Stephaan Silne

MA 331: Intermediate Statistics

Final Project

December 16th, 2020

Executive Summary

The following report outlines my findings on the relationship between the taste of cheese and the chemical acids that develop as the cheese matures (Acetic, Hydrogen Sulfide, Lactic). Using the CHEESE data file given, the statistical report includes singular and pairwise numerical and graphical descriptions of the data and a statistical methodology that includes several regression analyses. In closing, the multiple linear regression model $\widehat{Taste} = -27.592 + 19.887\ Lactic + 3.946\ H2S$, was shown to be the *best* model in quantifying the relationship between the taste of cheese and the chemical compounds within the food product.

Description of the Data

In the study regarding the maturation of cheddar cheese from LaTrobe Valley of Victoria, Australia, the several chemical compounds that develop during this process and the implications it imposes on the taste of the food product were investigated. To synthesize, 30 samples (labeled as 'Cases' in the dataset) of cheddar cheese were analyzed by their (1) Taste, and chemical concentration of: (2) Acetic Acid, (3) Hydrogen Sulfide, and (4) Lactic Acid. The variable Taste was obtained by collecting responses from several tasters for each individual sample. As for the chemical compounds, variables Acetic Acid and Hydrogen Sulfide were measured and transformed through the natural log function. The transformation of this data allows the use of procedures based on the Normal distributions to be better justified, therefore more accurate analysis. Contrastingly, the concentration of the variable Lactic Acid was not transformed. More in depth, through RStudio software, the header of the dataset is shown below.

case	taste	acetic	h2s	lactic
1	12.3	4.543	3.135	0.86
2	20.9	5.159	5.043	1.53
3	39.0	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

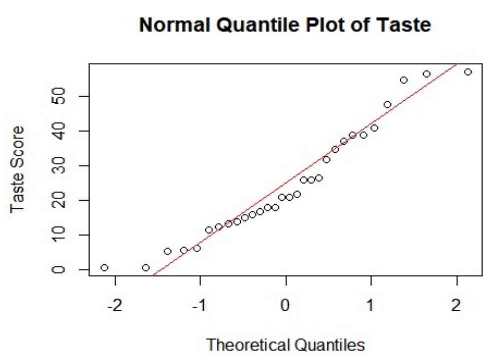
Preliminary Analysis of the Data

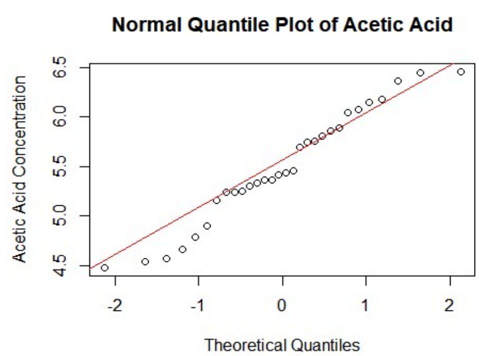
The first step of any type of data analysis are to carefully examine each variable. Therefore, for each of the four variables: Taste, Acetic Acid, Hydrogen Sulfide (H2S), Lactic Acid, their means, medians, standard deviations, and interquartile ranges were calculated through RStudio software.

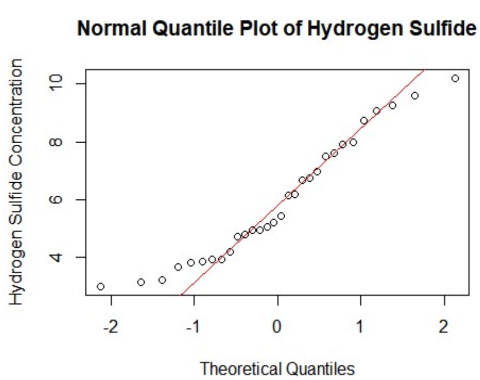
	$\bar{x}$	M	s	IQR
Taste	24.533	20.950	16.255	23.150
Acetic	5.498	5.425	0.570	0.645
H2S	5.941	5.329	2.127	3.597
Lactic	1.422	1.450	0.303	0.417

