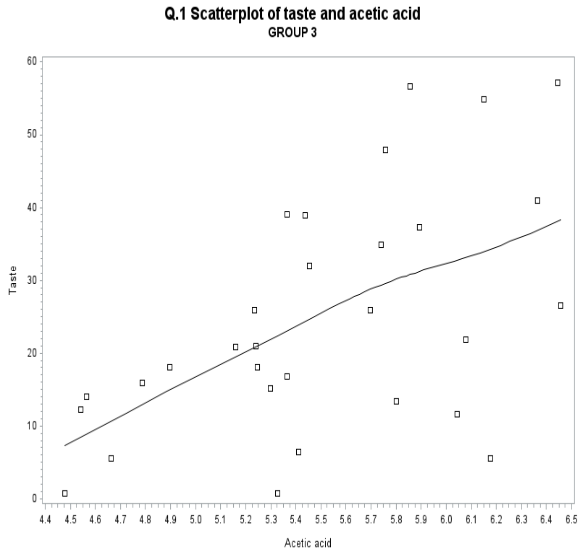
KARTHIK SAJEEV

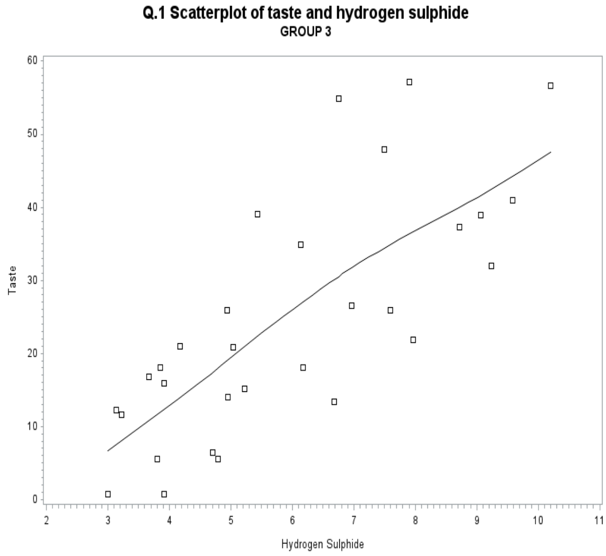
SANDESH GEORGE OOMMEN

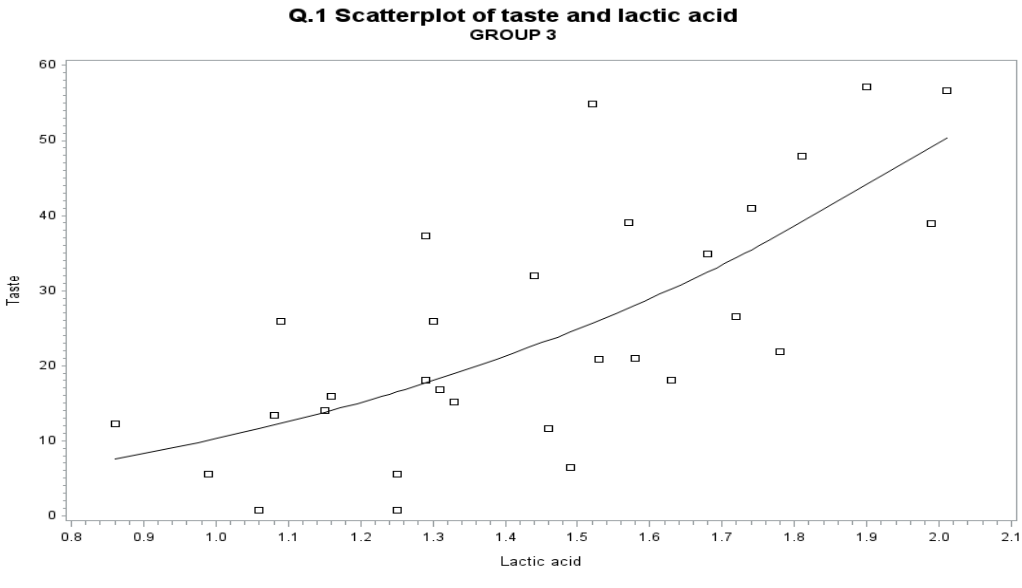
PART-1

Question 1:

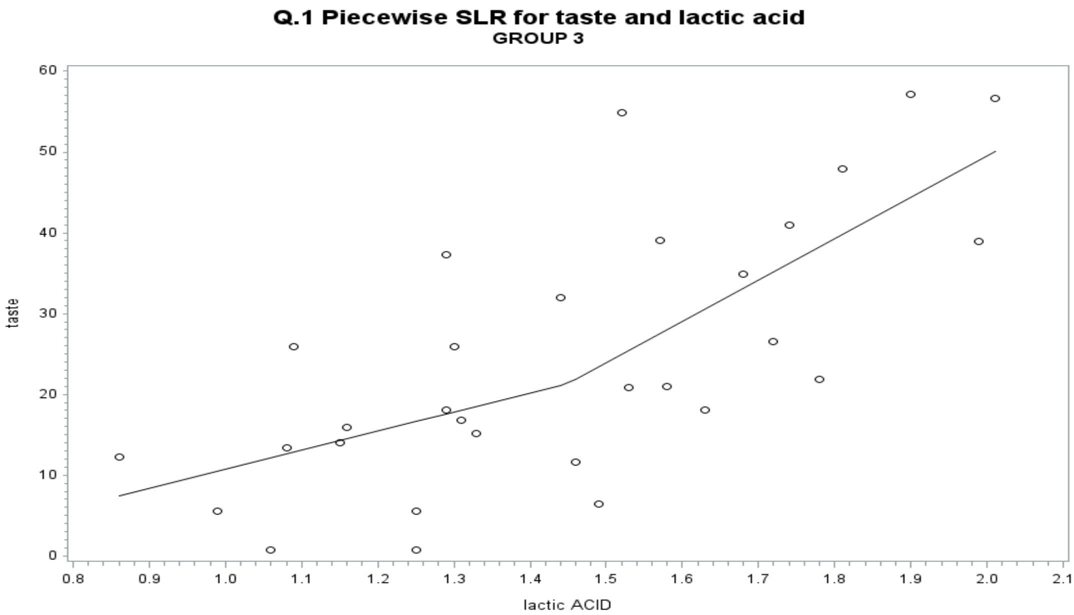
We must choose a predictor to perform Piecewise SLR and model its relationship with the response variable. To choose the predictor, we generate the individual scatterplots of the response variable with each of the explanatory variables.







On analysing the individual scatter plots, we observe that the relationship between taste and acetic acid and between taste and H2S is fairly linear. However, the relationship between taste and lactic acid is curved. Hence, we decide to run Piecewise Regression on Lactic Acid. We split the data at the value of lactic acid=1.45, as this is the center point and approximately where the relationship bends. We get the plot as follows:



The regression results of the Piecewise Regression are given below:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-12.64843	19.52769	-0.65	0.5226
lactic	1	23.41079	15.42478	1.52	0.1407
CSLOPE	1	28.06070	26.77792	1.05	0.3040

The piecewise regression gives us the following equation for the model:
taste= -12.64843 + 23.41079(lactic) + 28.06070 (cslope)
We define cslope such that:
cslope = 0 if lactic $\le$ 1.45
cslope = (lactic -1.45) if lactic > 1.45

Taste = $\beta_0 + \beta_1$*lactic +β_2*cslope

Using the definition of cslope we can break this equation into 2 different equations. These two equations represent the equation of the two lines on the piecewise plot. Using these two equations, we perform the sameline test to make sure that the two lines are not the same.

taste = $\beta_0 + \beta_1$ *lactic   if lactic $\le$ 1.45
taste = $\beta_0 + \beta_1$ *lactic +β_2 *(lactic - 1.45)   if lactic $\ge$ 1.45

For the two lines to be same, β_2 would be equal to 0. This would make the equation of both the lines same. The sameline test is performed and the results are as follows:

Q.1 Test to determine if lines are same
GROUP 3

The REG Procedure
Model: MODEL1

Test SAMELINE Results for Dependent Variable taste				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	150.95007	1.10	0.3040
Denominator	27	137.46439		

Let us state the hypotheses for the test:
Null hypothesis H_0: $\beta_2=0$
Alternate hypothesis H_a: $\beta_2 \neq 0$
Test statistic $F_{1,27} = 1.10$
The p-value is 0.3040> alpha (0.05).
Thus, we fail to reject the null hypothesis that $\beta_2=0$ for $\alpha=0.05$(95% confidence level).
So we do not have statistical evidence to show that $\beta_2 \neq 0$.
So we can conclude that both lines are in fact the same.

