Ahilan Subbaian

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

Statistics Final Project

Statistical Report – Part 1

I am a student at Stevens Institute of Technology in MA 331 studying the effect of different chemicals on the taste of cheese and their significance on the taste. My approach was to basically first check if all the data points were independent of each other to see if a simple random sample was replicated. After I found that the data points were independent of each other, my approach was to see how much of taste was described by these chemicals through multiple regressions.

Objectives and Data:

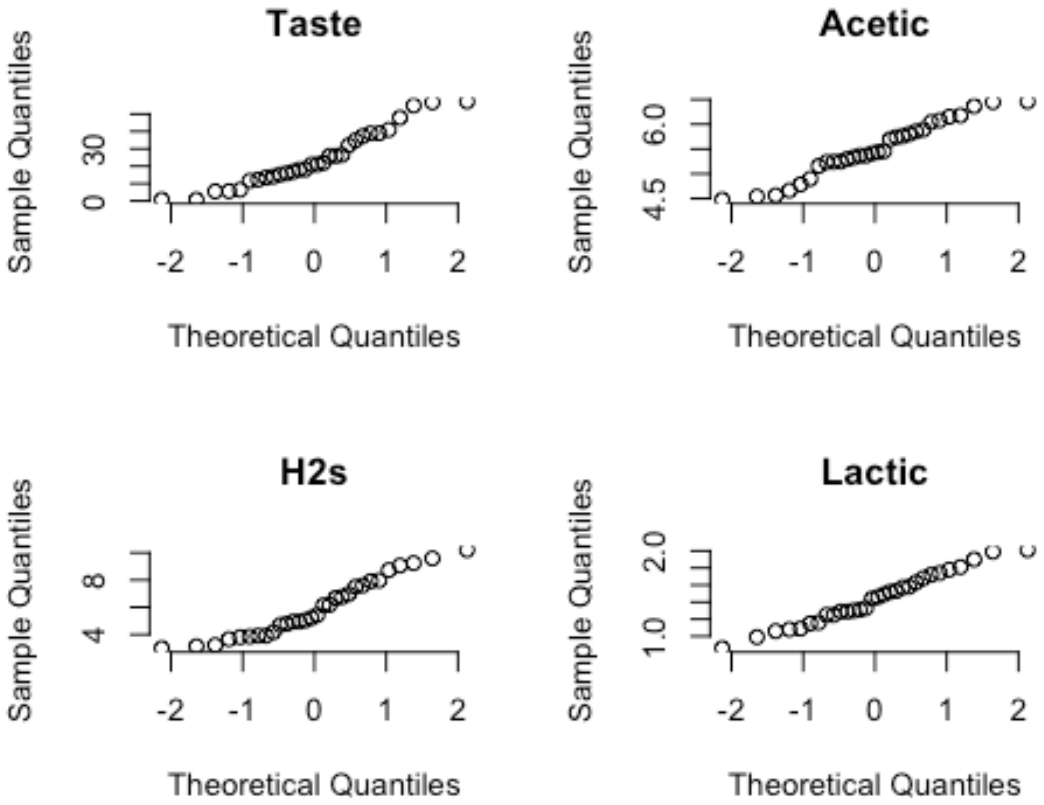
You supplied me with the data for the cheese and how each of the chemicals correlated with the taste according to the 30 different participants. You essentially wanted me to find the correlating between which different chemicals correlated to the change in taste the most. And how to effectively predict the quantitative change in the taste value due to each combination of chemicals.

Statistical Methodology:

My form of the Statistical methodology to solve the problem was to use regression analysis. I used singular regressions to see how each individual independent variable contributed to the taste value as well as multiple regression analysis in order to see which combinations of elements basically contribute to the taste factors the most. In our case, the independent variables were Acetic, H2S, and Lactic, in order to see which contributed to the taste most according to the thirty participants. The sample size is good as well because it is at least a size of thirty. Of course, there are a lot of calculations involved in calculating this, so I used R markdown to help me interpret the data using a linear regression.