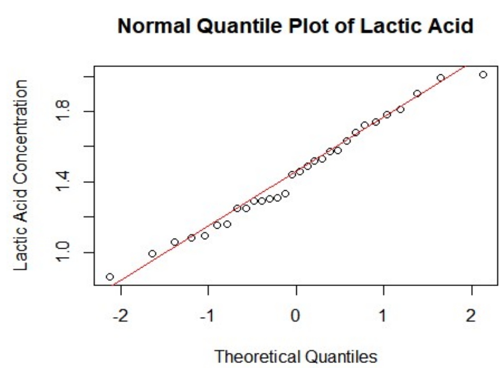
For the chemical concentrations, the transformed data of acetic acid and hydrogen sulfide have larger means and medians than the untransformed lactic acid. Of the three, hydrogen sulfide has the large spread in its data as well as the largest interquartile range, thus the data points are not as close to means as the other chemicals. The same can be said about the variable Taste in comparison to the three chemicals altogether, as it has larger variability (s = 16.225) among its dataset as well as a larger interquartile range (IQR = 23.150). Compared to the rest of the variables, lactic acid appears to have less spread, meaning it is more normal. Despite this, the largeness of the sample sizes for each ($n \geq 30$) is enough to assume Normality. For further confirmation, stemplots can be used to examine the shapes of their distributions and normal quantile plots can indicate whether the distributions appear Normal through RStudio software.

Taste	Acetic	H2S	Lactic
## 0 | 11666	## 44 | 846	## 2 |	## 8 | 69
## 1 | 223456788	## 46 | 69	## 3 | 01278999	## 10 | 68956
## 2 | 112667	## 48 | 0	## 4 | 27899	## 12 | 5599013
## 3 | 25799	## 50 | 6	## 5 | 024	## 14 | 4692378
## 4 | 18	## 52 | 4450377	## 6 | 1278	## 16 | 38248
## 5 | 577	## 54 | 146	## 7 | 0569	## 18 | 109
	## 56 | 046	## 8 | 07	## 20 | 1
	## 58 | 069	## 9 | 126	
	## 60 | 4858	## 10 | 2	
	## 62 | 7		
	## 64 | 56		







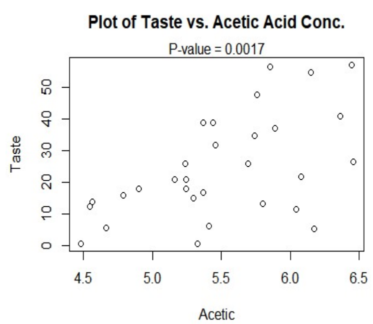


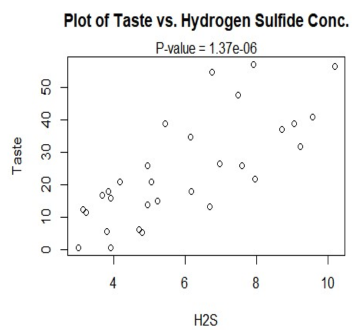
The variables exhibit minimal deviations from the Normality and the normal quantile plots appears to be linear. The variables, Taste and H2S are slightly skewed to the right while Acetic exhibits a bimodal distribution, as shown in the stem plot. As mentioned before, Lactic is closer to Normality given the singular peak in the stemplot about the center. There do not appear to be any extreme outliers in any of the variables analyzed. One may note that depending on the statistical methodology that will be conducted in the following section of the report, the normality of these variable may not be necessary. Despite this, examining these plots allows us to understand each variable alone before attempting any complex model of analysis; allowing us to check any extreme values that can be corrected.

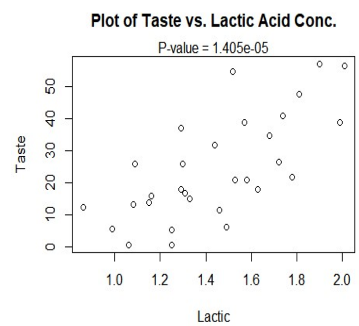
With the objective of this report being to investigate the implications that chemical compounds may have in the Taste of matured cheddar cheese, the second step in the analysis would be examine the relationships between all pairs of variables. For this, scatterplots and correlations can be administered for studying any two-variable relationships and obtain a fair representation of the data. Below is a correlation table that was created through R software, computing the sample correlation coefficient $r(x, y)$ for each pair.