Question 2

SAS outputs for the following predictions are shown:

(a). (i). taste using acetic:

Q.2a i. Predicting response using all explanatory variables
GROUP 3

The REG Procedure
Model: MODEL1
Dependent Variable: taste

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2314.14151	2314.14151	12.11	0.0017
Error	28	5348.74515	191.02661		
Corrected Total	29	7662.88667			

(ii). taste using SUM and acetic:

Q.2a ii. Predicting response using all explanatory variables and SUM
GROUP 3

The REG Procedure
Model: MODEL1
Dependent Variable: taste

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4695.89081	2347.94541	21.37	<.0001
Error	27	2966.99585	109.88874		
Corrected Total	29	7662.88667			

Extra sum of squares for the comparison may be computed as:
SSM(F-R) = SSM(F) - SSM(R) = SSM(SUM|acetic) = SSM(SUM,acetic) - SSM(acetic)
= 4695.8908 – 2314.1415 = 2381.7493
The general linear test statistic is obtained using:
F-statistic = MSM(F-R)/MSE(F)= (SSM(F-R)/DFM(F-R))/(SSE(F)/DFE(F))
Obtaining relevant values from the table:
DFM(F) = 2
DFM(R) = 1
DFM(F-R)= DFM(F)- DFM(R)=2-1=1
SSE(F) = 2966.9958
DFE(F) = n-p=30-3=27
Therefore, F statistic = (2381.7493 / 1) / (2966.9958 / 27) = 21.6742
Numerator degrees of freedom = DFM(F-R)=1
Denominator degrees of freedom = n–p = 30–3 = 27

(b). Using the test statement in proc reg yields the following:

Q.2b Test statistic using test statement in proc reg
GROUP 3

The REG Procedure
Model: MODEL1

Test sum_coefficient Results for Dependent Variable taste				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2381.74930	21.67	<.0001
Denominator	27	109.88874		

Null hypothesis H_0 : $\beta_{sum} = 0$ (regression coefficient of sum equals 0)
Alternate hypothesis H_a : $\beta_{sum} \neq 0$ (regression coefficient of sum is not equal to 0)
Test statistic, F = 21.67
Degrees of freedom = (1, 27)
p-value: < 0.0001
Conclusion: As the p-value is less than 0.05(alpha), we reject the null hypothesis that the regression coefficient of SUM is zero. This means that the SUM variable is a good predictor of taste when acetic is there in the model and so SUM should not be removed from the model when acetic is the only other predictor.

(c).

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25.86888	20.33960	-1.27	0.2143
SUM	1	5.06178	1.08726	4.66	<.0001
acetic	1	2.36943	4.44542	0.53	0.5984

From the figure above,
Individual t-test for SUM:
Test statistic, t = 4.66 , p-value = < 0.0001
Test statement in proc reg:
Test statistic, F = 21.67
We know that $F = t^2$. Therefore, $t = \sqrt{21.67} = 4.66$
p-value = < 0.0001
As we can see, both methods give identical results. This is because the test statement checks the null hypothesis that the SUM has regression coefficient equal to zero, and the individual t-test for the coefficient of SUM checks the same. Both confirm that regression coefficient of SUM is not equal to zero and so SUM is a good predictor when acetic is the only other predictor in the model.

Question 3 :

SAS output for predicting the response using all predictors except SUM is given below:

Q.3 TYPE I AND TYPE II SS- all predictors except SUM
GROUP 3

The REG Procedure
Model: MODEL1
Dependent Variable: taste

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4994.47558	1664.82519	16.22	<.0001
Error	26	2668.41109	102.63120		
Corrected Total	29	7662.88667			

Root MSE	10.13071	R-Square	0.6518
Dependent Mean	24.53333	Adj R-Sq	0.6116
Coeff Var	41.29364		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-28.87677	19.73542	-1.46	0.1554	18057	219.72726
acetic	1	0.32774	4.45976	0.07	0.9420	2314.14151	0.55427
h2s	1	3.91184	1.24843	3.13	0.0042	2147.01680	1007.65809
lactic	1	19.67054	8.62905	2.28	0.0311	533.31727	533.31727

The order of variables given in the model statement is : acetic, h2s, lactic.
SS1 sum = 2314.1415 + 2147.0168 + 533.3173 = 4994.4756
SS2 sum = 0.5543 + 1007.6581 + 533.3173 = 1541.5297
SSM = 4994.4756
Thus, the Type I sums of squares add up to the model sums of squares.
For the predictor 'lactic', Type I and Type II sums of squares are the same.

This may be explained as follows:
'lactic' is the last among the 3 predictor variables as per the specified order. SS1 is defined as the extra SS for each variable, given that all previous predictors are in the model. SS2 is defined as the extra SS for each variable, given all other variables in the model. For the last predictor, 'all other variables' and 'all previous variables' are equivalent. Hence, they have the same value.

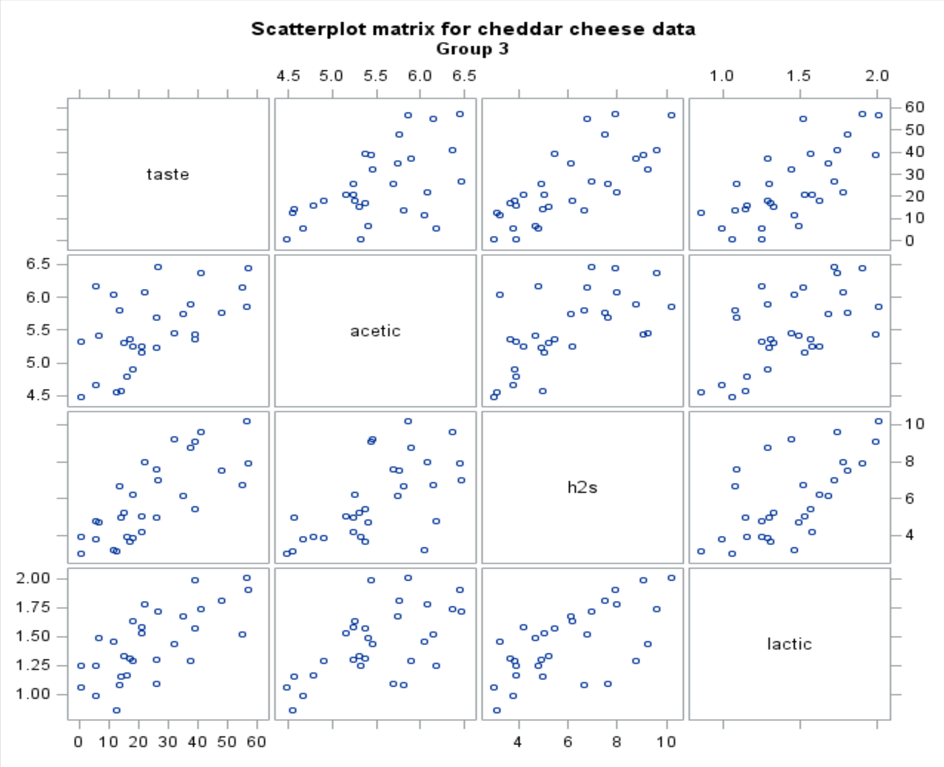
Question 4:

We run the regression to predict the response using a variety of combinations. SUM is considered an explanatory variable. We summarize the results by making a table giving the percentage of variation explained (R^2) by each model:

Sl. No.	**Explanatory variable(s)**	**R^2**	**Adj. R^2**
1	acetic	0.3020	0.2771
2	h2s	0.5712	0.5558
3	lactic	0.4959	0.4779
4	h2s lactic	0.6517	0.6259
5	acetic lactic	0.5203	0.4847
6	h2s acetic	0.5822	0.5512
7	SUM	0.6087	0.5948
8	SUM acetic	0.6128	0.5841
9	SUM h2s	0.6517	0.6259
10	SUM lactic	0.6517	0.6259
11	acetic h2s lactic	0.6518	0.6116

Part II

Question 1:



Correlation Matrix for cheddar cheese data
Group 3

The CORR Procedure

4 Variables:	taste acetic h2s lactic

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
taste	30	24.53333	16.25538	736.00000	0.70000	57.20000
acetic	30	5.49803	0.57088	164.94100	4.47700	6.45800
h2s	30	5.94177	2.12688	178.25300	2.99600	10.19900
lactic	30	1.44200	0.30349	43.26000	0.86000	2.01000

Pearson Correlation Coefficients, N = 30				
	taste	acetic	h2s	lactic
taste	1.00000	0.54954	0.75575	0.70424
acetic	0.54954	1.00000	0.61796	0.60378
h2s	0.75575	0.61796	1.00000	0.64481
lactic	0.70424	0.60378	0.64481	1.00000

From the scatter plot and correlation matrix, we can infer that taste has high correlation with h2s and lactic and also a good correlation with acetic. We can also infer that there exists high correlation among the predictor variables so there is a possibility of multicollinearity with the predictor variables.

