Final Project

Eric Altenburg **David Horowitz** **Lachlan Mountjoy**

December 6, 2019

Pledge: I pledge my honor that I have abided by the Stevens Honor System.

1 Report

1.1 Executive Summary

We are a group of students at Stevens Institute of Technology interested in statistics and its real world applications. We are analyzing how the flavor of cheese can be altered by several different chemicals found in the cheese. Our previous research into chemical compounds lead us to the idea that acetic acid, hydrogen sulfide, and lactic acid could be major contributors to the flavor of cheese. Those are the variables we explore in this research. We believe that studying what causes a cheese to taste better could allow us to develop more flavorful and marketable cheeses.

1.2 Data Set

Our data set contains 30 different measurements from a set of cheddar cheeses. The variables being measured result from chemical processes which occur when cheddar cheese matures. The cheese is from the LaTrobe Valley of Victoria in Victoria, Australia. “Taste” is the main response variable being explored which is related to the concentrations of various chemicals in the cheese. The taste values were measured by combining the taste rankings from several different participants. Three explanatory variables were measured and recorded from the cheese: acetic acid, hydrogen sulfide (H2S), and lactic acid. For acetic acid and hydrogen sulfide, logarithmic transformations were taken. Lactic acid did not have a logarithmic transformation. The variable "Case" corresponds to the observation number 1 to 30.

1.3 Software

We used the R programming language combined with the RStudio development environment. We used the base packages included with R for our analysis. We also used LaTeX to create the document.

1.4 Analysis

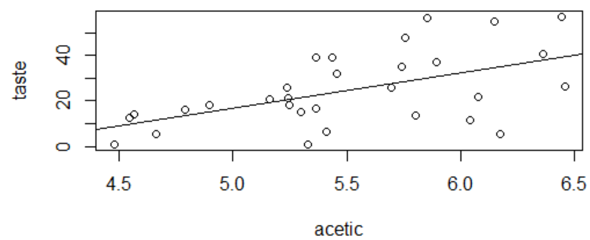
Our analysis is broken down into several components. We began with an overview of the data. This broad analysis shows the general statistics of the data such as the mean, standard deviation, and quartile ranges.

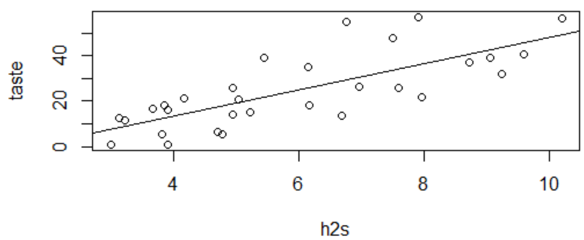
	Mean	Median	St. Dev.	IQR
Taste	24.53	20.93	16.26	23.15
Acetic	5.50	5.42	0.57	0.646
H2S	5.94	5.33	2.13	3.597
Lactic	1.44	1.45	0.3	0.417

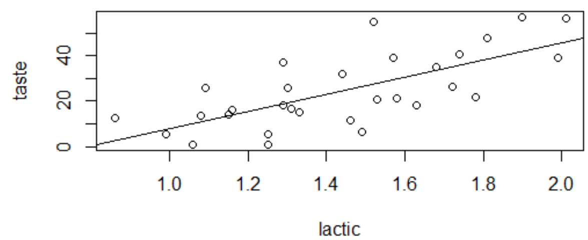
We calculated the correlations between each combination of variables, this gave us a good starting point as to which variables might be related and told us which multiple regressions would be the best.

	Taste	Acetic	H2S	Lactic
Taste	1.0000000	0.5495393	0.7557523	0.7042362
Acetic	0.5495393	1.0000000	0.6179559	0.6037826
H2S	0.7557523	0.6179559	1.0000000	0.6448123
Lactic	0.7042362	0.6037826	0.6448123	1.0000000

We also ran 3 linear regressions on the dataset to find whether or not the explanatory variables are linearly related.







