Max Shi and Hien Bui
MA 331
Professor Li
December 4, 2019
I pledge my honor that I have abided by the Stevens Honor System.
Final Project Textbook Problems

11.53:

```
     taste             acetic            h2s              lactic
 Min.   : 0.70   Min.   :4.477   Min.   : 2.996   Min.   :0.860
 1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978   1st Qu.:1.250
 Median :20.95   Median :5.425   Median : 5.329   Median :1.450
 Mean   :24.53   Mean   :5.498   Mean   : 5.942   Mean   :1.442
 3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575   3rd Qu.:1.667
 Max.   :57.20   Max.   :6.458   Max.   :10.199   Max.   :2.010
```

Standard deviation:
Taste: 16.25538
Acetic: 0.5708784
H2s: 2.126879
Lactic: 0.30349

Plots:

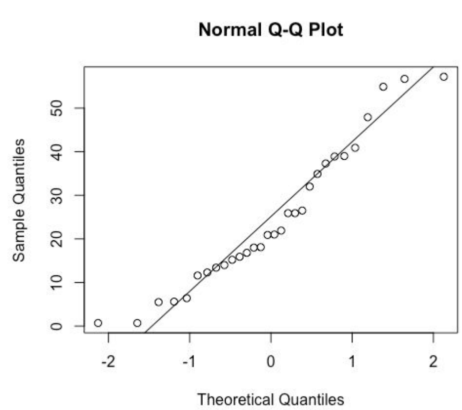
- Taste:

Stem plot:

```
The decimal point is 1 digit(s) to the right of the |

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

Taste is right-skewed

Normal QQ plot



Taste seems to follow Normal distribution

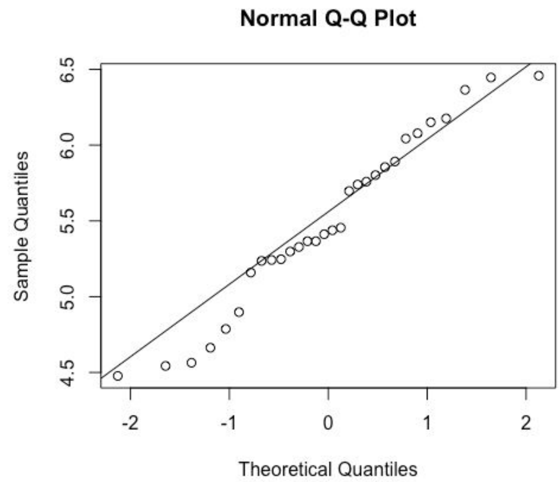
- Acetic:

Stem Plot

```
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
```

Seems to have two peaks.

Normal QQ plot



Seems to follow normal distribution

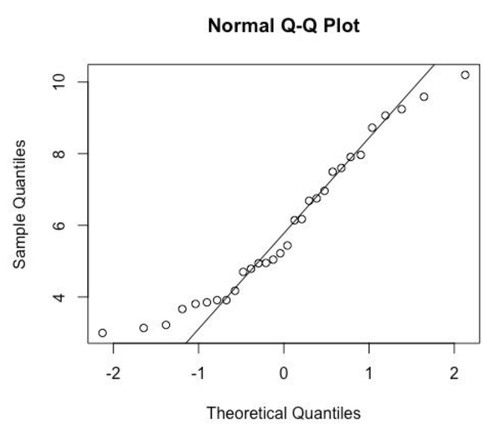
- H2s:

Stem Plot

```
 2 |
 3 | 01278999
 4 | 27899
 5 | 024
 6 | 1278
 7 | 0569
 8 | 07
 9 | 126
10 | 2
```

Seems to be right skewed

Normal QQ plot

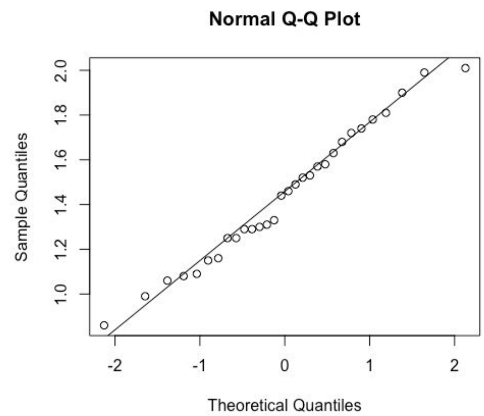


Seems to be normally distributed

- Lactic:

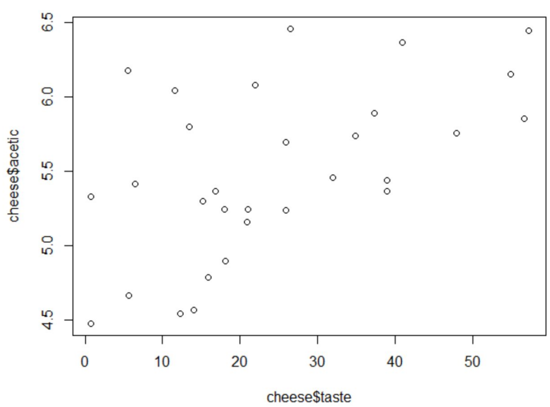
Stem Plot	Normal QQ Plot:

```
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
```



Seems to follow normal distribution

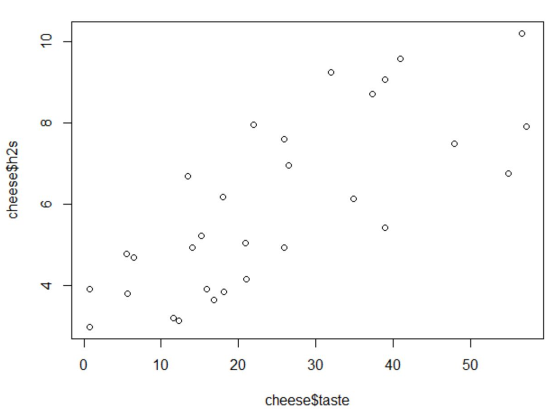
11.54.



Taste vs. Acetic

Taste seems to increase in a vaguely linear pattern with Acetic

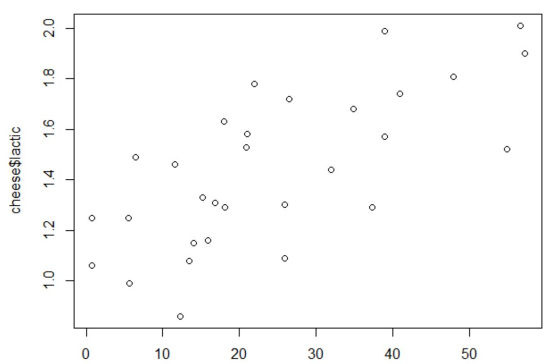
```
Pearson's product-moment correlation

data:  cheese$taste and cheese$acetic
t = 3.4806, df = 28, p-value = 0.001658
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.2359923 0.7594509
sample estimates:
      cor
0.5495393
```



Taste vs H2S

Taste seems to increase in a linear pattern with H2S

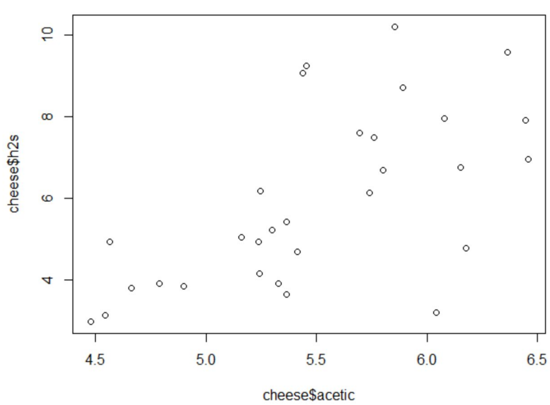
```
    Pearson's product-moment correlation
data:  cheese$taste and cheese$h2s
t = 6.1068, df = 28, p-value = 1.374e-06
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.5434507 0.8771862
sample estimates:
      cor
0.7557523
```



Taste vs Lactic

Taste seems to increase in a linear pattern with Lactic

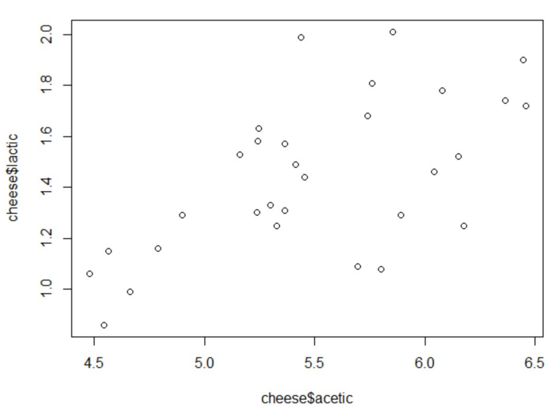
```
    Pearson's product-moment correlation
data:  cheese$taste and cheese$lactic
t = 5.2488, df = 28, p-value = 1.405e-05
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.4609054 0.8490811
sample estimates:
      cor
0.7042362
```



Acetic vs. H2S

Acetic seems to be positively correlated with H2S

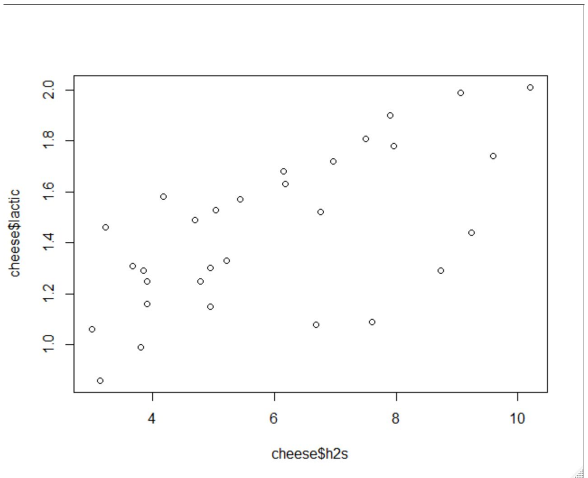
```
    Pearson's product-moment correlation
data:  cheese$acetic and cheese$h2s
t = 4.1591, df = 28, p-value = 0.0002739
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.3314855 0.8000988
sample estimates:
      cor
0.6179559
```



Acetic vs Lactic

Acetic seems to be positively correlated with Lactic