Question 2

Initially, to analyse the predictors' relationship with the response variable, regression analysis is done for taste using all the predictor variables: acetic, h2s and lactic.

Modelling with all three predictor variables
Group 3

The REG Procedure
Model: MODEL1
Dependent Variable: taste

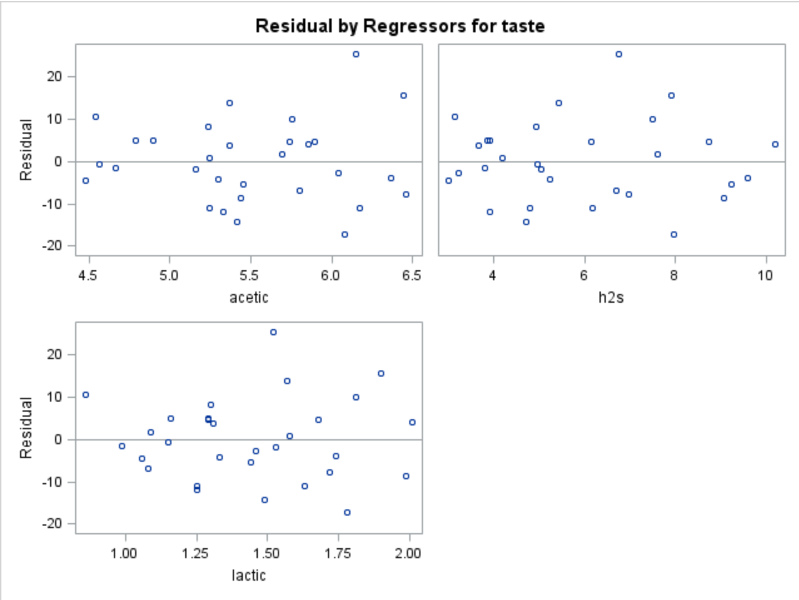
Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4994.47558	1664.82519	16.22	<.0001
Error	26	2668.41109	102.63120		
Corrected Total	29	7662.88667			

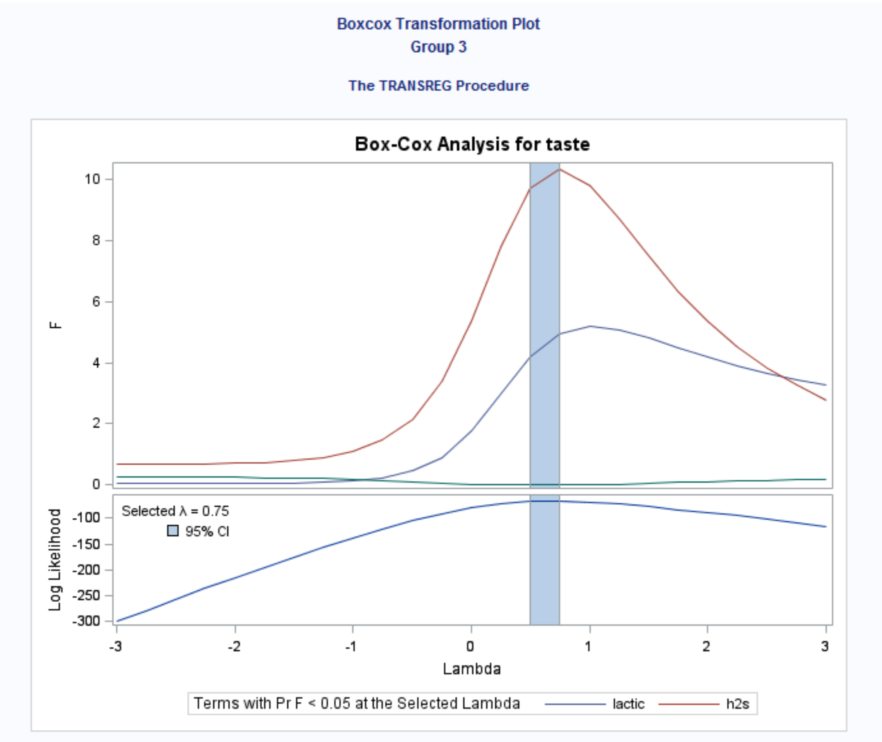
Root MSE	10.13071	R-Square	0.6518
Dependent Mean	24.53333	Adj R-Sq	0.6116
Coeff Var	41.29364		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-28.87677	19.73542	-1.46	0.1554
acetic	1	0.32774	4.45976	0.07	0.9420
h2s	1	3.91184	1.24843	3.13	0.0042
lactic	1	19.67054	8.62905	2.28	0.0311

From the above table, we can see that even though overall p-value is significant, individual p-values are not significant, especially for acetic. This means that there is a chance of multicollinearity among the predictor variables.
When taste is predicted using only one predictor variable, p value was significant for all the variables which implies that all variables have a linear relationship with taste.
Also, from the individual scatter plots for taste with each of the predictor variables(Part 1 Question 1), we know that all predictors have a linear relationship with taste. So there is no need to transform the predictor variables initially.



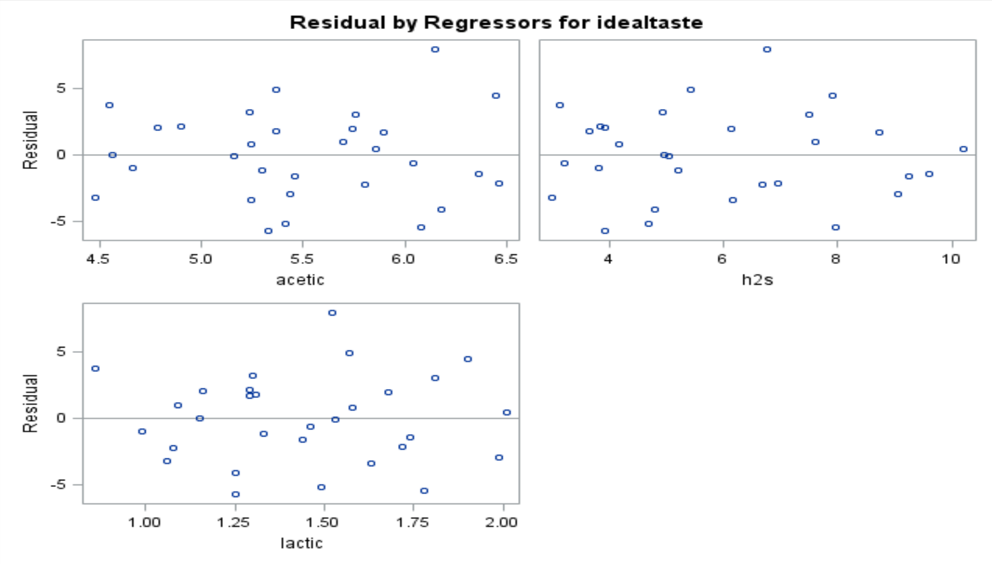
From the residual plots above, we can observe that acetic acid doesn't have constant variance. So, the response variable taste should be transformed using box-cox transformation.



From the above graph, an optimal λ = 0.75 is obtained. A convenient transformation of λ = 0.50 can also be done. So the transformations of $Y' = Y^{0.75}$ and $Y' = Y^{0.50}$ were done to solve the problem of constant variance.