The residuals from each of the linear regressions are close to linear which indicates that our model is a good fit. Afterwards, we computed multiple linear regressions to see if they were better models than the single regressions were. We found that the model that used H2S and Lactic was the best in the end, due to its low P value in comparison to the other models. The remainder of the calculations and results can be found in the second section of the document.

2 Data

11.53

	Mean	Median	St. Dev.	IQR
Taste	24.53	20.93	16.26	23.15
Acetic	5.50	5.42	0.57	0.646
H2S	5.94	5.33	2.13	3.597
Lactic	1.44	1.45	0.3	0.417

Taste STEM: The decimal points 1 digit(s) to the right of the |

```
0  | 11666
1  | 223456788
2  | 112667
3  | 25799
4  | 18
5  | 577
```

Acetic STEM: The decimal point is 1 digit(s) to the left of the |

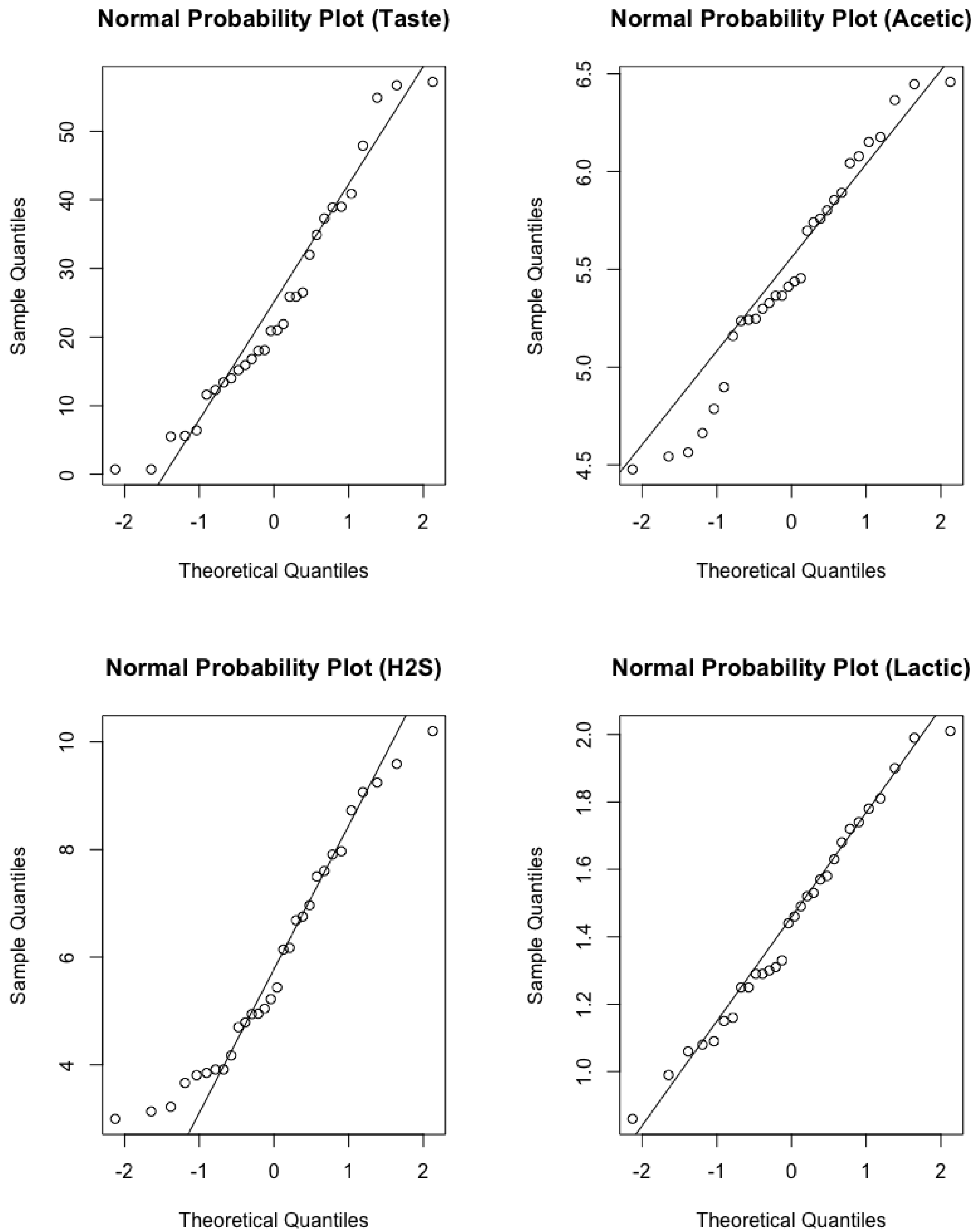
```
44 | 846
46 | 69
48 | 0
50 | 6
52 | 4450377
54 | 146
56 | 046
58 | 069
60 | 4858
62 | 7
64 | 56
```

