MA 331 - Intermediate Statistics
Aparajita Rana & Ming Lin
I pledge my honor that I have abided by the Stevens Honor System.

Section 1: Statistical Report

Executive Summary

We are students at Stevens Institute of Technology who are studying Intermediate Statistics. Our goal is to understand how the flavor of cheese can be affected by the concentration of certain chemicals found in it. The taste of matured cheese is related to the concentration of several chemicals in the final product.

Dataset

There are 30 observations in total. In the dataset, the case variable is only used to number the observations. taste is the response variable of interest. These scores were obtained through the combination of the scores of several tasters. The concentration of three chemicals were measured. These chemicals were acetic acid, hydrogen sulfide, and lactic acid. Natural Log transformations were taken on acetic acid and hydrogen sulfide. The explanatory variables are the untransformed concentration of lactic acid (which we call "Lactic") and the transformed concentrations of acetic acid (labelled as "acetic") and hydrogen sulfide (called "H2S").

Software

We utilized R in R Studio to produce our output to perform this statistical analysis

Analysis

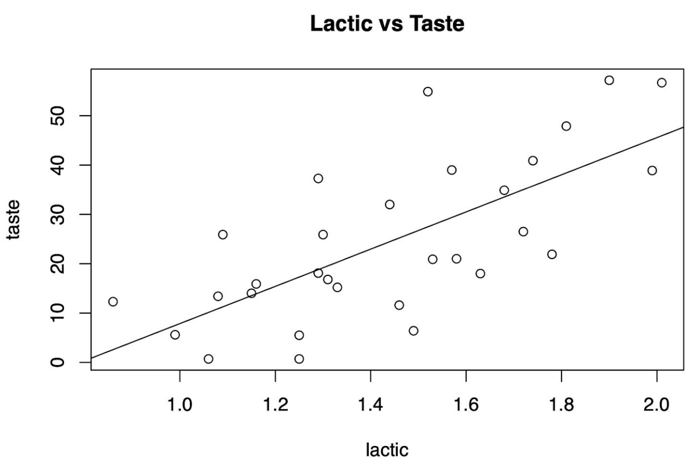
We will begin with a general analysis of the data. In the following table, the mean, median, standard deviation, and interquartile range for each of the variables is shown:

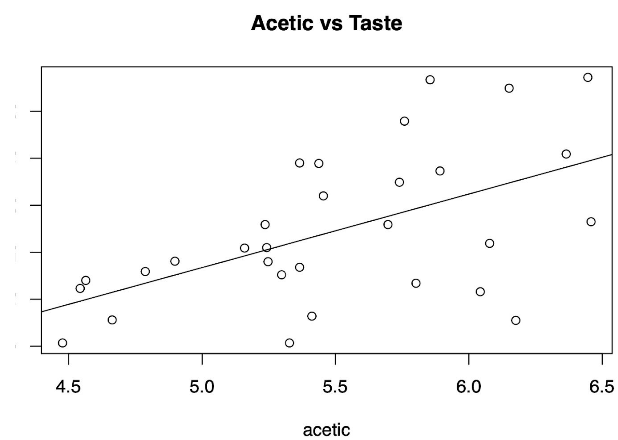
	Mean	Median	Standard Dev	IQR
Taste	24.53	20.95	16.25538	23.9
Acetic	5.498	5.425	0.571	0.656
H2S	5.942	5.329	2.127	3.689
Lactic	1.442	1.450	0.3035	0.43

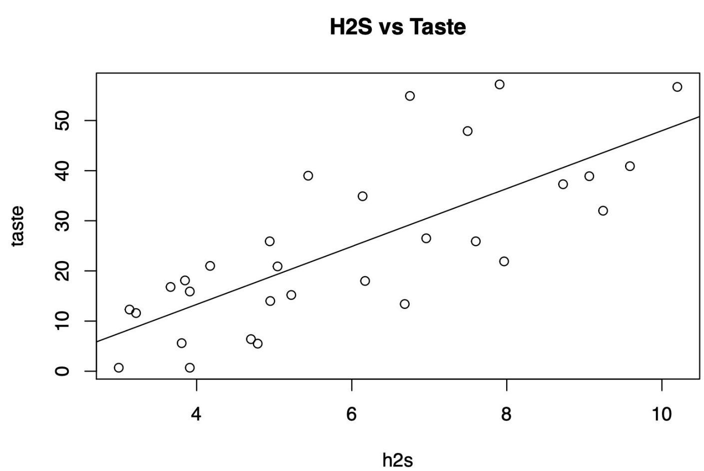
Now we run linear regression to find a correlation between the response variable (Taste) and Acetic, H2S, and Lactic. The linear regression models are as follows:
Taste = −61.6 + 15.6 x Acetic
Taste = −9.8 + 5.8 x H2S
Taste = −29.9 + 37.7 x Lactic

	f-statistic	p-value	R^2	S
Acetic vs Taste	12.11	0.00166	0.302	13.82
H2S vs Taste	37.29	1.374e-6	0.571	10.83
Latic vs Taste	27.55	1.405e-5	0.4959	11.75