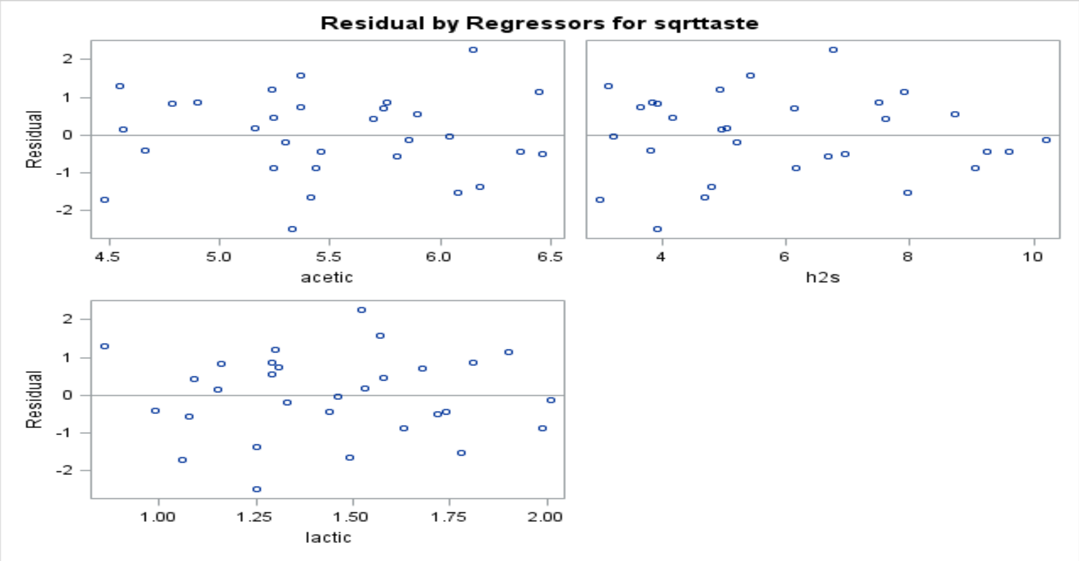
Regression analysis after transformation to 0.75 (Ideal Transformation) Group 3

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-7.35556	6.77557	-1.09	0.2876
acetic	1	0.03189	1.53112	0.02	0.9835
h2s	1	1.37864	0.42861	3.22	0.0035
lactic	1	6.59880	2.96253	2.23	0.0348



Regression analysis after transformation to Sqrt(taste) Group 3

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-0.92376	2.25627	-0.41	0.6856
acetic	1	-0.00049785	0.50986	-0.00	0.9992
h2s	1	0.44499	0.14273	3.12	0.0044
lactic	1	2.01849	0.98652	2.05	0.0510



From the residual plots above, we can see that the constant variance assumption holds after both the transformations for all the predictors.

From the tables above, we can still see insignificant p value for acetic which confirms that it has multicollinearity issues even after transformation. This can be concluded from different SS1 and SS2 values for acetic in the table for Part 1 Question 3. So a general linear test is carried out to determine whether it is possible to eliminate the variable acetic from the model.

Testing models - lactic and h2s only
Group 3

The REG Procedure
Model: MODEL1

Test lactich2sonly Results for Dependent Variable taste				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.55427	0.01	0.9420
Denominator	26	102.63120		

Let us state the hypotheses for the test:
Null hypothesis H_0: $\beta_{acetic}=0$
Alternate hypothesis H_a: $\beta_{acetic} \neq 0$
Test statistic $F_{1,26} = 0.01$
The p-value is 0.94200> alpha (0.05).
Thus, we fail to reject the null hypothesis that $\beta_{acetic}=0$ for $\alpha=0.05$(95% confidence level).
So it is possible to eliminate acetic from the model.

When acetic is eliminated from the model, there is no need to transform the response variable taste since transformation was done to remove non constant variance of acetic variable.

So we can conclude that only two predictors h2s and lactic are required in the model and there is no need of transformation for either the response or the explanatory variables.

Question 3 :

The Cp values for different subsets of variables are shown below:

Selection of best model using cp criterion
Group 3

The REG Procedure
Model: MODEL1
Dependent Variable: taste

C(p) Selection Method

Number of Observations Read	30
Number of Observations Used	30

Number in Model	C(p)	R-Square	Parameter Estimates			
			Intercept	acetic	h2s	lactic
2	2.0054	0.6517	-27.59182	.	3.94627	19.88720
3	4.0000	0.6518	-28.87677	0.32774	3.91184	19.67054
1	6.0189	0.5712	-9.78684	.	5.77609	.
2	7.1964	0.5822	-26.93973	3.80120	5.14560	.
1	11.6346	0.4959	-29.85883	.	.	37.71995
2	11.8182	0.5203	-51.36603	5.57139	.	31.39229
1	26.1162	0.3020	-61.49861	15.64777	.	.

Out of the different subsets of variables, only two models satisfy the criterion $C_p \leq p$:

- h2s and lactic-
  Number in model=2. p=3. $C_p = 2.0054 \leq 3$
- acetic, h2s and lactic-
  Number in model=3. p=4. $C_p = 4.0000 \leq 4$

Now, for deciding between these two models, we compare the C_p values and observe that the first model has a much lower value compared to the second. Also, the R^2 value does not show any significant improvement when moving from the first model to the second. Also, the first model contains only two predictor variables which is a simpler model. For these reasons, we conclude that the model having only h2s and lactic is the best based on C_p criterion.

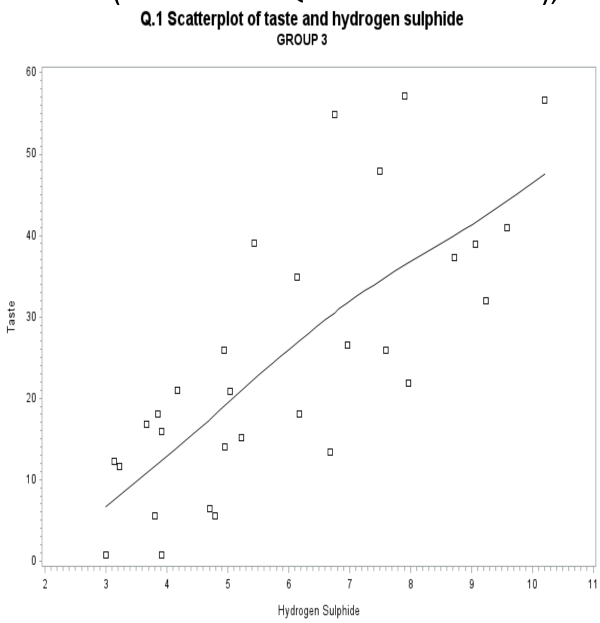
Question 4 :

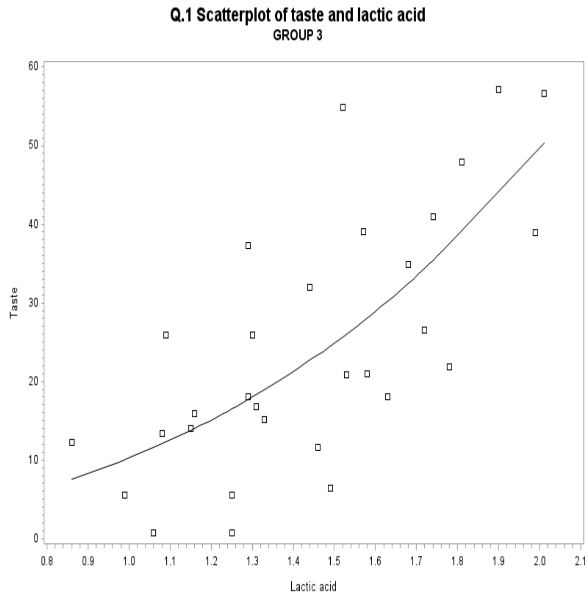
Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	h2s		1	0.5712	0.5712	6.0189	37.29	<.0001
2	lactic		2	0.0805	0.6517	2.0054	6.24	0.0188

From stepwise model selection method, we got the model with only two predictors, h2s and lactic, as the best model. This result confirms the result of Cp criterion model selection method and so we can conclude that it is indeed the best model.