H2S STEM: The decimal point is at the |

```
2  |
3  | 01278999
4  | 27899
5  | 024
6  | 1728
7  | 0569
8  | 07
9  | 126
10 | 2
```

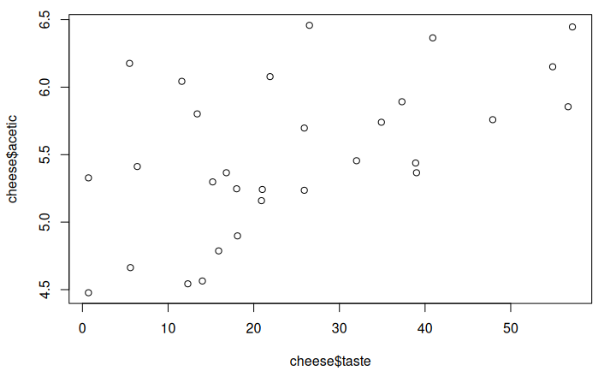
Lactic STEM: The decimal point is 1 digit(s) to the left of the |

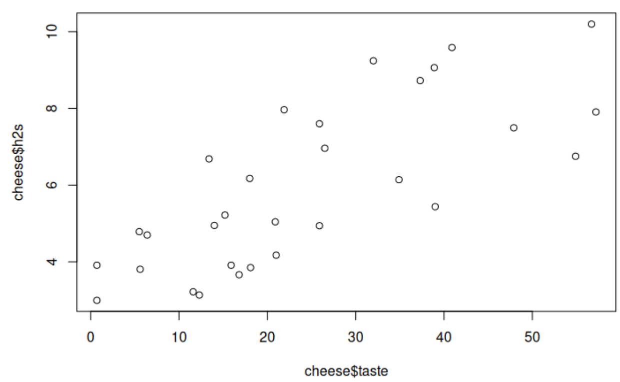
```
8  | 69
10 | 68956
12 | 5599013
14 | 4692378
16 | 38248
18 | 109
20 | 1
```

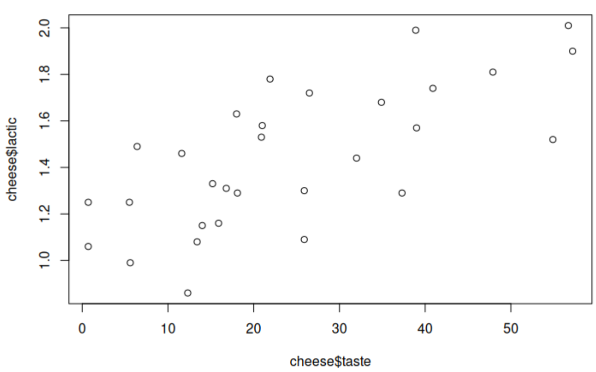


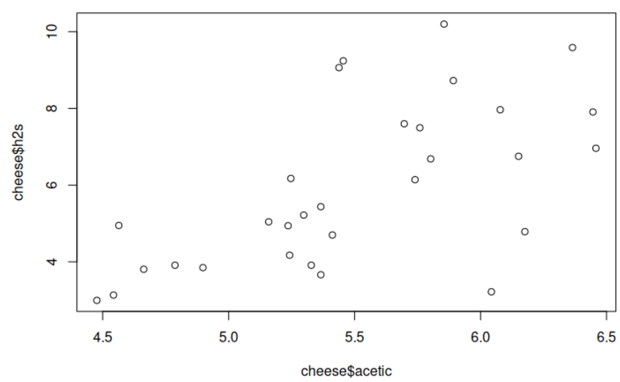
All of the above plots showed normality among all data sites as the plotted observations were close to the line.