Looking at these models, we see that all three chemicals have a positive association with Taste. Acetic has the weakest relationship with Taste with a P-value of 0.00166 and an R2 value of 0.302 while H2S has the strongest relationship with Taste, with a P-value of 1.374e-6 and an R2 value of 0.571. We can see this conclusion through our linear regression plots as well:







Section 2: Study Details

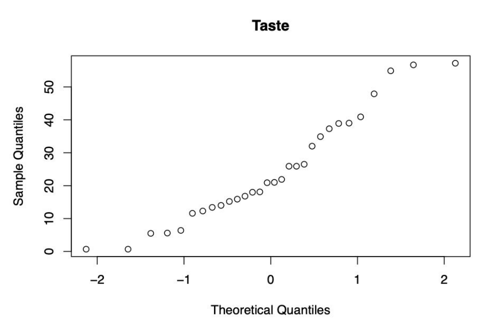
11.53)

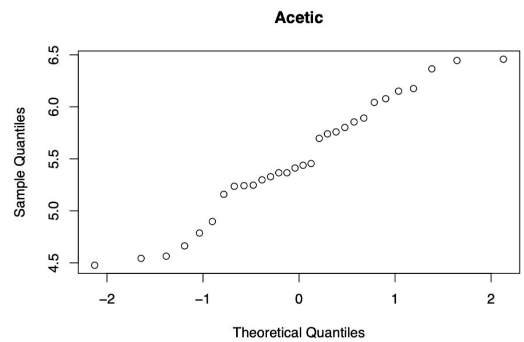
Summaries in the following order: Taste, Acetic, H2S, and Lactic

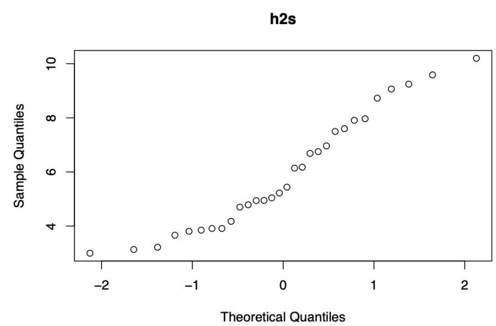
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.70   13.55   20.95   24.53   36.70   57.20

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.477   5.237   5.425   5.498   5.883   6.458

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.996   3.978   5.329   5.942   7.575  10.199