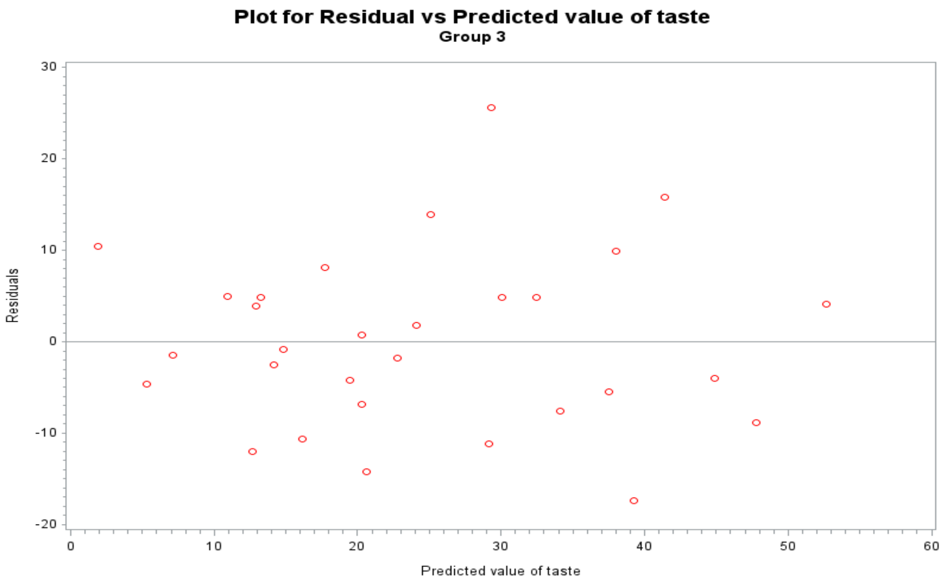
Question 5:

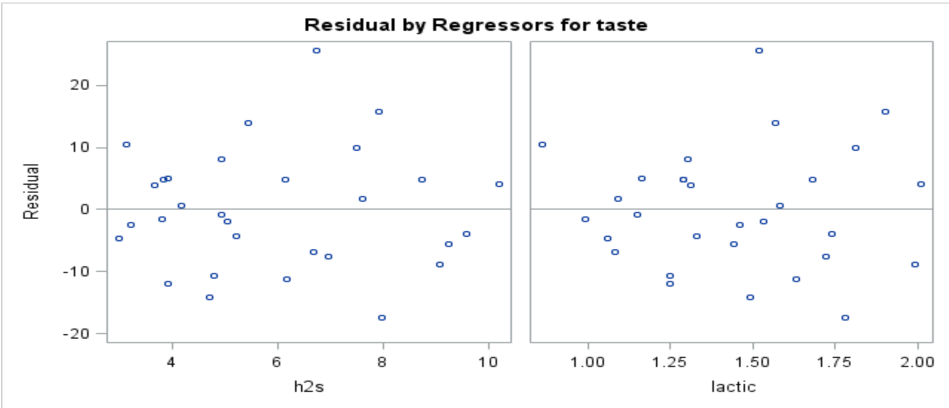
From the individual scatter plots for the response variable taste and the predictors lactic and H2S (as shown in Question 1 of Part 1), linearity assumption seems to hold.



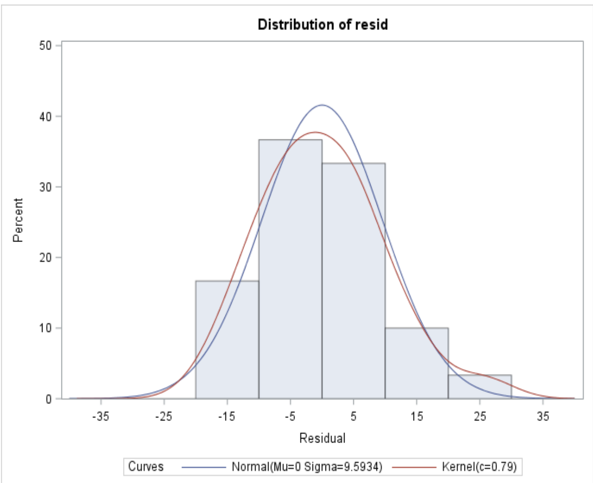


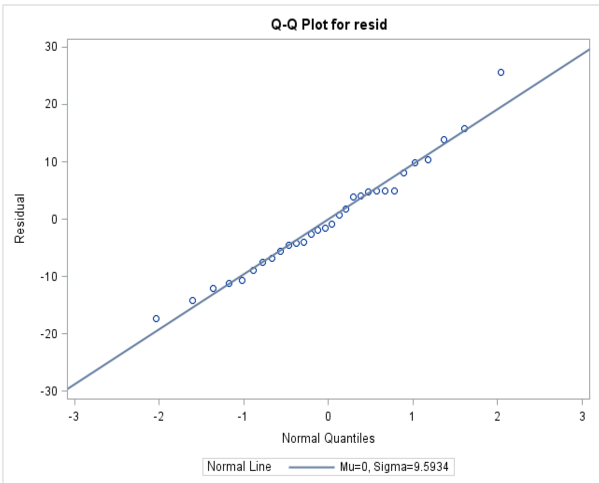
From the residual plots, we observed that the assumption of constant variance seems to hold since the residuals do not follow any pattern and no outlier is present.



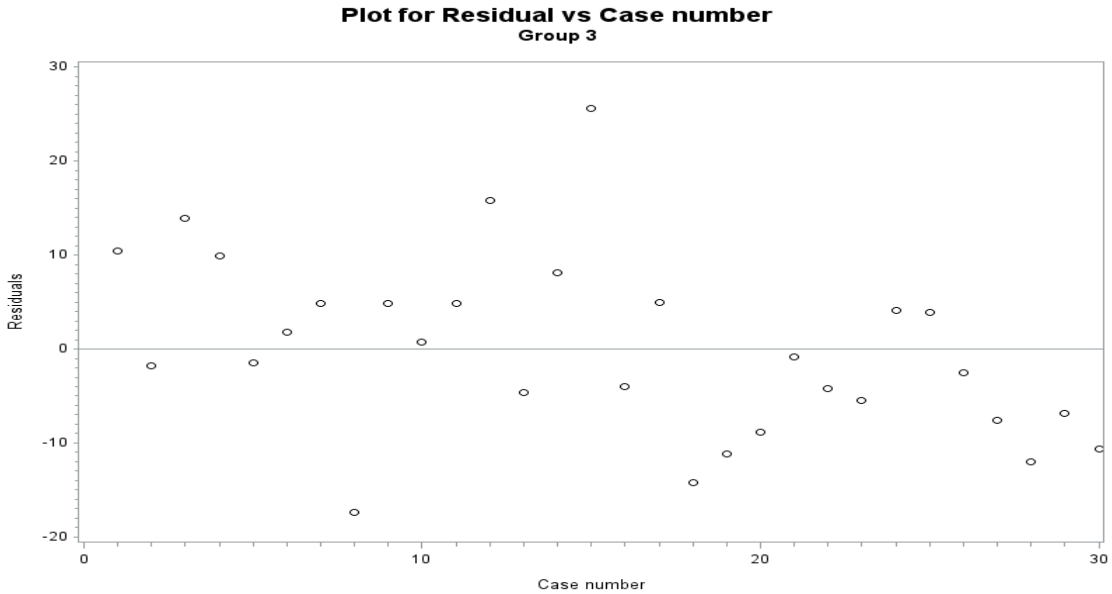


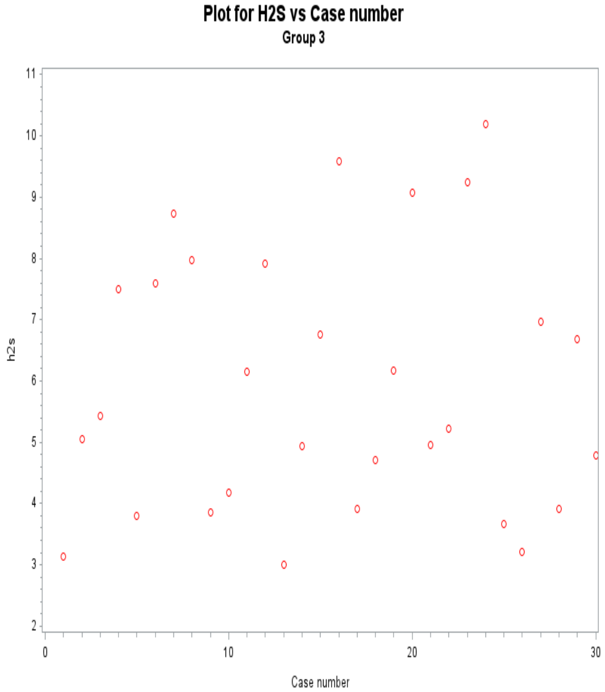
Based on the QQ-Plot, we can say that the points fit the line well. The histogram also appears to show a normal distribution. Thus, the normality assumption seems to hold.