```
    Pearson's product-moment correlation
data:  cheese$acetic and cheese$lactic
t = 4.0079, df = 28, p-value = 0.0004114
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.3112089 0.7918132
sample estimates:
      cor
0.6037826
```



H2S vs Lactic

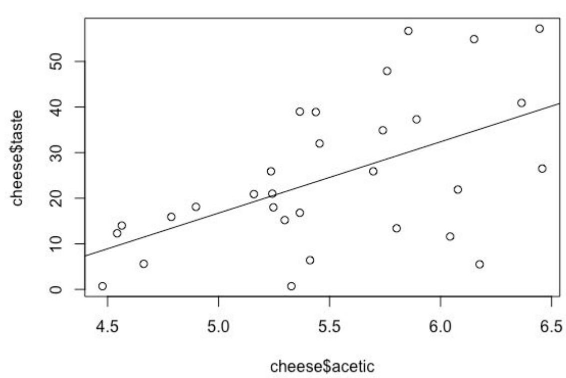
H2S seems to be positively correlated with Lactic

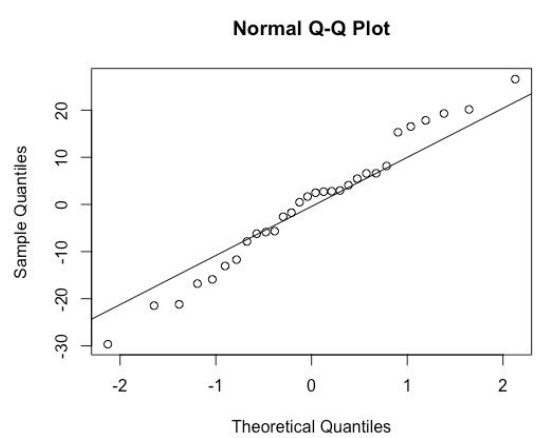
```
      Pearson's product-moment correlation
data:  cheese$h2s and cheese$lactic
t = 4.464, df = 28, p-value = 0.0001198
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.3706465 0.8156103
sample estimates:
      cor
0.6448123
```

11.55:

```
Call:
lm(formula = taste ~ acetic, data = cheese)