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.860   1.250   1.450   1.442   1.667   2.010
```

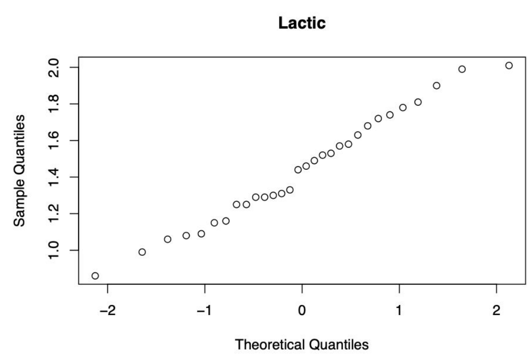
	Mean	Median	Stand Dev	IQR
Taste	24.53	20.95	16.255	23.15
Acetic	5.498	5.425	0.571	0.6453
H2S	5.942	5.329	2.127	3.600
Lactic	1.442	1.450	0.304	0.418

<table>
<tr>
<td>Taste</td>
<td>The decimal point is 1 digit(s) to the right of the |

0 | 11666
1 | 223456788
2 | 112667
3 | 25799
4 | 18
5 | 577</td>
<td>
</td>
</tr>
</table>

<table>
<tr>
<td>Acetic</td>
<td>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</td>
<td>
</td>
</tr>
<tr>
<td>H2S</td>
<td>The decimal point is at the |

2 |
3 | 01278999
4 | 27899
5 | 024
6 | 1278
7 | 0569
8 | 07
9 | 126
10 | 2</td>
<td>
</td>
</tr>
<tr>
<td>Latic</td>
<td>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</td>
<td>
</td>
</tr>
</table>

Observations:

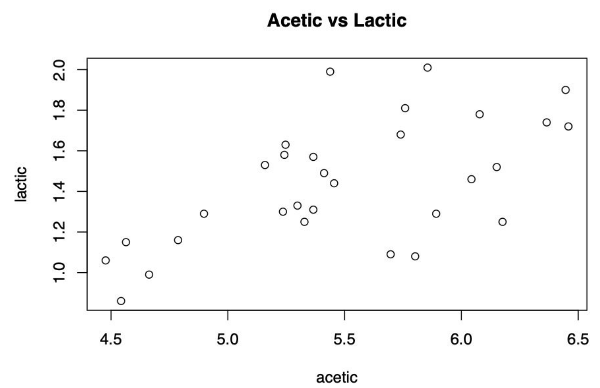
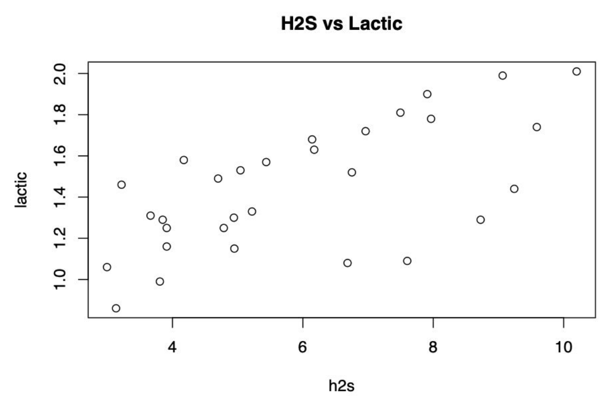
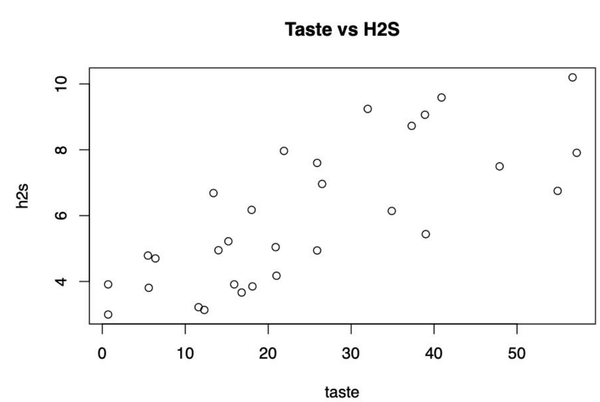
Taste: Relatively normal, skewed right, & no outliers

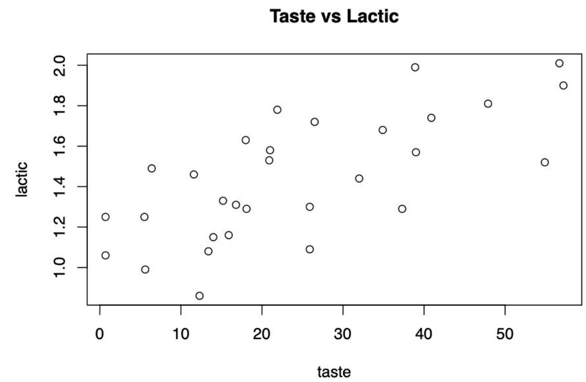
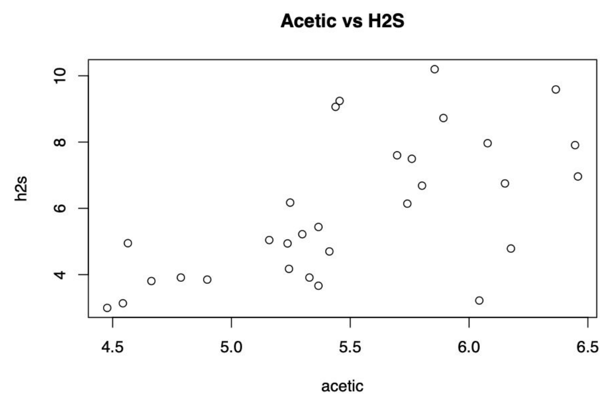
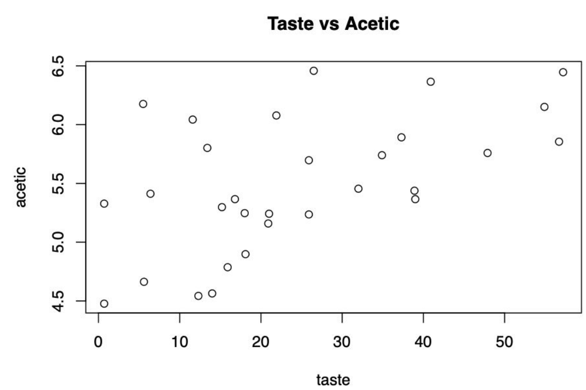
Acetic: Relatively normal, 2 peaks, & no outliers

H2S: Relatively normal, skewed right, & no outliers

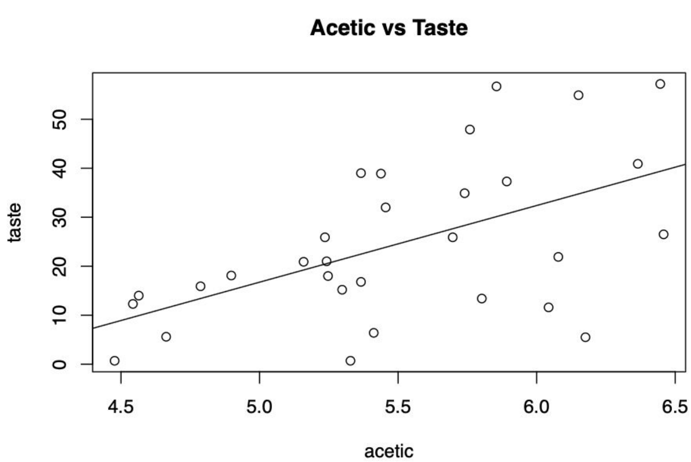
Lactic: Relatively normal & no outliers