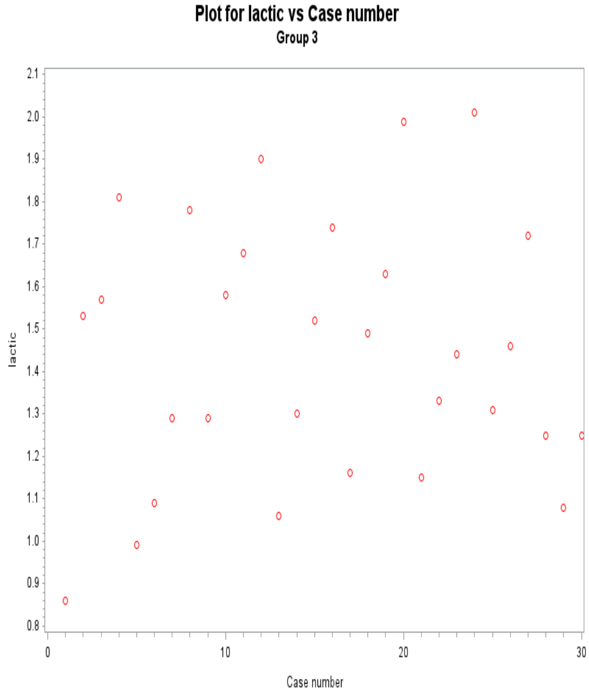




Sequence plots also revealed that the residuals are independent and predictor variables also does not depend on the sequence of observations.





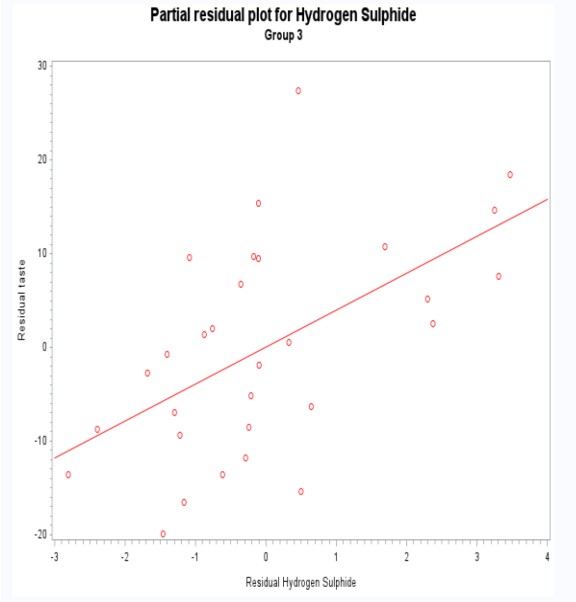


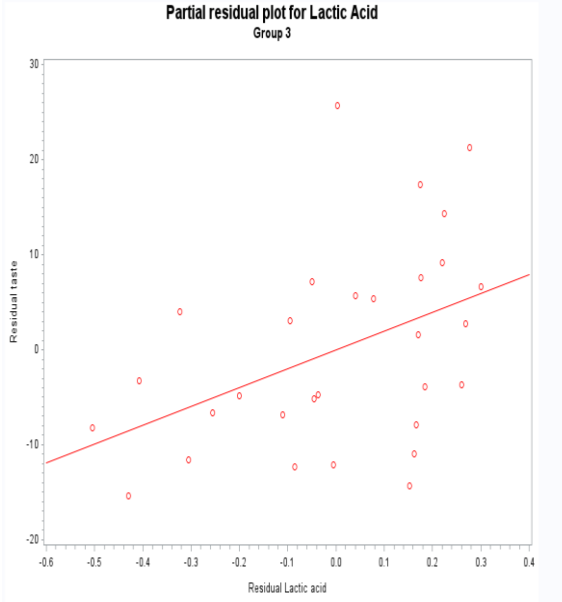
Thus the assumptions of linearity, constant variance, normality and independence seem to hold for the best model i.e model with only h2s and lactic as predictors.

Question 6:

Checking partial residual plots

From the partial residual plots for h2s and lactic, we can see both plots show a linear pattern and, so we can conclude that both h2s and lactic are significant factors in predicting taste.





Checking for multicollinearity using vif and tolerance

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-27.59182	8.98183	-3.07	0.0048	.	0
h2s	1	3.94627	1.13569	3.47	0.0017	0.58422	1.71169
lactic	1	19.88720	7.95901	2.50	0.0188	0.58422	1.71169

Both the VIF values are less than 10 and both tolerance values are greater than 0.1. So there does not exist any multicollinearity issue between h2s and lactic.

Output Statistics												DFBETAS		
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	h2s	lactic
1	12.3	1.8827	3.9838	10.4173	9.109	1.144	0.083	1.1504	0.1606	1.1495	0.5031	0.4812	-0.0254	-0.3254
2	5.6	7.1200	3.2965	-1.5200	9.380	-0.162	0.001	-0.1591	0.1099	1.2545	-0.0559	-0.0516	0.0018	0.0345
3	0.7	5.3116	3.2579	-4.6116	9.393	-0.491	0.010	-0.4839	0.1074	1.2213	-0.1678	-0.1306	0.0714	0.0455
4	13.4	20.2672	3.9300	-6.8672	9.133	-0.752	0.035	-0.7457	0.1562	1.2456	-0.3209	-0.2216	-0.2206	0.2797
5	25.9	24.0808	4.6371	1.8192	8.795	0.207	0.004	0.2031	0.2175	1.4244	0.1071	0.0630	0.0853	-0.0928
6	14.0	14.8085	2.5667	-0.8085	9.605	-0.084	0.000	-0.0826	0.0666	1.1989	-0.0221	-0.0183	-0.0032	0.0137
7	15.9	10.9151	2.6409	4.9849	9.585	0.520	0.007	0.5129	0.0706	1.1690	0.1413	0.1060	-0.0459	-0.0406
8	0.7	12.7050	2.5301	-12.0050	9.615	-1.249	0.036	-1.2622	0.0648	1.0018	-0.3321	-0.1853	0.1733	0.0055
9	5.5	16.1580	2.1830	-10.6580	9.700	-1.099	0.020	-1.1032	0.0482	1.0257	-0.2483	-0.1688	0.0371	0.0776
10	18.1	13.2558	2.5876	4.8442	9.600	0.505	0.006	0.4975	0.0677	1.1676	0.1341	0.0598	-0.0827	0.0167

<u>Checking for outliers using:</u>

1. Studentized residual t-test

The critical value for the t test = ±3

Since none of the student residual values (test statistic) are greater than 3 or less than -3, we can conclude there is no outlier.

2. Studentized deleted residual

The critical value for the test=$t_{n-p-1,\alpha/2n}$=$t_{30-3-1,0.05/60}$=$t_{26,0.00083}$=3.4766

Since none of the studentized deleted residuals (Rstudent) values are outside the range of [-3.4766,3.4766], we can conclude there does not exist any outlier.

<u>Checking for influential observations using:</u>

1. Cook's distance

The critical F value for the test=$F_{p,\ n-p}$ (0.5) = $F_{3,27}$(0.5)=0.8089

Since none of the cook's d values are greater than 0.89, there exist no influential observation.

2. Hat matrix diagonals

The critical value for the test=2p/n=6/30=0.2

Since 5th observation has a value of 0.2175, it might be an influential observation. We can check for DFFITS and DFBETAS values to confirm it.

3. DFFITS

Since our dataset is small (30 observations), cut off value for test=±1

Since none of the DFFITS values are outside the range, there exists no influential observation.

4. DFBETAS

Since our dataset is small (30 observations), cut off value for test=±1

Since none of the DFBETAS values are outside the range, there exists no influential observation.