Regression Analysis:

First, I needed to analyze all the three chemicals in order to see if they were random and had no correlation to one another so I made QQ plots for each of the four variables: 3 elements and the taste variable. All showed to follow a normal distribution and after a scatterplot graph was done to test the residuals of each element to taste with all of the other variables it was seen that the data points are completely independent to one another which ensured that I can proceed further and make conclusions based on the data if I found relationships. I am attaching the QQ-Plot graphs to this report below.



In order to see the best relationship between the multiple independent variables which in this case were the chemicals, I performed multiple regressions, and I am adding in the regression equations of different combinations, however I found that the combination of all three chemicals helped explain the taste value the most.

The first regression equation is the model I chose, and I proceed to talk about why in my conclusion.

Taste ~ 28.877 + 0.3277 * acetic + 3.9118 * h2s + 19.6705 * lactic.

Taste ~ 26.94 + 3.801 * acetic + 5.15 * h2s

Taste ~ 27.592 + 19.887 * acetic + 3.946 * h2s.

Taste ~ 37.72*lactic - 29.86

Taste ~ 16.65*acetic - 61.499

Taste ~ 5.78*H2s - 9.79

Conclusion:

I found significant evidence that all three chemicals had a significant impact on the taste value. Because I had evidence that they had a significant impact on the value of taste, I believe that the best model to accurately predict the value of taste was the linear regression that included all three chemicals. I have copied below the linear regression of taste with respect to all three chemicals.

```
##
## Call:
## lm(formula = taste ~ (acetic + h2s + lactic))
##
## 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
```

The coefficient of determination is .651 which means 65.1% of the change in taste can be explained by lactic and h2s and acetic. For every increase in lactic by 1 the taste value increases by 19.7. For every increase in h2s by 1 the taste value increases by 3.91. For every increase in acetic by 1 the taste value increases by .0327. I believe this model is a better predictor of taste. The metrics are near similar. Even though the coefficient for acetic is near 0, we previously showed that taste is correlated to acetic, meaning there is significant evidence that taste is affected by the acetic value.

The model that describes Taste is Taste ~ 28.877 + 0.3277 * acetic + 3.9118 * h2s + 19.6705 * lactic.

It is worrying that the p-value for acetic is .94198, however, given that we have significant evidence that taste is affected by acetic and that the p-value of the model is 3.81e-06 which is still below alpha and the R^2 value is higher in this model which means more of taste can be explained in this model. Therefore, I conclude that the model with all three chemicals best predicts taste.

Final Project

Ahilan Subbaian

5/14/2020

11.53)

```
##              Mean     Median Standard Deviation        IQR
## Taste  24.5333333 20.9500000          16.2553828 23.1500000
## Acetic  5.4980333  5.4250000           0.5708784  0.6452500
## H2s     5.9417667  5.3290000           2.1268792  3.5972500
## Lactic  1.4420000  1.4500000           0.3034900  0.4175000
```

STEM PLOTS:

Taste Slightly skewed to the right

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 11666
##   1 | 223456788
##   2 | 112667
##   3 | 25799
##   4 | 18
##   5 | 577
```

Acetic Uniform distribution