11.54)

Acetic vs Lactic: *Correlation: 0.604* There is a moderately strong positive correlation. H0:β =0 H1:$\beta \neq$0 α = .05 t-value = 4.008 P-value = 0.000411 < .05 Reject the null hypothesis.	
H2S vs Lactic: *Correlation: 0.645* There is a moderately strong positive correlation. H0:β =0 H1:$\beta \neq$0 α = .05 t-value = 4.464 P-value = 0.00012 < .05 Reject the null hypothesis.	
Taste vs H2S: *Correlation: 0.7558* There is a strong positive correlation. H0:β =0 H1:$\beta \neq$0 α = .05 t-value = 6.107 P-value = 1.37 ∗ 10-. < .05 Reject the null hypothesis.	

Taste vs Lactic: *Correlation: 0.7042* There is a strong positive correlation. H0:β =0 H1:$\beta \neq$0 α = .05 t-value = 5.249 P-value = 1.41 $*$ 10-0 < .05 Reject the null hypothesis.	
Acetic vs H2S: *Correlation: 0.6179* There is a moderately strong positive correlation. H0:β =0 H1:$\beta \neq$0 α = .05 t-value = 4.159 P-value = 0.000274 < .05 Reject the null hypothesis.	
Taste vs Acetic: *Correlation: 0.5495* There is a moderate positive correlation between the two. H0:β =0 H1:$\beta \neq$0 α = .05 t-value = 3.481 P-value = .00166 < . 05 Reject the null hypothesis.	

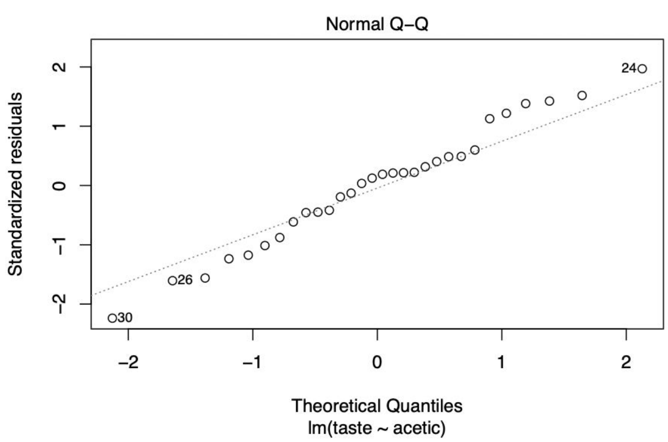
11.55)

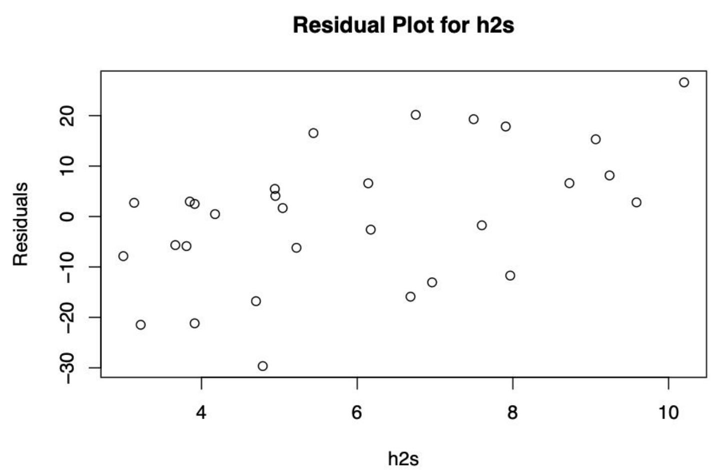


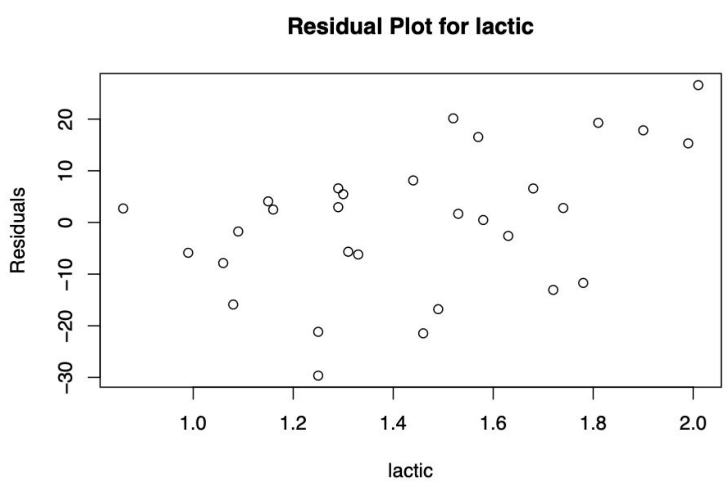
Summary of Linear Regression:

```
Residuals:
    Min      1Q  Median      3Q     Max

-29.642  -7.443   2.082   6.597  26.581

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
```

H0: $\beta 1 = 0$
Ha: $\beta 1 \neq 0$
Linear regression equation: Taste = −61.6 + (15.6 * Acetic)
t-value = 3.48, p-value = 0.002
Because the p-value < $\alpha = 0.05$, we reject H0.
There is statistically significant evidence that $\beta 1 \neq 0$.
Thus, there is evidence of a linear relationship between taste and acetic acid.







Based on the QQ plot, the residuals have a Normal distribution. The residuals seem to be positively associated with H2S (0.498) and Lactic (0.4457).

11.56)

Summary of Linear Regression:

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.426  -7.611  -3.491   6.420  25.687

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
```

H0: $\beta 1 = 0$

Ha: $\beta 1 \neq 0$

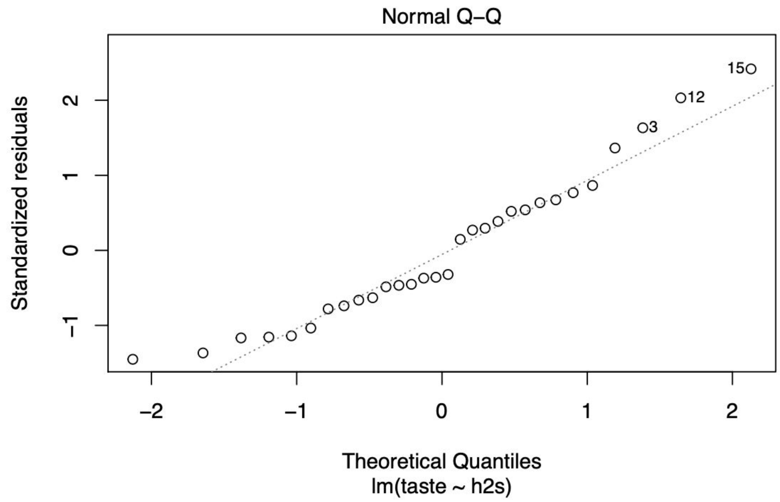
Linear regression equation: Taste = −9.78 + (5.77 * H2S)