Hat matrix diagonal values suggest 5th observation as an influential observation but other tests do not suggest the same, so it is affordable to ignore it as an influential observation.

Question 7 :

(a) The equation for the regression model is:
taste = -27.59182+3.94627*h2s+19.88720*lactic

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	-27.59182	8.98183	-3.07	0.0048	-42.89045	-12.29318
h2s	1	3.94627	1.13569	3.47	0.0017	2.01186	5.88068
lactic	1	19.88720	7.95901	2.50	0.0188	6.33072	33.44369

(b) &(c)
90% confidence interval for mean of response variable (**90% CL Mean** columns) and 90% prediction interval for individual observations (**90% CL Predict** columns) are shown in the table.

Estimate of 90% confidence interval for Taste
Group 3

The REG Procedure
Model: MODEL1
Dependent Variable: taste

Output Statistics										
Obs	h2s	lactic	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	3.14	0.86	12.3000	1.8827	3.9838	-4.9028	8.6683	-16.3609	20.1263	10.4173
2	3.81	0.99	5.6000	7.1200	3.2965	1.5051	12.7348	-10.7213	24.9612	-1.5200
3	3.00	1.06	0.7000	5.3116	3.2579	-0.2374	10.8607	-12.5090	23.1323	-4.6116
4	6.69	1.08	13.4000	20.2672	3.9300	13.5732	26.9611	2.0575	38.4769	-6.8672
5	7.60	1.09	25.9000	24.0808	4.6371	16.1825	31.9792	5.3948	42.7669	1.8192
6	4.95	1.15	14.0000	14.8085	2.5667	10.4366	19.1804	-2.6814	32.2985	-0.8085
7	3.91	1.16	15.9000	10.9151	2.6409	6.4170	15.4133	-6.6068	28.4371	4.9849
8	3.91	1.25	0.7000	12.7050	2.5301	8.3956	17.0144	-4.7694	30.1794	-12.0050
9	4.79	1.25	5.5000	16.1580	2.1830	12.4396	19.8763	-1.1801	33.4961	-10.6580
10	3.85	1.29	18.1000	13.2558	2.5876	8.8484	17.6633	-4.2431	30.7547	4.8442
11	8.73	1.29	37.3000	32.4978	4.4374	24.9397	40.0559	13.9530	51.0426	4.8022
12	4.94	1.30	25.9000	17.7640	2.0510	14.2705	21.2575	0.4727	35.0553	8.1360
13	3.66	1.31	16.8000	12.9195	2.7542	8.2283	17.6108	-4.6529	30.4920	3.8805
14	5.22	1.33	15.2000	19.4577	1.9543	16.1290	22.7864	2.1989	36.7164	-4.2577
15	9.24	1.44	32.0000	37.5172	4.1737	30.4081	44.6262	19.1508	55.8835	-5.5172
16	3.22	1.46	11.6000	14.1465	3.6672	7.9002	20.3929	-3.9034	32.1965	-2.5465
17	4.70	1.49	6.4000	20.5876	2.4748	16.3723	24.8029	3.1361	38.0390	-14.1876
18	6.75	1.52	54.9000	29.2819	1.9469	25.9658	32.5981	12.0256	46.5383	25.6181
19	5.04	1.53	20.9000	22.7366	2.3978	18.6525	26.8208	5.3164	40.1569	-1.8366
20	5.44	1.57	39.0000	25.0909	2.3263	21.1285	29.0533	7.6988	42.4830	13.9091
21	4.17	1.58	21.0000	20.3017	3.3728	14.5569	26.0465	2.4191	38.1843	0.6983
22	6.17	1.63	18.0000	29.1886	2.2571	25.3441	33.0331	11.8230	46.5542	-11.1886
23	6.14	1.68	34.9000	30.0567	2.5257	25.7546	34.3587	12.5840	47.5293	4.8433
24	6.96	1.72	26.5000	34.0881	2.4954	29.8377	38.3385	16.6281	51.5480	-7.5881
25	9.59	1.74	40.9000	44.8487	3.6609	38.6132	51.0843	26.8025	62.8950	-3.9487
26	7.97	1.78	21.9000	39.2434	2.8002	34.4738	44.0130	21.6498	56.8369	-17.3434
27	7.50	1.81	47.9000	37.9852	2.8848	33.0716	42.8989	20.3521	55.6184	9.9148
28	7.91	1.90	57.2000	41.4010	3.3274	35.7334	47.0685	23.5430	59.2589	15.7990
29	9.06	1.99	38.9000	47.7527	3.8661	41.1676	54.3378	29.5827	65.9227	-8.8527
30	10.20	2.01	56.7000	52.6294	4.3498	45.2205	60.0384	34.1449	71.1139	4.0706

(d)
From table for problem 7(a),

90% confidence interval for intercept = [-42.89045,-12.29318]
90% confidence interval for regression coefficient of h2s = [2.01186, 5.88068]
90% confidence interval for regression coefficient of lactic = [6.33072, 33.44369]