Residuals:
   Min     1Q Median     3Q    Max
-29.64  -7.44   2.08   6.60  26.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -61.5       24.9   -2.48   0.0196 *
acetic          15.7        4.5    3.48   0.0017 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.8 on 28 degrees of freedom
Multiple R-squared:  0.302,     Adjusted R-squared:  0.277
F-statistic: 12.1 on 1 and 28 DF,  p-value: 0.00166
```

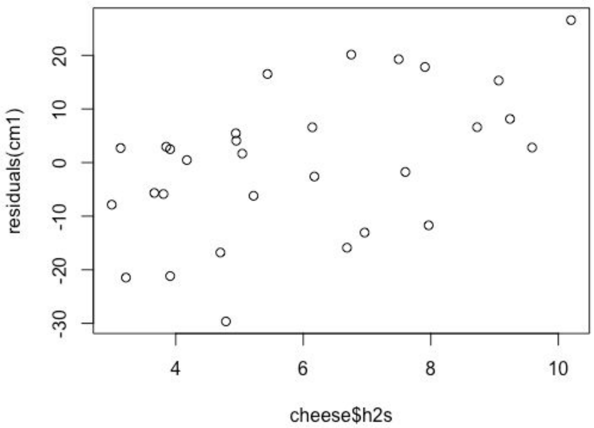
Least squares regression line through scatter plot and QQ Norm plot of residuals:

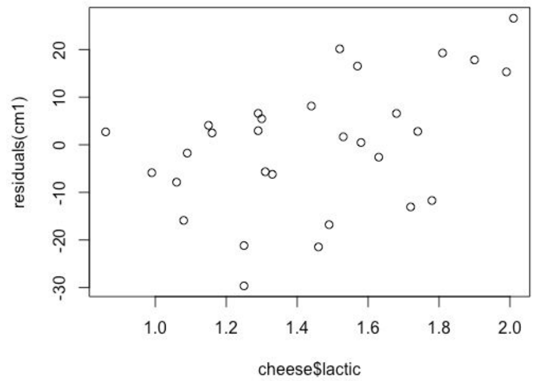




It looks like the residuals have a normal distribution based on the QQ plot.

Residuals vs other lurking variables:



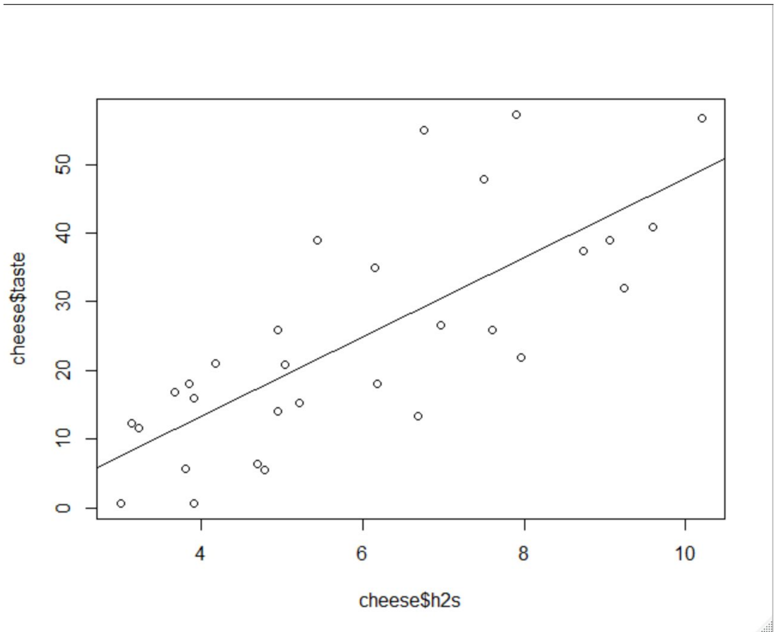


The residuals seem to be positively associated with both h2s and lactic

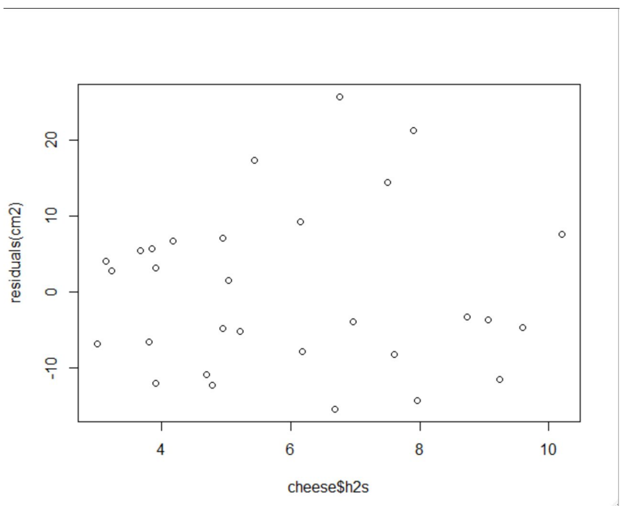
11.56.

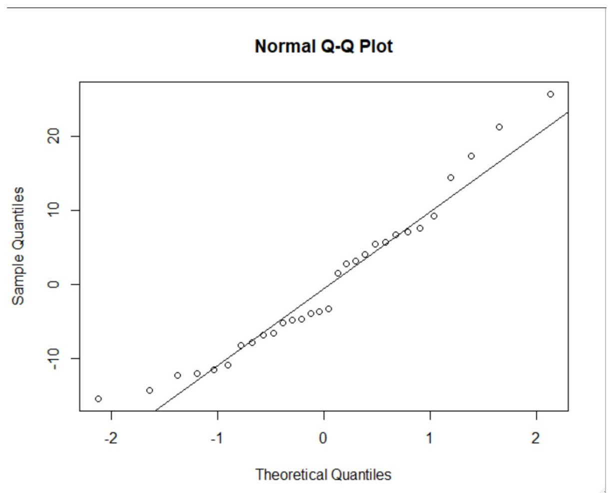
```
Call:
lm(formula = taste ~ h2s, data = cheese)