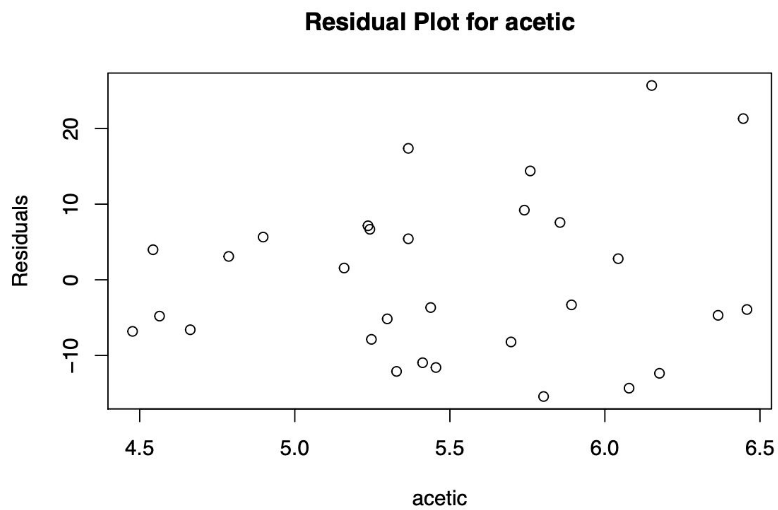
t-value = -1.64, p-value = 1.37 * 10^-6

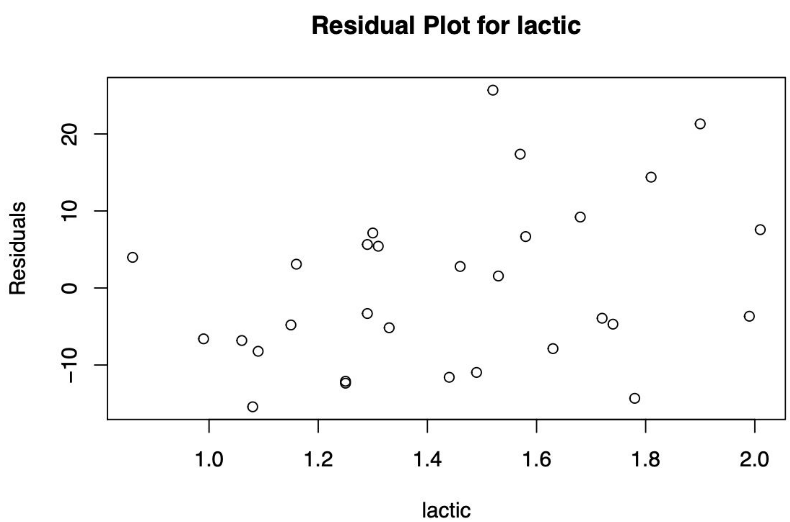
Because the p-value < $\alpha = 0.05$, we reject H0.

There is statistically significant evidence that $\beta 1 \neq 0$.

Thus, there is evidence of a linear relationship between taste and H2S acid.

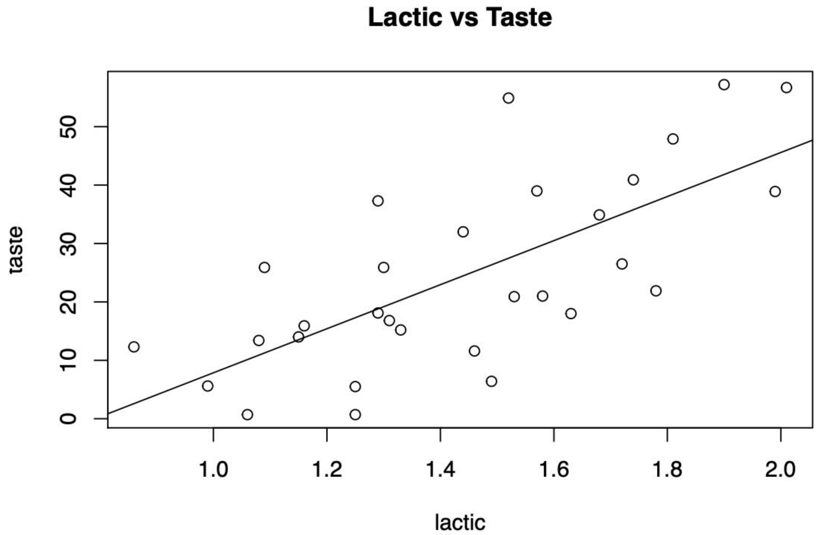






Based on the QQ plot, the residuals have a Normal distribution. The residuals do not seem to have a strong correlation with either Acetic and Lactic.

11.57)




```
Residuals:
     Min       1Q   Median       3Q      Max
-19.9439  -8.6839  -0.1095   8.9998  27.4245

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
```