```
##
##   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 Right skewed

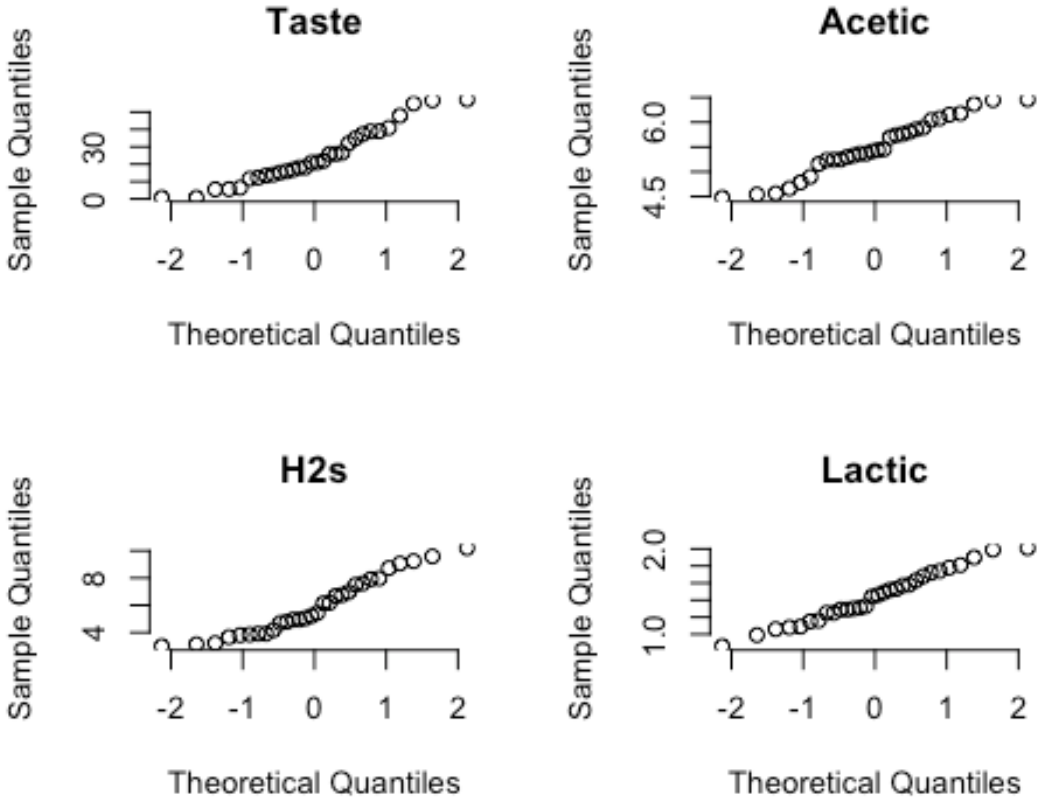
```
##
##   The decimal point is at the |
##
##    2 |
```

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

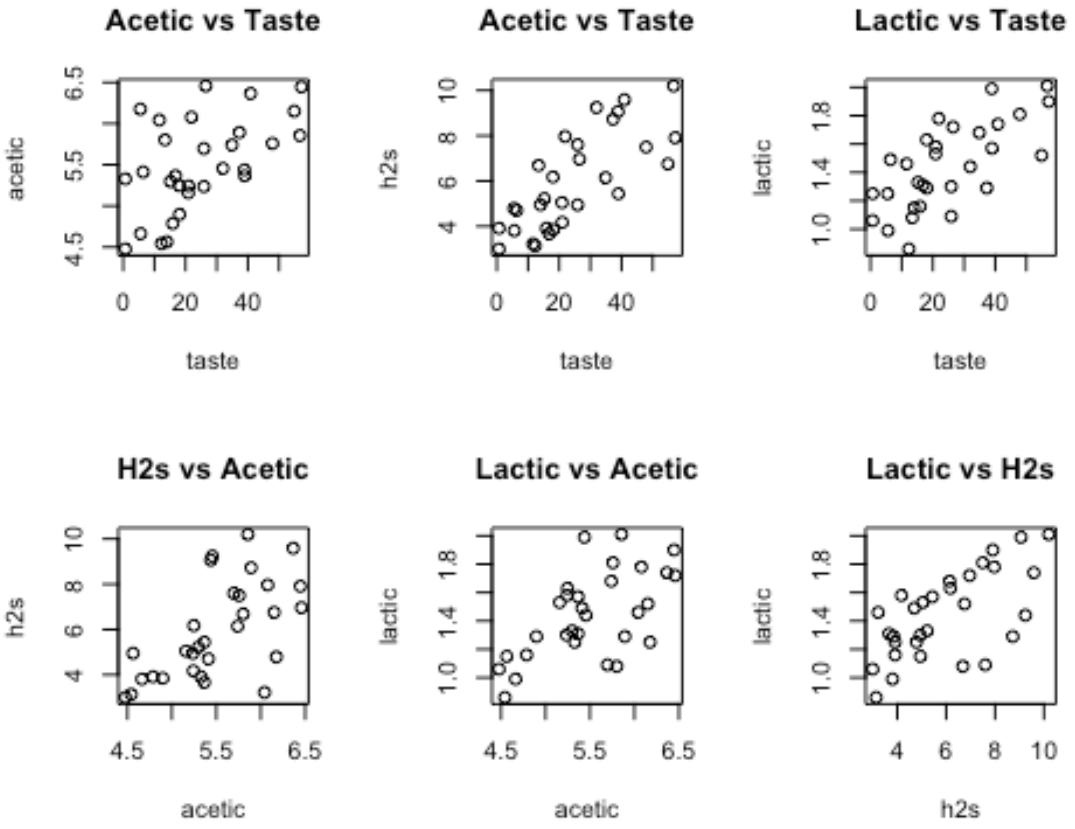
Lactic Normally distributed

```
##
##   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
```