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
```

Data with least squares regression line:



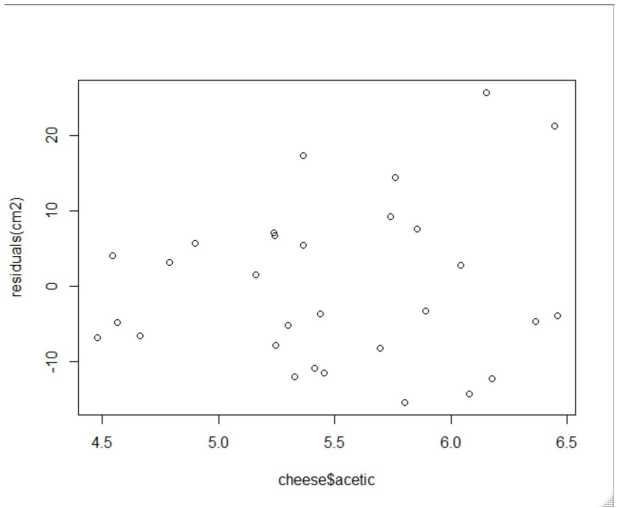
Residuals (with qq norm plot):

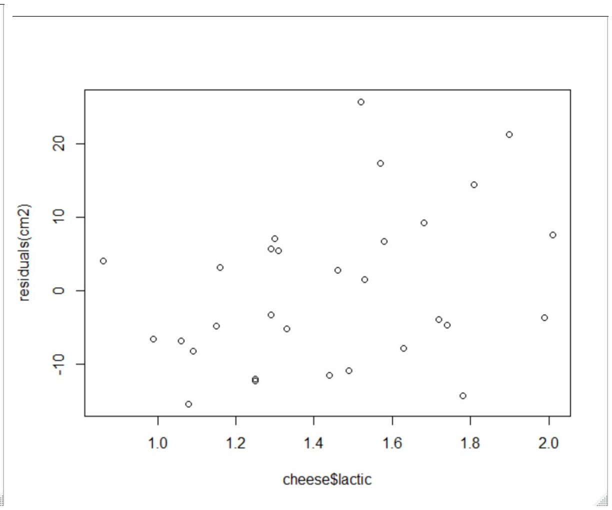




The residuals look normal based on the normal QQ plot, with no obvious associated correlation between H2S and its residuals.

Residuals vs other lurking variables:



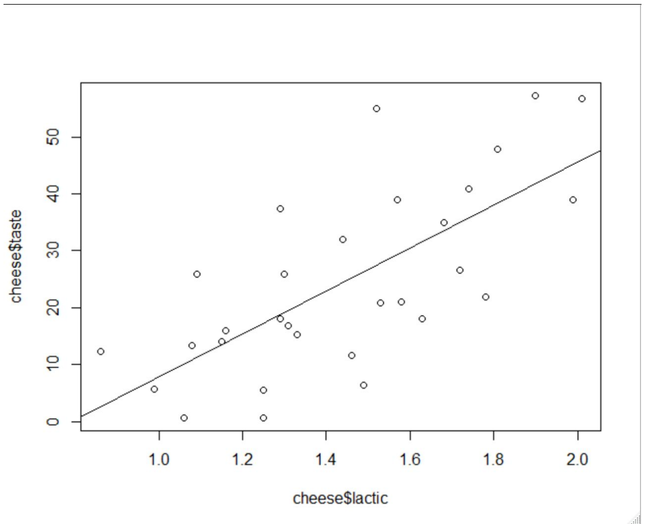


There seems to be a stronger positive association with the residuals and the lactic acid.

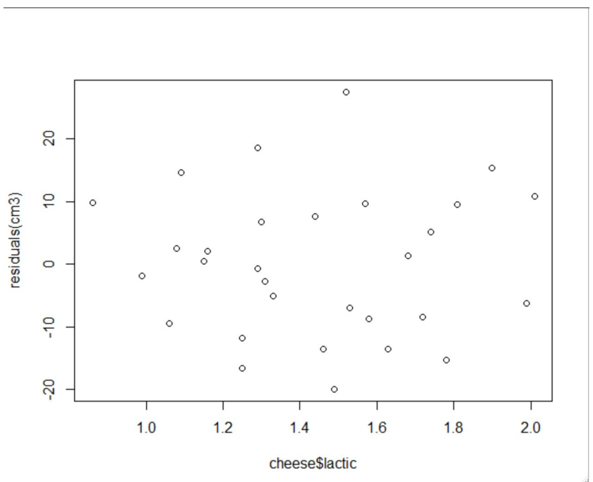
11.57:

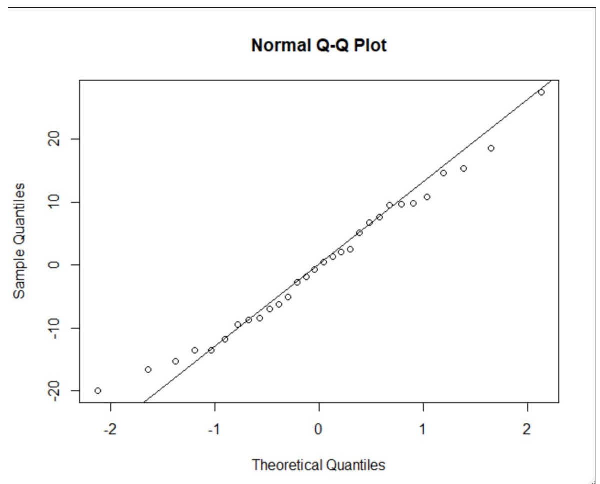
```
Call:
lm(formula = taste ~ lactic, data = cheese)