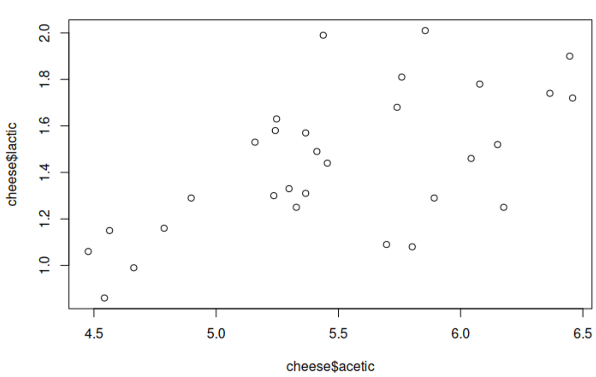
11.54

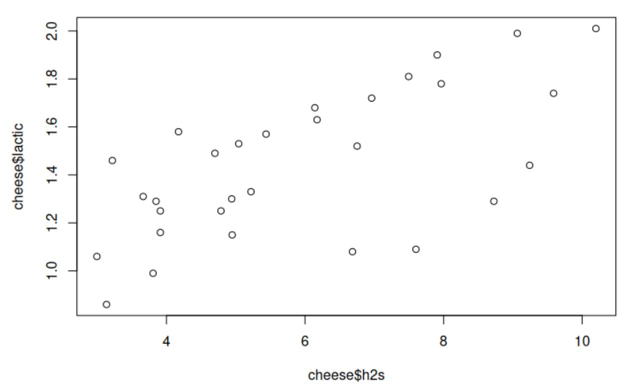












For all the above graphs, the relationships are positive.

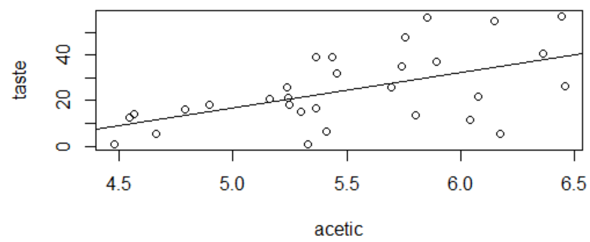
Coorelation

	Taste	Acetic	H2S	Lactic
Taste	1.0000000	0.5495393	0.7557523	0.7042362
Acetic	0.5495393	1.0000000	0.6179559	0.6037826
H2S	0.7557523	0.6179559	1.0000000	0.6448123
Lactic	0.7042362	0.6037826	0.6448123	1.0000000

P-value

Taste, Acetic	0.001658192
Taste, H2S	$1.373783 * 10^{-6}$
Taste, Lactic	$1.405117 * 10^{-5}$
Acetic, H2S	0.0002739173
Acetic, Lactic	0.0004113657
H2S, Lactic	0.0001198401