QQ PLOT:



Taste - slightly off the line Actic - slightly off the line H2s - appearns normal Lactic - appearns normal 11.54)




```
##                   Correlation  T-Statistic      P-Value
## Acetic vs Taste  5.495393e-01 3.480551e+00 1.658192e-03
## Acetic vs Taste  7.557523e-01 6.106770e+00 1.373783e-06
## Lactic vs Taste  7.042362e-01 5.248799e+00 1.405117e-05
## H2s vs Acetic    6.179559e-01 4.159072e+00 2.739173e-04
## Lactic vs Acetic 6.037826e-01 4.007930e+00 4.113657e-04
## Lactic vs H2s    6.448123e-01 4.464010e+00 1.198401e-04
```

The correlations can be described as Moderately Strong Positively, Strong Postively, Strong Postively, Moderately Strong Positively, Moderately Strong Positively and Moderately Strong Positively respectively

H_0: rho = 0 H_a: rho != 0 rho = population correlation coefficient DF = 30 - 2 = 28

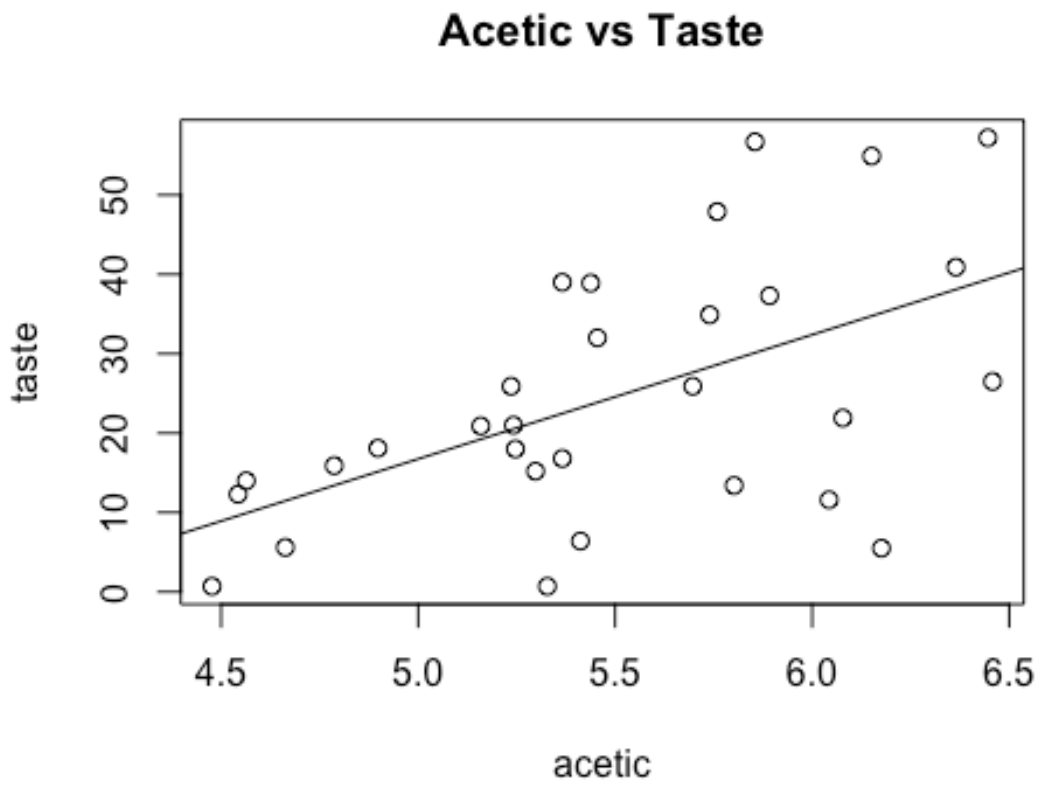
Because all p-values are less than alpha which is .05, we can reject the null hypothesis with significant evidence that there exists a linear correlation between each of the two variables.