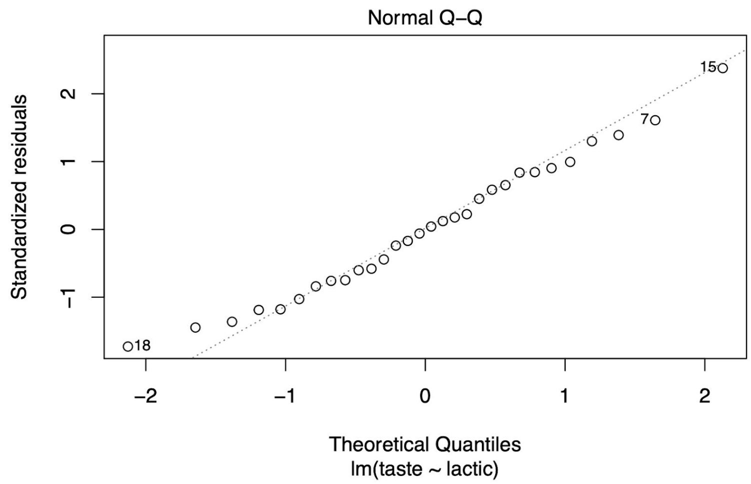
H0: $\beta 1 = 0$

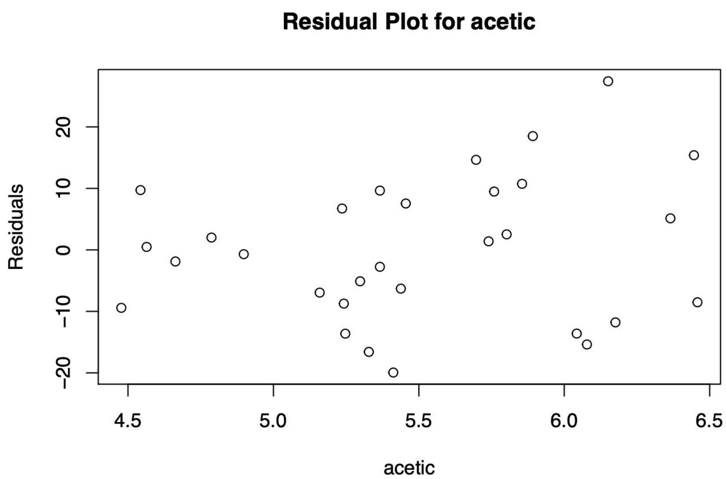
Ha: $\beta 1 \neq 0$

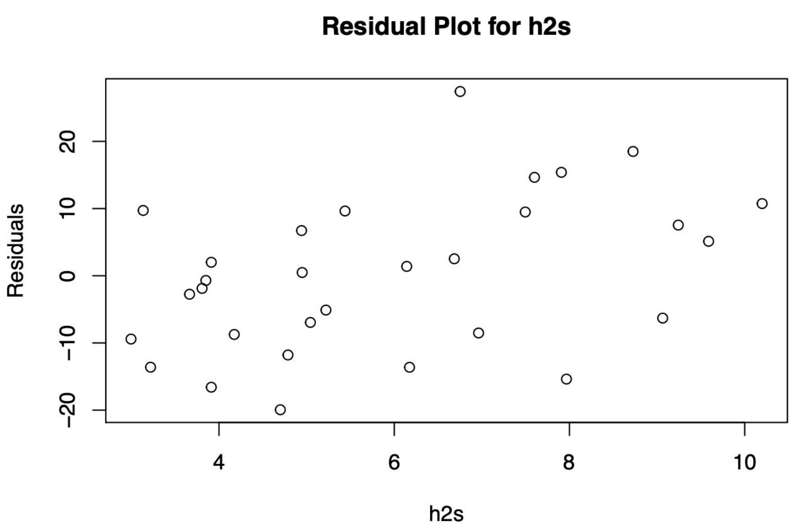
Linear regression equation: Taste = −29.9 + 37.7 * Lactic

t-value = -2.82, p-value = 1.41 * 10^-5

Because the p-value < $\alpha = 0.05$, we reject H0. There is statistically significant evidence that $\beta 1 \neq 0$. Thus, there is evidence of a linear relationship between taste and Lactic acid.







Based on the QQ plot, the residuals have a Normal distribution. The residuals do not seem to have a strong correlation with either Acetic or H2S.

11.58)

	F-Statistic	P-value	R^2	S	Equation
Acetic v Taste	12.11	0.001658	0.2771	13.82	−61.6 + 15.6 * Acetic
H2S v Taste	37.29	1.374e-6	0.571	10.83	−9.79 + 5.78 * H2S
Lactic v Taste	27.55	1.405e-5	0.4959	11.75	−29.9 + 37.7 * Lactic

The intercepts in the 3 equations are different because all of the variables have slopes and intercepts to predict the model. Accordingly, some models may be better than others to use to predict a response variable.

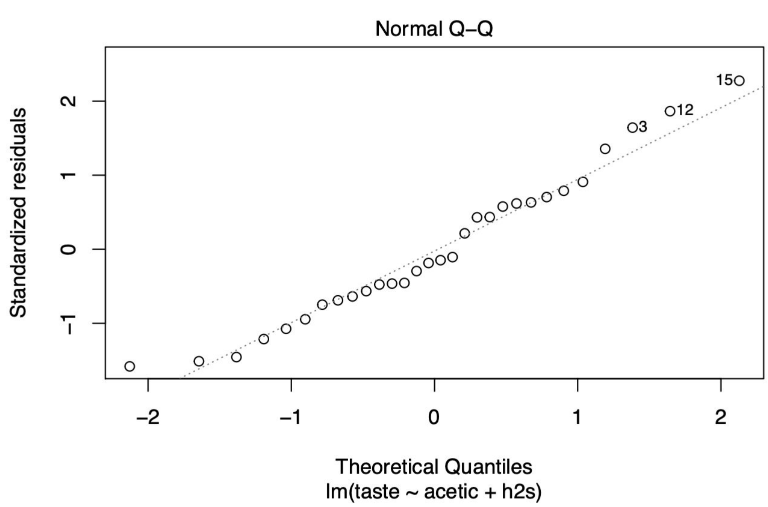
11.59)