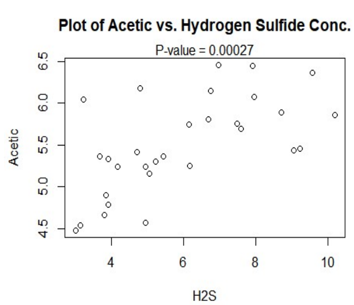
Correlation Table				
	Taste	Acetic	H2S	Lactic
Taste	1.000000	0.549539	0.755752	0.704236
Acetic	0.549539	1.000000	0.617956	0.603783
H2S	0.755752	0.617956	1.000000	0.644812
Lactic	0.704236	0.603782	0.644812	1.000000

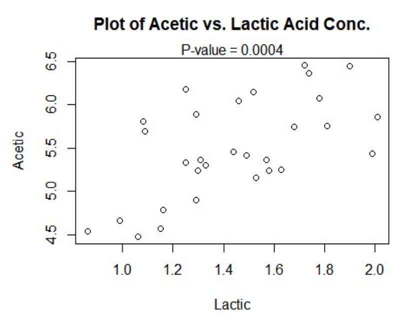
The correlation table exhibits a diagonal of $r(x, y) = 1$, given that it marks the correlation between a variable with itself. More importantly, the variables Taste and Hydrogen Sulfide have the largest correlation between each other $(r(x, y) = 0.756)$ while Taste and Acetic have the lowest correlation between each other $(r(x, y) = 0.550)$. In plotting each variable against one another, a P-value for the hypothesis t-test was calculated, where the null hypothesis states the population correlation is 0 ($H_0: \rho(x, y) = 0$) and the two-sided alternative hypothesis states $H_a: \rho(x, y) \neq 0$.

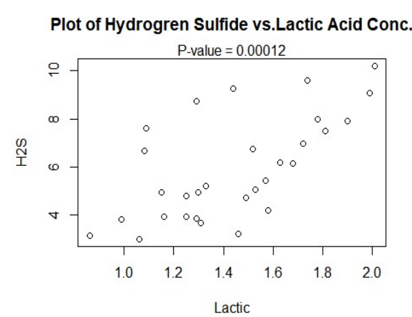












All the six plots show a positive correlation between the variables. All correlations are thus positive, and the corresponding P-values for each case are significantly different from 0, as for each, $p \ll 0.05$. It can be concluded that the P-values are correct given that the two variables under each investigation are normally distributed.

Statistical Methodology

To understand the relationship between the taste of cheese and the chemical compounds that develop within the food product during maturation, multiple linear regression is a good method to consider for our study. With Taste acting as a response variable Y and the three chemicals acting as explanatory covariates $(x_{i,1}, \ldots, x_{i,k})$, for $i =$

$1, \ldots, n$, a regression model can be formulated to quantify this relationship. In assessment of this methodology, doing multiple linear regression is advantageous as we will be able to determine the relative influence that one or more explanatory variables has on the response variable, allowing us to study the influence of the chemical composition on Taste instead of inaccurately representing Taste as an effect of a single chemical. While multiple linear regression aligns with the objective of the study, it may prove to be disadvantageous if the correlations between certain variables affects the strength and goodness of the linear model. However, the regression model can be refined to eradicate these possible errors, thus multiple linear regression is our method of choice.

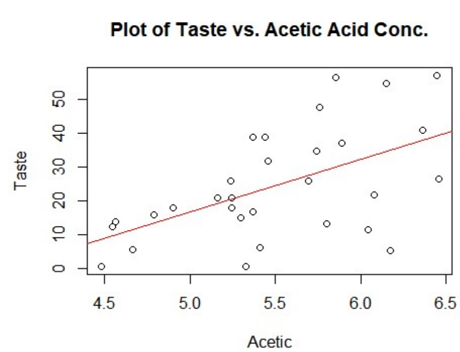
Regression Analysis

To affirm that multiple linear regression is the appropriate method, conducting simple linear regression for Taste with respect to each of the three chemical and comparing it to the multiple regression model should be done. For simple linear regression, it must be assumed that the response variable $Y \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ is of normal distribution, the mean of the response variable y depends on the explanatory variable x, and the random errors $\epsilon_i \sim N(0, \sigma^2)$ are normally distributed. With the β's being determined through method of least squares; the linear models are computed through RStudio software.