11.55)

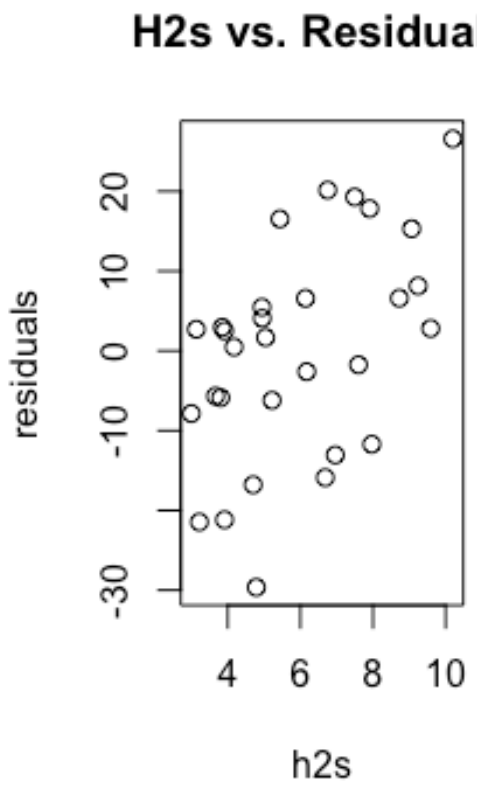
```
##
## Call:
## lm(formula = taste ~ acetic)
##
## Residuals:
```

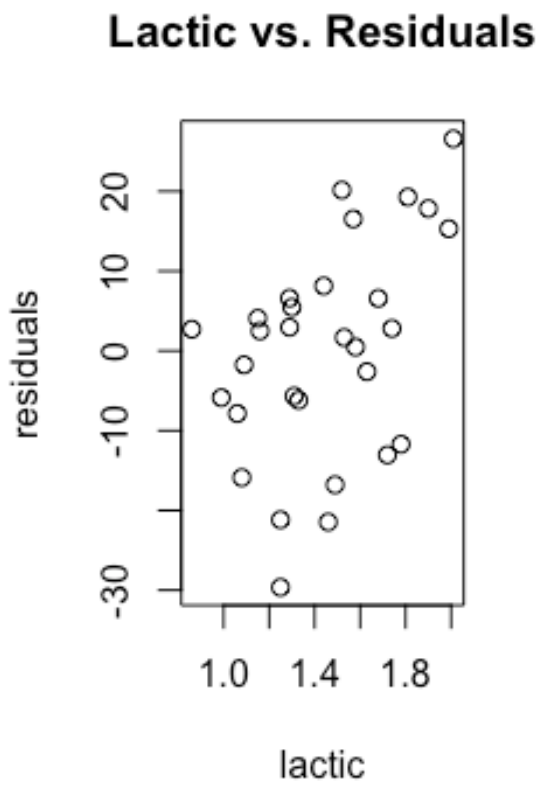
```
##     Min      1Q  Median      3Q     Max
## -29.642  -7.443   2.082   6.597  26.581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -61.499     24.846  -2.475  0.01964 *
## acetic        15.648      4.496   3.481  0.00166 **
## ---
## Signif. codes:  0 '***' 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
```

The coefficient of determination is .302 which means 30% of the change in taste can be explained by acetic. For every increase in acetic by 1 the taste value increases by 15.6.