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
```

Scatter plot with least squares regression line:



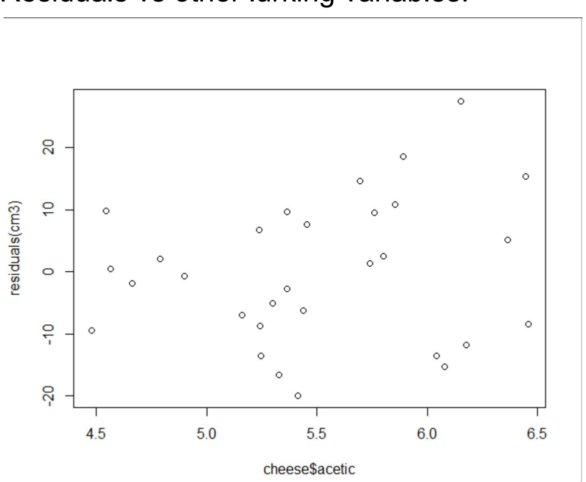
Residuals (with qq norm line):

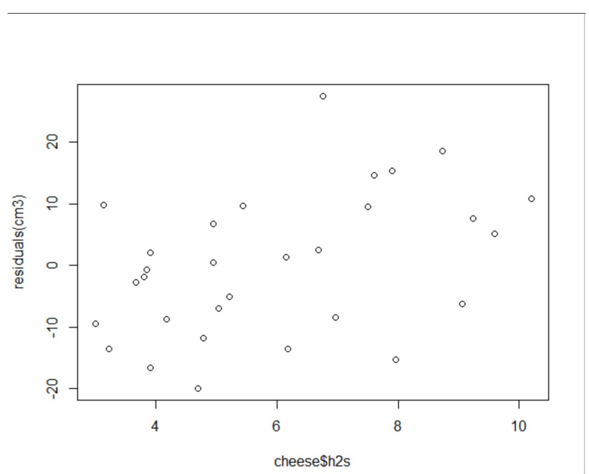




The residuals look normal based on the normal QQ plot.

Residuals vs other lurking variables:





There are no striking patterns revealed in these graphs of the lurking variables and the residuals.

11.58:

	F statistics	P-value	R^2	Estimate std dev
Taste vs Acetic	12.1 on 1 and 28 DF	0.002	0.302	13.8 on 28 df
Taste vs H2S	37.29 on 1 and 28 DF	1.374e-06	0.571	10.83 on 28 degrees of freedom
Taste vs Lactic	27.55 on 1 and 28 DF	1.405e-05	0.496	11.75 on 28 degrees of freedom

Regression equation for:

- Taste vs Acetic model: taste = -61.5 + 15.7*acetic
- Taste vs H2S model: taste = -9.79 + 5.78*h2s
- Taste vs Lactic model: taste = -29.859 + 37.720*lactic

The intercepts in the 3 equations are different because the 3 models use different explanatory variables and those variables have different values.

11.59:

```
Call:
lm(formula = taste ~ acetic + h2s, data = cheese)

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
```

In this model, the p value of acetic acid is 0.406, compared to the P value in the simple linear regression model of 0.0017. We prefer this model over the simple linear regression of just acetic acid due to the smaller P value on H2S, however, this P value is still greater compared to the simple linear regression of H2S by itself, but not by much (multiple linear regression p-value = 7.645e-6, h2s alone p-value = 1.374e-6). Thus, it seems that acetic acid, being positively correlated with H2S already, does not contribute much information to the model.

11.60:

```
Call:
lm(formula = taste ~ h2s + lactic, data = cheese)

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
```

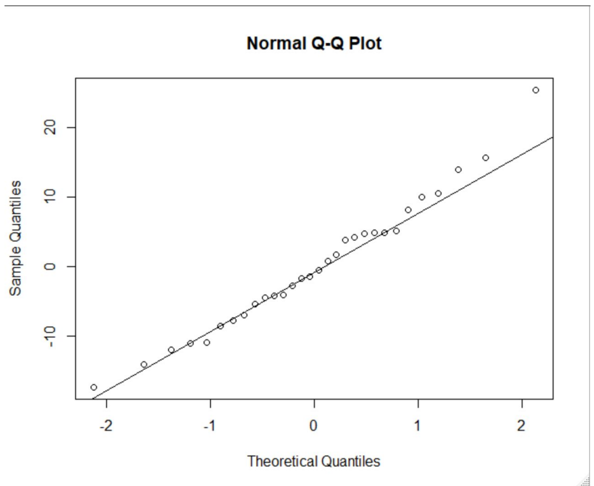
This model seems like the best model so far. If we compare the values to the separate simple linear regressions of taste vs H2S and taste vs lactic acid, we see residual standard error values of 10.83 and 11.75, respectively. These are both higher compared to the new error, which is 9.942. This also translates into a lower p-value for the multiple linear regression -- here it is 6.551e-7 compared to 1.374e-6 and 1.405e-5 in the H2S and lactic acid models, respectively. Furthermore, this model produces a stronger positive correlation of 0.6517 compared to 0.5712 and 0.4959, also supporting the idea of a strong model.

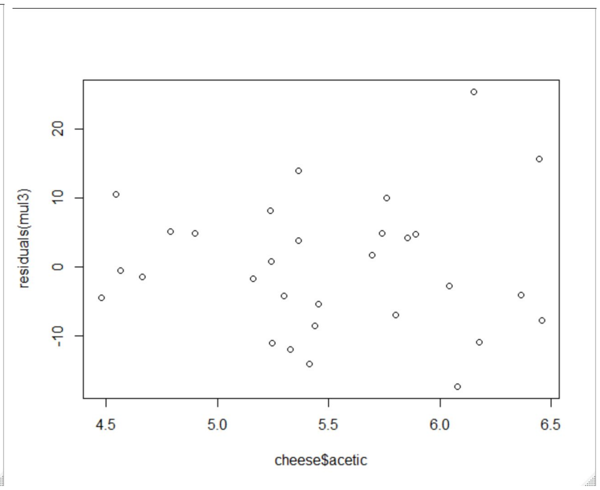
11.61:

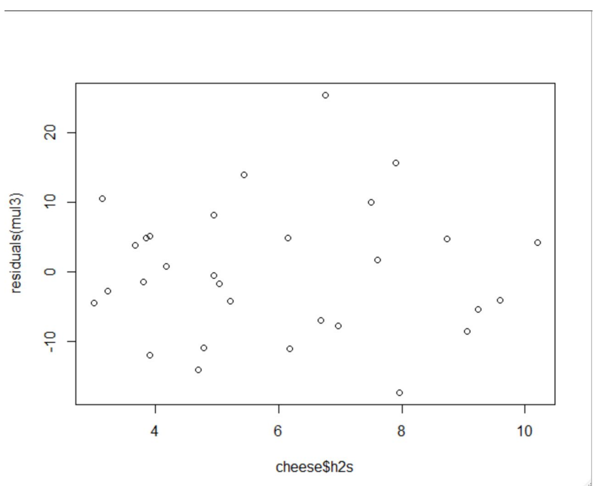
```
Call:
lm(formula = taste ~ acetic + h2s + lactic, data = cheese)