Our first simple linear regression model treats Taste as a response variable and Acetic as an explanatory variable.

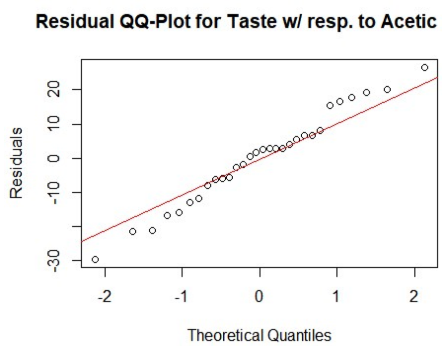
```
##
## Call:
## lm(formula = chz$taste ~ chz$acetic, data = chz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.642  -7.443   2.082   6.597  26.581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -61.499     24.846  -2.475  0.01964 *
## chz$acetic    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
```

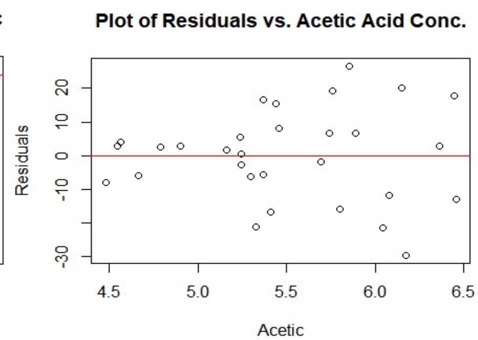
The regression equation is $\widehat{Taste} = -61.50 + 15.648\,Acetic.$ In terms of summary statistics, the model has a standard deviation $s = 13.82$ and a coefficient of determination $R^2 = 0.302.$ For the significance testing on regression parameter β_1, the testing statistic and corresponding P-value are given, respectively $(t = 3.48, P = 0.00166)$. To represent this linear relationship, the scatterplot below depicts the plot of Taste with respect to Acetic concentration.

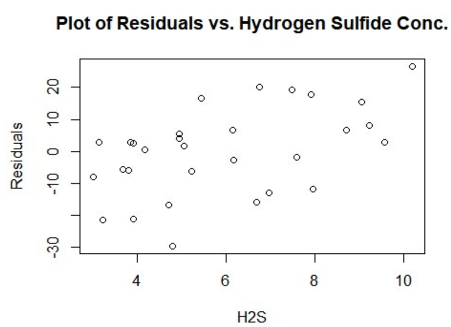


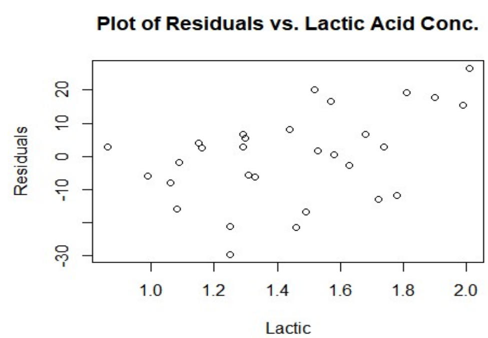
It is usual to plot the residuals versus the predicted values ŷ and versus each of the explanatory variables, looking any unusual observations such as a curved relationship (rather than linear) or extreme outliers. Under the assumption that the deviations ε in the model are Normally distributed, the residuals should be as well as any unusual observation would be a violation of our analysis. Also, to reflect Normality, the stemplot of the residuals is computed through the software.