The residuals in both cases seem normally distributed and are positively associated with H2S and Lactic.



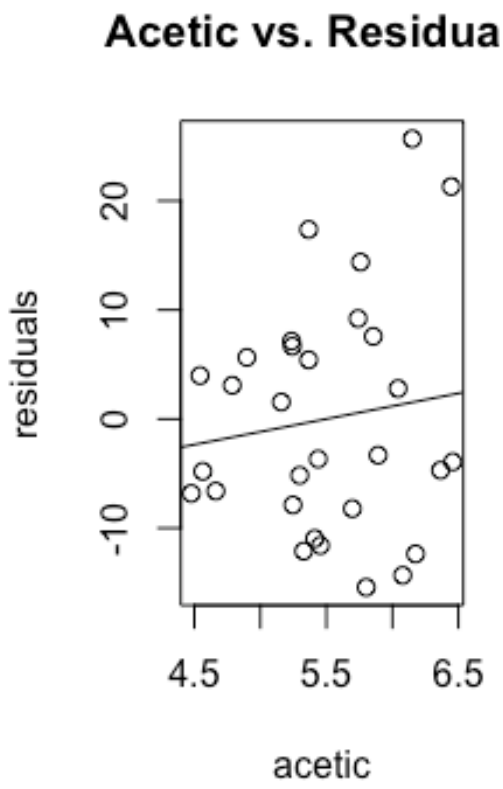


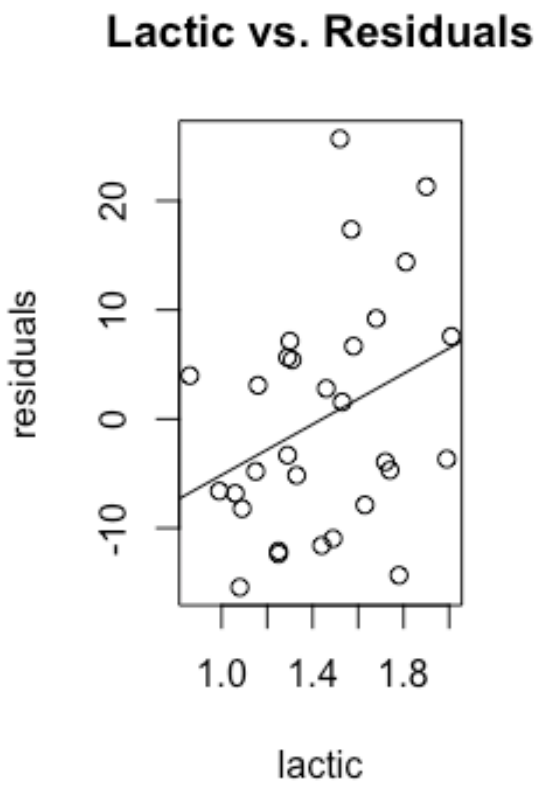
11.56)

```
##
## Call:
## lm(formula = taste ~ h2s)
##
## 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
```

The coefficient of determination is .571 which means 57.1% of the change in taste can be explained by h2s. For every increase in h2s by 1 the taste value increases by 5.77.

The residuals in both cases seem normally distributed as the residuals are randomly plotted and are slightly positively associated with Acetic and a little more positively associated with Lactic as seen through the line of best fit.