Taste = −26.9 + 3.80 * Acetic + 5.15 * H2S

For the coefficient of Acetic, t = 0.84 and P = 0.406

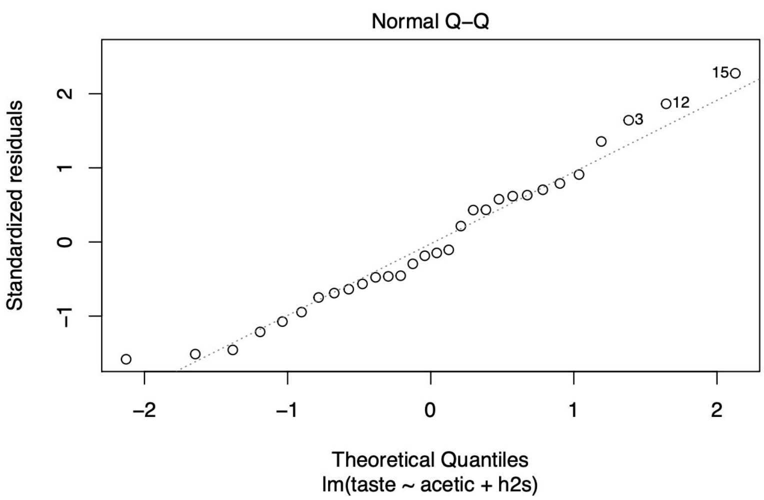
This model is slightly better than the model with only H2S alone; Acetic and H2S are correlated (r = 0.618), so Acetic does not add significant information if H2S is included. The reason that acetic may appear to be a good model but with H2S isn't is because h2s and acetic have a strong correlation. As a result, acetic does not add much of a difference when paired with H2S.

```
Residuals:
    Min      1Q  Median      3Q     Max
-16.113  -6.893  -1.673   6.592  23.715

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -26.940     21.194  -1.271 0.214536
acetic         3.801      4.505   0.844 0.406245
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
```



11.60)

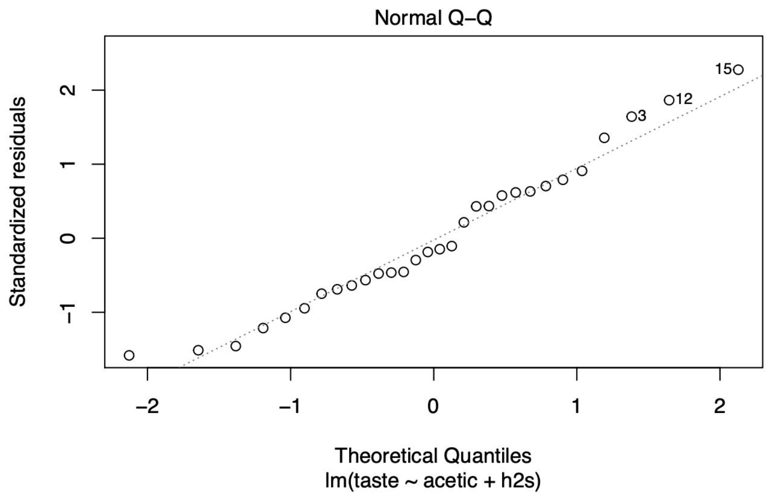
Taste = −27.59 + 3.95 * H2S+ 19.89 * lactic
lm(taste~h2s+lactic)
lactic t-value: 2.499
lactic p-value: 0.01885
Adj. R2 of model: 0.6259
For the coefficient of H2S, t = 3.475 and P = 0.00174
The P-value is much smaller. This means it has a higher chance of being accurate. Therefore, this model is better than only using a single variable.




```
Residuals:
    Min      1Q  Median      3Q     Max
-17.343  -6.530  -1.164   4.844  25.618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
h2s            3.946      1.136   3.475  0.00174 **
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
```

11.61)
Taste = −28.9 + 0.33 * Acetic + 3.91 * H2S + 19.7 * Lactic
The coefficient of Acetic is not significantly different from 0 (P = 0.942).
The coefficient of h2s is significantly different from 0 (P = 0.00425).
The coefficient of lactic is not significantly different from 0 (P = 0.03108).

Based on the R_2 values we can conclude the model of using the H2S and lactic model to predict taste is the best model. Although there are slight differences, it is similar to our current model where we use all 3 variables to predict taste. From the residuals plot we can see that the residuals for the model with all 3 variables still follow a relatively normal distribution.




```
Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
h2s           3.9118     1.2484   3.133  0.00425 **
lactic       19.6705     8.6291   2.280  0.03108 *
---
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
```