11.55



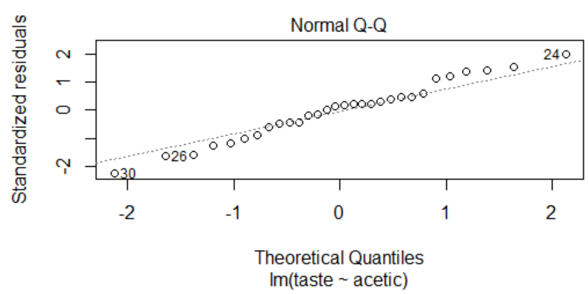
Summary

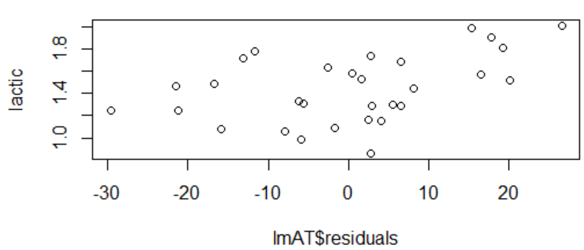
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -61.499     24.846  -2.475  0.01964 *
acetic        15.648      4.496   3.481  0.00166 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.82 on 28 degrees of freedom
Multiple R-squared:  0.302,     Adjusted R-squared:  0.2771
F-statistic: 12.11 on 1 and 28 DF,  p-value: 0.001658
```

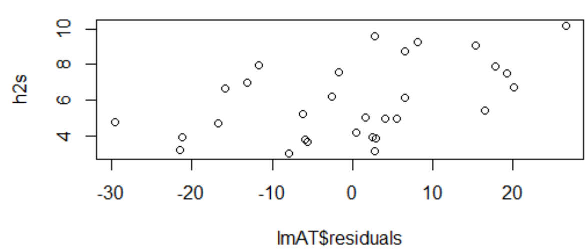
$$\text{H}_0 : \beta_1 = 0\text{, Acetic and Taste are not related}$$

$$\text{H}_a : \beta_1 \neq 0\text{, Acetic and Taste are related}$$

Since the p-value of $0.001658 < 0.05$, we can reject H_0 and say that the model is statistically significant in that acetic and taste variables are related.

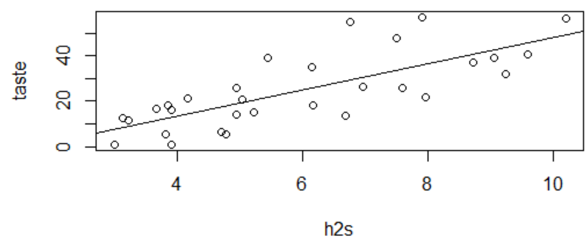






Based on the above three graphs, the residuals seem to be relatively normal, and have some positive association with the other two variables.

11.56



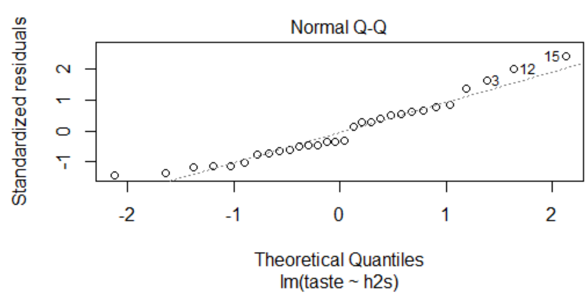
Summary

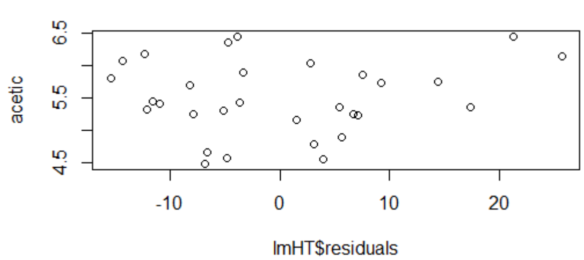
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.7868     5.9579  -1.643    0.112
h2s           5.7761     0.9458   6.107 1.37e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.83 on 28 degrees of freedom
Multiple R-squared:  0.5712,    Adjusted R-squared:  0.5558
F-statistic: 37.29 on 1 and 28 DF,  p-value: 1.374e-06
```

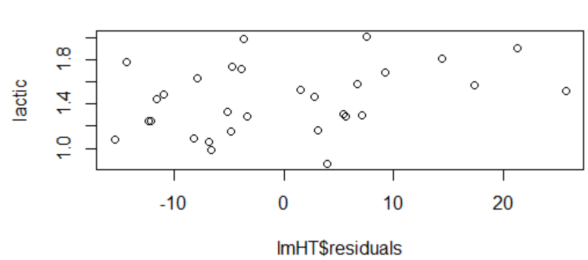
$$\mathrm{H}_0 : \beta_1 = 0, \text{ Taste and H2S are not related}$$

$$\mathrm{H}_a : \beta_1 \neq 0 \text{ Taste and H2S are related}$$

Since the p-value of $1.374 * 10^{-6} < 0.05$, we can reject H_0 and say that the model is statistically significant in that taste and H2S are related.