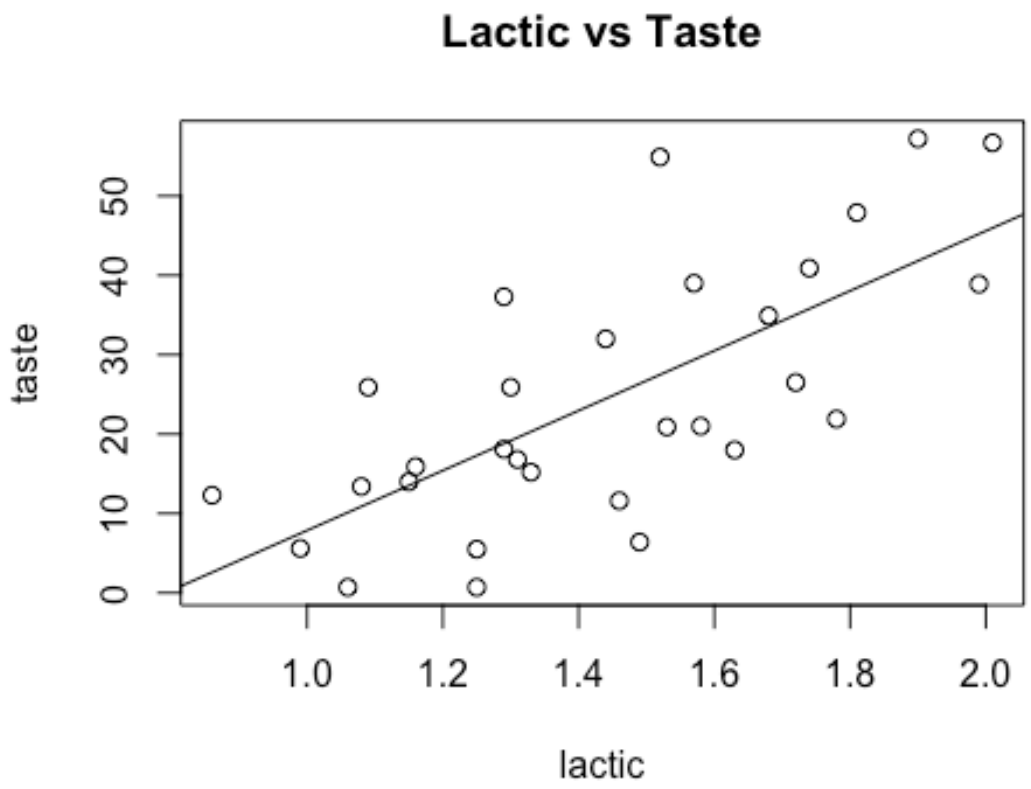




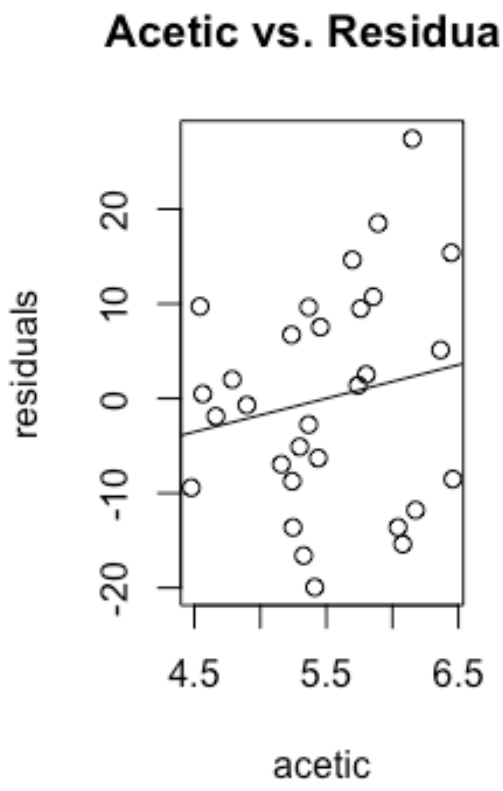
11.57)