Stem Plot of Residuals
-3 | 0
-2 | 11
-1 | 7632
-0 | 866632
0 | 02233345778
1 | 5789
2 | 07









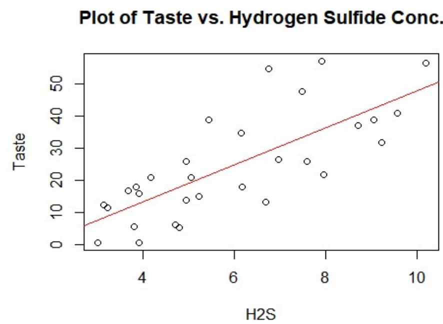
Based on this, the slope is significantly different from 0. Based on the unimodal stemplot and the linear QQ-plot, the residuals appear to have a Normal distribution. The scatterplot of residuals exhibits a positive association when in respect to H2S and Lactic Acid. The plot of residuals against Acetic suggest that a greater scatter in the residuals for the large Acetic values.

Our second simple linear regression model treats Taste as a response variable and Hydrogen Sulfide as an explanatory variable.

```
## Call:
## lm(formula = chz$taste ~ chz$h2s, data = chz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.426  -7.611  -3.491   6.420  25.687
##
## Coefficients:
```

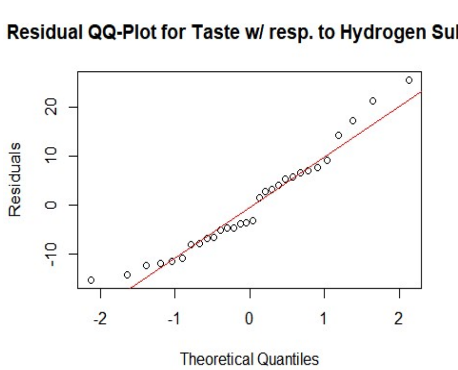
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.7868     5.9579  -1.643    0.112
## chz$h2s       5.7761     0.9458   6.107 1.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.83 on 28 degrees of freedom
## Multiple R-squared:  0.5712, Adjusted R-squared:  0.5558
## F-statistic: 37.29 on 1 and 28 DF,  p-value: 1.374e-06
```

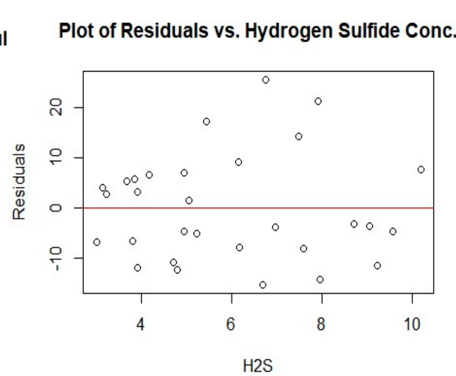
The regression equation is: $\widehat{Taste} = -9.787 + 5.7761\ H2S$. In terms of summary statistics, the model has a standard deviation $s = 10.83$ and a coefficient of determination $R^2 = 0.571$. For the significance testing on regression parameter β_1, the testing statistic and corresponding P-value are given, respectively $(t = 6.11, P = 1.37 \times 10^{-6} \ll 0.05)$. To represent this linear relationship, the scatterplot below depicts the plot of Taste with respect to Hydrogen Sulfide concentration.

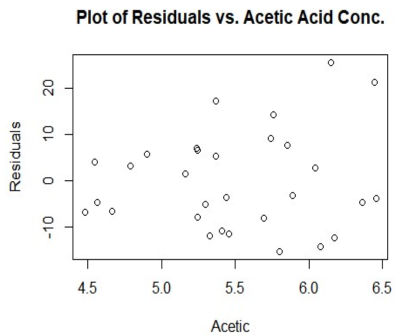


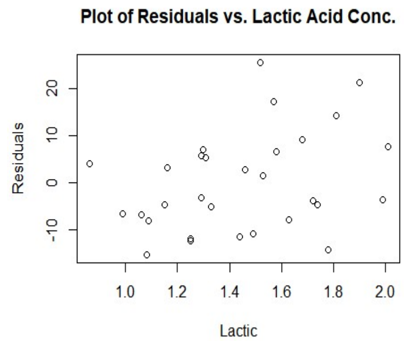
To examine residual normality, the following plots are highlighted below.