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 * 
```

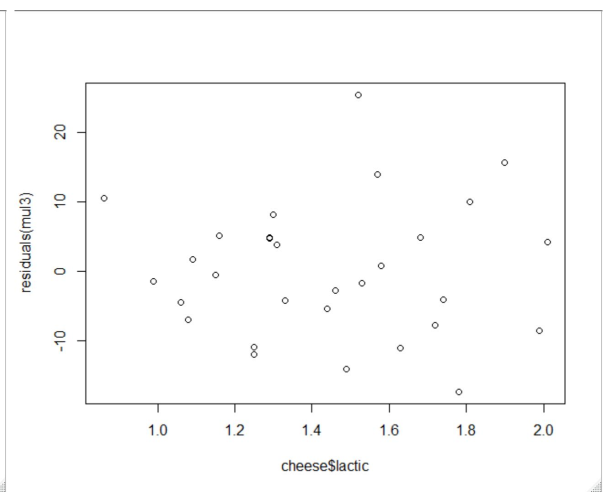
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

Normal QQ plot of residuals, and plots of explanatory variables vs residuals:









From each of these plots, we see a normal distribution of the residuals, with no distinct patterns that signify some sort of correlation with the explanatory variables and the residuals. Interestingly enough, the coefficient associated with acetic acid is very close to 0, also signified with a P-value of 0.94, showing that acetic acid contributes almost no information to this model. Finally, we conclude that the multiple regression model with H2S and Lactic Acid is best, due to that model having the lowest p-value out of all calculated regressions.

Statistical Report of Cheese Data Set

Analysis of Chemical Predictors for the Taste of Cheese

Max Shi and Hien Bui

December 7, 2019

Objectives and Data:

The data contained four variables: Taste (obtained by combining the scores from several tasters), Acetic (concentration of acetic acid whose log transformations were taken), H2S (concentration of hydrogen sulfide whose log transformations were taken), and Lactic (untransformed concentration of lactic acid). We suspect that Taste is directly related to those three chemicals (Acetic, H2S, and Lactic), and we'll try to quantify this relationship to predict which chemical(s) is the best predictor for taste.

Preliminary analysis of the data:

Using tools in R, we created Stem plots and Normal quantile plots for each variable in the data set and found that the data seemed to follow normal distribution.

We then created pairwise scatterplots to analyze the relationship between Taste and each of the three chemicals. For the Taste vs Acetic plot, we found that the points follow a vague linear pattern from bottom left to upper right. However for the Taste vs H2S and Taste vs Lactic plots, this pattern is much clearer. The correlation between Taste and Acetic is also lower than that of the other two pairs. This means that taste tend to increase linearly with H2S and Lactic, and slightly less likely to increase linearly with Acetic.

We also created pairwise scatterplots to analyze the relationship between our three explanatory variables (Acetic, H2S, and Lactic). We found that for all three plots, the points follow a vague linear pattern from bottom left to upper right. However, the correlation is not strong enough to suspect multicollinearity.

Regression Analysis

Regression Tests Done (T = Taste, A = Acetic, H = H2S, L = Lactic)						
Relationship	T=A	T=H	T=L	T=A+H	T=H+L	T=A+H+L
P-value	0.00166	1.37e-6	1.40e-5	7.64e-6	6.55e-7	3.81e-6
R-square (adjusted)	0.277	0.5558	0.4779	0.5512	0.6259	0.6116

Standard Error	13.8	10.83	11.75	10.89	9.942	10.13

The summaries of the simple linear regression and multiple linear regression tests tell us the story of the best regression in this dataset. From the outputs, the lowest p-value is generated when using H2S and lactic acid as predictors for taste, with the regression equation printed here:

Taste = -27.592 + 3.946 H2S + 19.887 lactic

Where taste is a response variable, and H2S is a natural log of the concentration of this chemical and lactic acid is the untransformed concentration of this chemical in the cheese. Two other important summary values are the R-square and the standard error, which are 0.6259 and 9.942, respectively.

To interpret these variables, we can take a look at the coefficients of H2S and lactic. Both of them being positive implies that there is a positive correlation between the concentrations of these two chemicals and how good cheese tastes. Because of the natural log transformation of H2S, it is possible that there are diminishing returns on the effect of the concentration of H2S in the response variables. However, because lactic has been untransformed, it is possible that there is a linear association between the concentration of lactic acid and the taste response variable.

I would be careful to assign these relationships directly to these explanatory variables, however. The R-square lies at 62.59%, which implies that although these two factors explain a majority of the scores in the dataset, it still leaves ~37% of the data unexplained by other factors not included. Thus, I would explore other factors, such as the chemical concentrations of other compounds, as well as other transformations on the data of this set in order to gain a better R-square value and make more accurate predictions.

As for predictions, the standard error value gives insight into the confidence intervals we can create out of this regression. When predictions are made using this regression, we can be 68% sure that the true value will lie within 1 standard error of our prediction, and 95% sure the true value will lie within two standard errors of our prediction.

Thus, we conclude in this report that in order to maximize how good cheese tastes, we should aim to maximize the concentrations of H2S and lactic acid. However, we recommend searching for other factors to create a better model, as our model here has room for improvement, and requires more data points or explanatory variables to increase the R-squared value for a better fit.