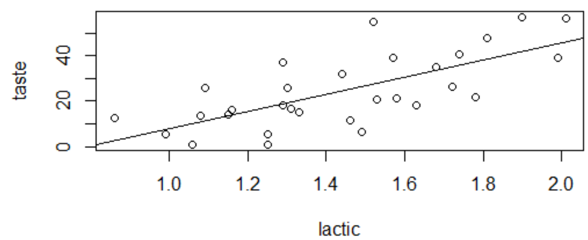






Based on the above three graphs, the residuals seem to be relatively normal, and no noticeable association with the other two variables.

11.57



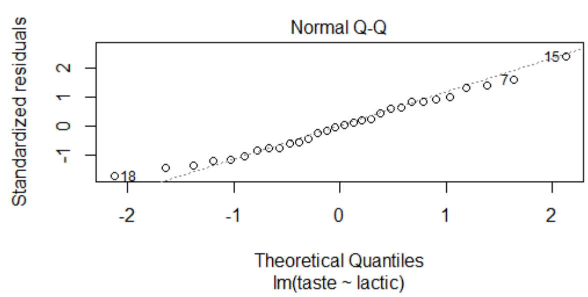
Summary

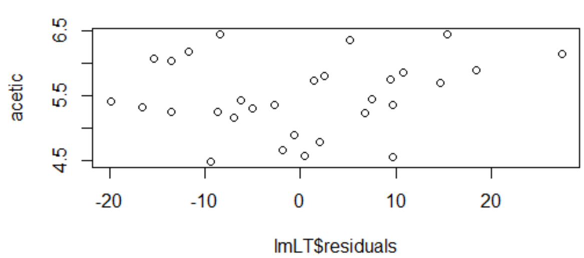
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -29.859     10.582  -2.822  0.00869 ** 
lactic        37.720      7.186   5.249 1.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 28 degrees of freedom
Multiple R-squared:  0.4959,    Adjusted R-squared:  0.4779 
F-statistic: 27.55 on 1 and 28 DF,  p-value: 1.405e-05
```

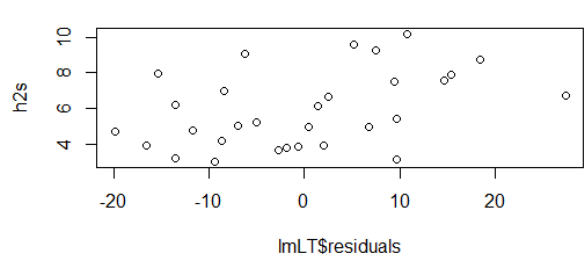
$\mathrm{H}_0 : \beta_1 = 0$, Taste and Lactic are not related

$\mathrm{H}_a : \beta_1 \neq 0$ Taste and Lactic are related

Since the p-value of $1.405 * 10^{-5} < 0.05$, we can reject H_0 and say that the model is statistically significant in that taste and lactic are related.







Based on the above three graphs, the residuals seem to be relatively normal, and no noticeable association with the other two variables.

11.58

Regression Model	F Statistic	P-Value	R^2	S
Acetic	12.11	0.001658	0.2771	13.82
H2S	37.29	$1.37 * 10^{-6}$	0.5558	10.83
Lactic	27.55	$1.41 * 10^{-5}$	0.4779	11.75

$$\hat{taste} = -61.499 + 15.648 * acetic$$

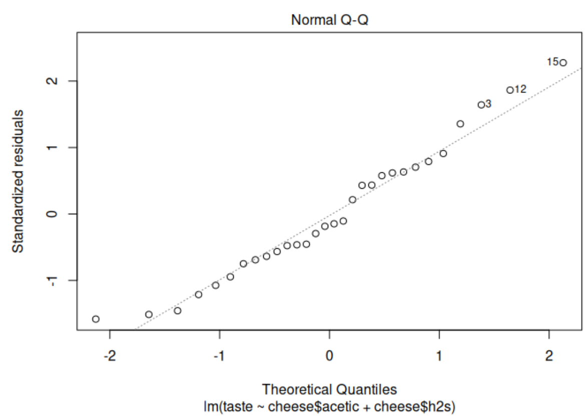
$$\hat{taste} = -9.7868 + 5.7761 * H2S$$

$$\hat{taste} = -29.859 + 37.720 * lactic$$

The above three equations' intercepts are different because they are using three different explanatory variables.

11.59

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -26.940     21.194  -1.271 0.214536
cheese$acetic    3.801      4.505   0.844 0.406245
cheese$h2s       5.146      1.209   4.255 0.000225 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.89 on 27 degrees of freedom
Multiple R-squared:  0.5822,	Adjusted R-squared:  0.5512
F-statistic: 18.81 on 2 and 27 DF,  p-value: 7.645e-06
```