```
Stem Plot of Residuals
##   -1 | 5
##   -1 | 42221
##   -0 | 8877555
##   -0 | 443
##    0 | 2334
##    0 | 567789
##    1 | 4
##    1 | 7
##    2 | 1
##    2 | 6
```







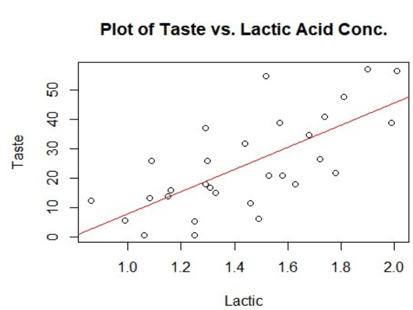


Based on this, the slope is significantly different from 0. Based on the stemplot and the QQ- plot, the residuals may be slightly skewed right, but do not differ too greatly from a normal distribution. The scatterplot of the residuals shows weak positive associations between the residuals and both Acetic and Lactic Acid. The plot of the residuals against H2S suggest greater scatter in the residuals for large H2S values.

Our third simple linear regression model treats Taste as a response variable and Lactic as an explanatory variable.

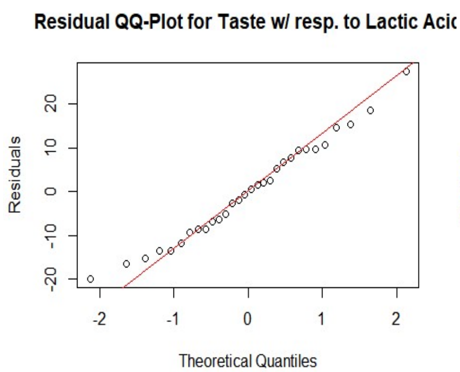
```
## Call:
## lm(formula = chz$taste ~ chz$lactic, data = chz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9439  -8.6839  -0.1095   8.9998  27.4245
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -29.859     10.582  -2.822  0.00869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

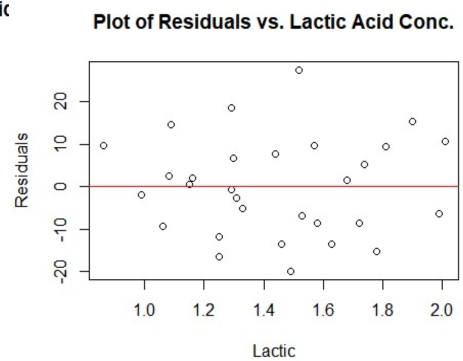
The regression equation is: $\widehat{Taste} = -29.86 + 37.720\ H2S$. In terms of summary statistics, the model has a standard deviation $s = 11.75$ and a coefficient of determination $R^2 = 0.496$. For the significance testing on regression parameter β_1, the testing statistic and corresponding P-value are given, respectively $(t = 5.25, P = 1.41 \times 10^{-5} \ll 0.05)$. To represent this linear relationship, the scatterplot below depicts the plot of Taste with respect to Lactic concentration.

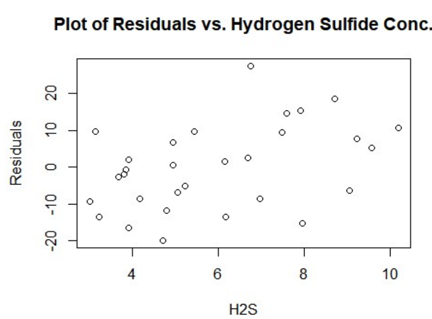


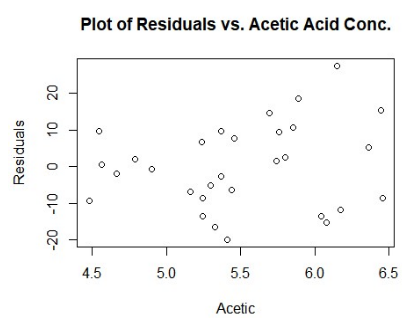
To legitimize the normality of the residuals and affirm our assumptions, the following plots are computed below.