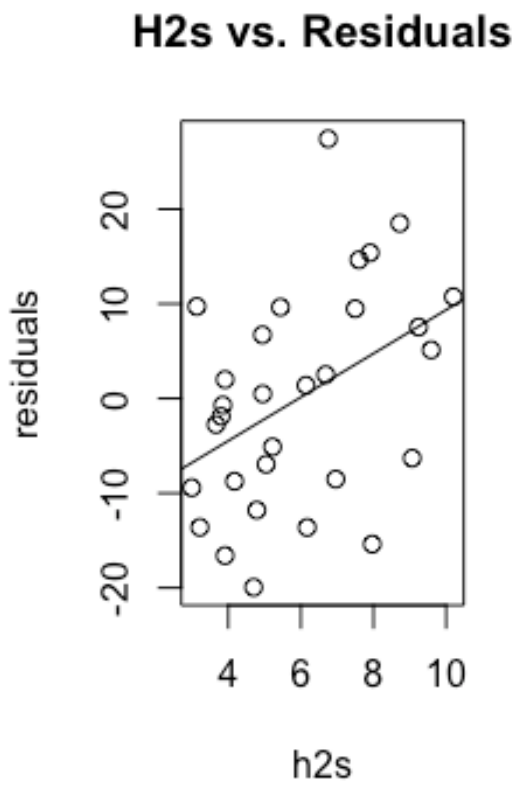
```
##
## Call:
## lm(formula = taste ~ lactic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9439  -8.6839  -0.1095   8.9998  27.4245
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -29.859     10.582  -2.822  0.00869 **
## lactic        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
```

The coefficient of determination is .496 which means 49.6% of the change in taste can be explained by lactic. For every increase in lactic by 1 the taste value increases by 37.7.



The residuals in both cases seem normally distribute as the residuals are randomly plotted and are slightly positively associated with Acetic and a little more positively associated with h2s as seen through the line of best fit.





11.58)

```
##          F-statistic           R^2 Standard Dev      P-value
## Lactic 27.549891146  0.495948607 11.745042567  0.000014050
## Acetic 12.114236229  0.301993441 13.821237740  0.001568000
## H2S    37.292645284  0.571161501 10.833382298  0.000001374
```

Taste = 37.72 * lactic - 29.86

Taste = 16.65 * acetic - 61.499

Taste = 5.78 * H2s - 9.79

For each of the regression equations the y-intercepts are negative and can be ignored as a negative value has no context. This occurs because there has to be a minimum value of each of the three elements in cheese, which is obtained from the x-intercept of each equation.

11.59)

```
##
## Call:
## lm(formula = taste ~ acetic + 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
```

The coefficient of determination is .582 which means 58.2% of the change in taste can be explained by acetic and h2s. For every increase in acetic by 1 the taste value increases by 3.8. For every increase in h2s by 1 the taste value increases by 5.14. Acetic in this model does not really help that much in predicting the taste with h2s as acetic is already correlated with the h2s 61.796%. So this model isn't much better than the alternative that was asked to compare.

The model that describes Taste is Taste ~ 26.94 + 3.801 * acetic + 5.15 * h2s

11.60)

```
##
## Call:
## lm(formula = taste ~ lactic + h2s)
##
## 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 **
## lactic        19.887      7.959   2.499  0.01885 *
## h2s            3.946      1.136   3.475  0.00174 **
## ---
## 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
```

The coefficient of determination is .651 which means 65.1% of the change in taste can be explained by lactic and h2s. For every increase in lactic by 1 the taste value increases by 19.8. For every increase in h2s by 1 the taste value increases by 3.94. We get a more accurate predictive model when taking into account of two predictors rather than just

using 1, because the taste value definitely relies on both. This is seen through an improvement residual value as the residual average, or error average is very close to zero at -4.996004e-16.

The model that describes Taste is Taste ~ 27.592 + 19.887 * acetic + 3.946 * h2s.

11.61)

```
##
## Call:
## lm(formula = taste ~ (acetic + h2s + lactic))
##
## 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
```

The coefficient of determination is .651 which means 65.1% of the change in taste can be explained by lactic and h2s and acetic. For every increase in lactic by 1 the taste value increases by 19.7. For every increase in h2s by 1 the taste value increases by 3.91. For every increase in acetic by 1 the taste value increases by .0327. I believe this model is a better predictor of taste. The metrics are near similar. Even though the coefficient for acetic is near 0, we previously showed that taste is correlated to acetic, meaning there is significant evidence that taste is affected by the acetic value.

The model that describes Taste is Taste ~ 28.877 + 0.3277 * acetic + 3.9118 * h2s + 19.6705 * lactic.