$H_0 : \beta_1 = 0$, Acetic and H2S are not related

$H_a : \beta_1 \neq 0$ Acetic and H2S are related

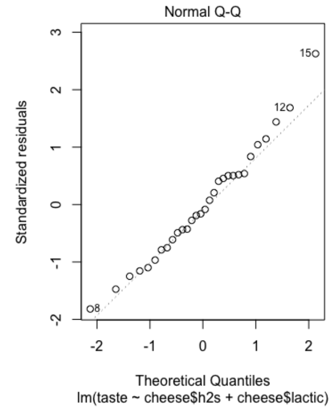
Since the p-value of $7.645 * 10^{-6} < 0.05$, we can reject H_0 and say that the model is statistically significant in that acetic and H2S are related.

In terms of the residual, it appears to be relatively normal.

This model does not seem to be better than the other model containing H2S. Since acetic and H2S are correlated, acetic doesn't add any significance.

11.60

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -27.592      8.982  -3.072  0.00481 **
cheese$h2s       3.946      1.136   3.475  0.00174 **
cheese$lactic   19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared:  0.6517,    Adjusted R-squared:  0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```



$H_0 : \beta_1 = 0$, H2S and Lactic are not related

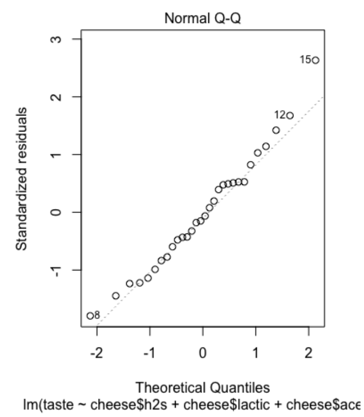
$H_a : \beta_1 \neq 0$ H2S and Lactic are related

Since the p-value of $6.551 * 10^{-7} < 0.05$, we can reject H_0 and say that the model is statistically significant in that H2S and lactic are related.

The residual appear to be relatively normal. The p-value is significantly lower in comparison to the previous two variables alone, therefore, it is more significant and a better model.

11.61

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -28.8768    19.7354  -1.463  0.15540
cheese$h2s      3.9118     1.2484   3.133  0.00425 **
cheese$lactic  19.6705     8.6291   2.280  0.03108 *
cheese$acetic   0.3277     4.4598   0.073  0.94198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,    Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```



$H_0 : \beta_1 = 0$, H2S, Lactic, and Acetic are not related

$H_a : \beta_1 \neq 0$, H2S, Lactic, and Acetic are related

Since the p-value of $3.81 * 10^{-6} < 0.05$, we can reject H_0 and say that the model is statistically significant in that H2S, lactic, and acetic are related.

It would appear that the multiple regression model combining H2S and lactic is the best because it has the smallest p-value and is statistically the most significant.