SAS code

```
DATA cheese;
INPUT  case taste acetic h2s lactic;
CARDS;
1      12.3   4.543 3.135 0.86
2      20.9   5.159 5.043 1.53
3      39           5.366 5.438 1.57
4      47.9   5.759 7.496 1.81
5      5.6          4.663 3.807 0.99
6      25.9   5.697 7.601 1.09
7      37.3   5.892 8.726 1.29
8      21.9   6.078 7.966 1.78
9      18.1   4.898 3.85  1.29
10     21           5.242 4.174 1.58
11     34.9   5.74  6.142 1.68
12     57.2   6.446 7.908 1.9
13     0.7          4.477 2.996 1.06
14     25.9   5.236 4.942 1.3
15     54.9   6.151 6.752 1.52
16     40.9   6.365 9.588 1.74
17     15.9   4.787 3.912 1.16
18     6.4          5.412 4.7   1.49
19     18           5.247 6.174 1.63
20     38.9   5.438 9.064 1.99
21     14           4.564 4.949 1.15
22     15.2   5.298 5.22  1.33
23     32           5.455 9.242 1.44
24     56.7   5.855 10.199      2.01
25     16.8   5.366 3.664 1.31
26     11.6   6.043 3.219 1.46
27     26.5   6.458 6.962 1.72
28     0.7          5.328 3.912 1.25
29     13.4   5.802 6.685 1.08
30     5.5          6.176 4.787 1.25
PROC PRINT DATA=cheese;
RUN;
*Part I;
****************** Q1**************;
TITLE1 'Q.1 Scatterplot of taste and acetic acid';
TITLE2 'GROUP 3';
SYMBOL1 V=SQUARE I=SM70;
PROC SORT DATA=cheese;
BY acetic;
AXIS1 LABEL=('Acetic acid');
AXIS2 LABEL=(ANGLE=90 'Taste');
PROC GPLOT DATA=cheese;
PLOT taste*acetic / HAXIS=AXIS1 VAXIS=AXIS2;
RUN;
TITLE1 'Q.1 Scatterplot of taste and hydrogen sulphide';
```

```
TITLE2 'GROUP 3';
PROC SORT DATA=cheese;
BY h2s;
AXIS1 LABEL=('Hydrogen Sulphide');
AXIS2 LABEL=(ANGLE=90 'Taste');
PROC GPLOT DATA=cheese;
PLOT taste*h2s / HAXIS=AXIS1 VAXIS=AXIS2;
RUN;
TITLE1 'Q.1 Scatterplot of taste and lactic acid';
TITLE2 'GROUP 3';
PROC SORT DATA=cheese;
BY lactic;
AXIS1 LABEL=('Lactic acid');
AXIS2 LABEL=(ANGLE=90 'Taste');
PROC GPLOT DATA=cheese;
PLOT taste*lactic / HAXIS=AXIS1 VAXIS=AXIS2;
RUN;
DATA piecewise;
SET cheese;
IF lactic LE 1.45
THEN CSLOPE=0;
IF lactic GT 1.45
THEN CSLOPE=lactic-1.45;
RUN;
PROC REG DATA=piecewise;
MODEL taste=lactic CSLOPE;
OUTPUT OUT=pieceout P=tastehat;
RUN;
TITLE1 'Q.1 Piecewise SLR for taste and lactic acid';
TITLE2 'GROUP 3';
SYMBOL1 V=CIRCLE I=NONE C=BLACK;
SYMBOL2 V=NONE I=JOIN C=BLACK;
PROC SORT DATA=pieceout;
BY lactic;
AXIS1 LABEL=('lactic ACID');
AXIS2 LABEL=(ANGLE=90 'taste');
PROC GPLOT DATA=pieceout;
PLOT (taste tastehat)*lactic/OVERLAY HAXIS=AXIS1 VAXIS=AXIS2;
RUN;
TITLE1 'Q.1 Test to determine if lines are same';
TITLE2 'GROUP 3';
PROC REG DATA=piecewise;
MODEL taste = lactic CSLOPE;
SAMELINE: TEST CSLOPE;
RUN;
****************** Q2**************;
DATA cheese;
SET cheese;
SUM=h2s + lactic;
TITLE1 'Q.2a i. Predicting response using all explanatory variables';
TITLE2 'GROUP 3';
PROC REG DATA=cheese;
MODEL taste=acetic;
RUN;
TITLE1 'Q.2a ii. Predicting response using all explanatory variables and
SUM';
TITLE2 'GROUP 3';
PROC REG DATA=cheese;
MODEL taste=SUM acetic;
RUN;
TITLE1 'Q.2b Test statistic using test statement in proc reg';
```

```
TITLE2 'GROUP 3';
PROC REG DATA=cheese;
MODEL taste=SUM acetic;
TEST SUM;
RUN;
****************** Q3**************;
TITLE1 'Q.3 TYPE I AND TYPE II SS-all predictors except SUM';
TITLE2 'GROUP 3';
PROC REG DATA=cheese;
MODEL taste=acetic h2s lactic /SS1 SS2;
RUN;
****************** Q4**************;
TITLE1 'Q.4 COMPARISON OF R2 AND ADJUSTED R2 VALUES';
TITLE2 'GROUP 3';
PROC REG DATA=cheese;
MODEL taste=acetic;
MODEL taste=h2s;
MODEL taste=lactic;
MODEL taste=h2s lactic;
MODEL taste=acetic lactic;
MODEL taste=h2s acetic;
MODEL taste=SUM;
MODEL taste=SUM acetic;
MODEL taste=SUM h2s;
MODEL taste=SUM lactic;
MODEL taste=acetic h2s lactic;
RUN;

*Part II;
****************** Q1**************;
TITLE1 'Scatterplot matrix for cheddar cheese data';
TITLE2 'Group 3';
PROC SGSCATTER DATA=cheese;
MATRIX taste acetic h2s lactic;
RUN;
TITLE1 "Correlation Matrix for cheddar cheese data";
TITLE2 'Group 3';
PROC CORR DATA=cheese NOPROB;
VAR taste acetic h2s lactic;
RUN;
****************** Q2**************;
PROC REG DATA=cheese;
TITLE1 'Modelling with all three predictor variables';
TITLE2 'Group 3';
MODEL taste=acetic h2s lactic;
RUN;
PROC TRANSREG DATA = cheese;
TITLE1 'Boxcox Transformation Plot';
TITLE2 'Group 3';
MODEL BOXCOX(taste)=IDENTITY(lactic h2s acetic);
RUN;
DATA box;
SET cheese;
idealtaste=taste**0.75;
sqrttaste=taste**0.5;
RUN;
SYMBOL I=NONE;
PROC REG DATA=box;
TITLE1 'Regression analysis after transformation to 0.75 (Ideal
Transformation)';
TITLE2 'Group 3';
```

```
MODEL idealtaste=acetic h2s lactic;
OUTPUT OUT=resid R=res;
RUN;
SYMBOL I=NONE;
PROC REG DATA=BOX;
TITLE1 'Regression analysis after transformation to Sqrt(taste)';
TITLE2 'Group 3';
MODEL sqrttaste=acetic h2s lactic;
OUTPUT OUT=resid R=res;
RUN;
PROC REG DATA=cheese;
TITLE1 'Testing models - lactic and h2s only';
TITLE2 'Group 3';
MODEL taste=lactic h2s acetic;
lactich2sonly: TEST acetic;
RUN;
****************** Q3**************;
PROC REG DATA=cheese;
TITLE1 'Selection of best model using cp criterion';
TITLE2 'Group 3';
MODEL taste=acetic h2s lactic/SELECTION=CP B;
RUN;
****************** Q4**************;
PROC REG DATA=cheese;
TITLE1 'Selection of best model using stepwise';
TITLE2 'Group 3';
MODEL taste=acetic h2s lactic/SELECTION=STEPWISE B;
RUN;
****************** Q5**************;
PROC REG DATA=cheese;
TITLE1 'Checking of assumptions for best model';
TITLE2 'Group 3';
MODEL taste=h2s lactic;
OUTPUT OUT=check R=resid P=pred;
RUN;
SYMBOL V=CIRCLE I=NONE C=RED;
PROC GPLOT DATA=check;
TITLE1 'Plot for Residual vs Predicted value of taste';
TITLE2 'Group 3';
AXIS1 LABEL=('Predicted value of taste');
AXIS2 LABEL=(ANGLE=90 'Residuals' );
PLOT resid*pred/ HAXIS=AXIS1 VAXIS=AXIS2 VREF=0;
RUN;
PROC GPLOT DATA=check;
TITLE1 'Plot for Residual vs Case number';
TITLE2 'Group 3';
AXIS1 LABEL=('Case number');
AXIS2 LABEL=(ANGLE=90 'Residuals' );
PLOT resid*case/HAXIS=AXIS1 VAXIS=AXIS2 VREF=0;
PROC GPLOT DATA=check;
TITLE1 'Plot for H2S vs Case number';
TITLE2 'Group 3';
AXIS1 LABEL=('Case number');
AXIS2 LABEL=(ANGLE=90 'h2s' );
PLOT h2s*case/HAXIS=AXIS1 VAXIS=AXIS2 VREF=0;
PROC GPLOT DATA=check;
TITLE1 'Plot for lactic vs Case number';
TITLE2 'Group 3';
AXIS1 LABEL=('Case number');
AXIS2 LABEL=(ANGLE=90 'lactic' );
PLOT lactic*case/HAXIS=AXIS1 VAXIS=AXIS2 VREF=0;
```

```
RUN;
PROC UNIVARIATE DATA=check PLOT NORMAL;
VAR resid;
TITLE1 'Histogram plot';
TITLE2 'Group 3';
HISTOGRAM resid / NORMAL KERNEL(L=2);
RUN;
PROC UNIVARIATE DATA=check PLOT NORMAL;
VAR resid;
TITLE1 'QQ plot';
TITLE2 'Group 3';
QQPLOT resid / NORMAL (L=1 MU=EST SIGMA=EST);
RUN;
****************** Q6**************;
PROC REG DATA=cheese;
MODEL taste=h2s lactic/R P INFLUENCE TOL VIF;
RUN;
TITLE1 'Partial residual plot for Hydrogen Sulphide';
TITLE2 'Group 3';
SYMBOL1 V=CIRCLE I=RL;
AXIS1 LABEL=('Residual Hydrogen Sulphide');
AXIS2 LABEL=(ANGLE=90 'Residual taste');
PROC REG DATA=cheese;
MODEL taste h2s = lactic;
OUTPUT OUT=partialh2s R=restaste resh2s;
PROC GPLOT DATA=partialh2s;
PLOT restaste*resh2s / HAXIS=AXIS1 VAXIS=AXIS2;
RUN;
TITLE1 'Partial residual plot for Lactic Acid';
TITLE2 'Group 3';
SYMBOL1 V=CIRCLE I=RL;
AXIS1 LABEL=('Residual Lactic acid');
AXIS2 LABEL=(ANGLE=90 'Residual taste');
PROC REG DATA=cheese;
MODEL taste lactic = h2s;
OUTPUT OUT=partiallactic R=restaste reslactic;
PROC GPLOT DATA=partiallactic;
PLOT restaste*reslactic / HAXIS=AXIS1 VAXIS=AXIS2;
RUN;
****************** Q7**************;
PROC REG DATA=cheese;
TITLE1 'Estimate of 90% confidence interval for Taste';
TITLE2 'Group 3';
MODEL taste=h2s lactic/CLB CLM CLI ALPHA=0.10;
ID h2s lactic;
RUN;
```