Stem Plot of Residuals
-2 | 0
-1 | 75
-1 | 442
-0 | 999765
-0 | 321
0 | 0123
0 | 5789
1 | 001
1 | 559
2 |
2 | 7









Based on this, the slope is significantly different from 0. Based on the stemplot and the linear QQ-plot, the residuals appear to be roughly Normal. The scatterplots do not reveal any striking patterns or curves for the residuals vs. Acetic and H2S that would violate any assumptions.

To further analyze how these linear models compare, a table featuring all simple linear models for Taste with respect to each chemical as explanatory variables is shown below.

	Taste	F	P	R^2	s
Acetic	-61.50 + 15.648x	12.11	0.00166	30.2%	13.82
H2S	-9.787 + 5.7761x	37.29	1.37×10^{-6}	57.1%	10.83
Lactic	-29.86 + 37.720x	27.55	1.41×10^{-5}	49.6%	11.75

The intercepts differ from model to model given that they represent different explanatory variables, representing a number where the specific explanatory being analyzed equals 0 with respect to the response variable Taste.

Now, multiple linear regression analysis can begin. First, we Taste with respect to Acetic and Hydrogen Sulfide as the correlation between them was moderate in comparison to the other two-variable correlations. Again, since the regression coefficients β_i , for $i = 1, \ldots, p$,are obtained by the complicated method of least squares, RStudio software was administered.

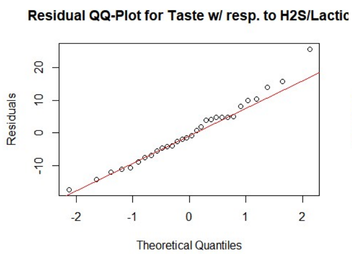
```
## Call:
## lm(formula = chz$taste ~ chz$acetic + chz$h2s, data = chz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.113  -6.893  -1.673   6.592  23.715
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -26.940     21.194  -1.271 0.214536
## chz$acetic     3.801      4.505   0.844 0.406245
## chz$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
```

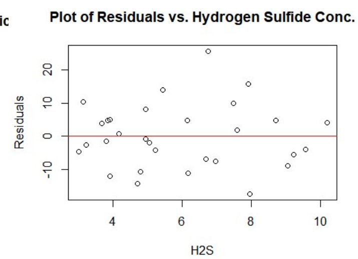
The regression equation is: $\widehat{Taste} = -26.94 + 3.801\,Acetic + 5.146\,H2S$ with standard deviation $s = 10.89$ and coefficient of determination of $R^2 = 0.582$. For the hypothesis testing on the regression parameters β_i, the testing statistic for the coefficient of Acetic and corresponding P-value respectively are $(t = 0.84, P = 0.406 \gg 0.05)$, thus indicating the covariate Acetic is not significant in predicting the response variable Taste. This can also be noted given that Acetic Acid and H2S are highly correlated $(r = 0.6179559)$, exhibiting multicollinearity which complicated the interpretation of the regression equation. Although this model does better than any of the three simple linear regression models, it is not significantly better than H2S model alone, as 58.2 is not much larger than 57.1, as expected given the t-test on Acetic.

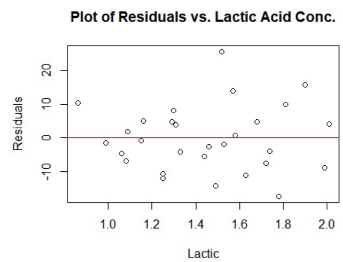
The second multiple linear regression model examines Taste with respect to Hydrogen Sulfide and Lactic Acid, executed by RStudio software below.

```
## Call:
## lm(formula = chz$taste ~ chz$lactic + chz$h2s, data = chz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.343  -6.530  -1.164   4.844  25.618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.592      8.982  -3.072  0.00481 **
## chz$lactic    19.887      7.959   2.499  0.01885 *
## chz$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 regression equation is: $\widehat{Taste} = -27.592 + 19.887\ Lactic + 3.946\ H2S$. The coefficient of determination $R^2 = 0.652$ is much larger than the R^2 of each respective explanatory variable alone. Furthermore, both coefficients are significantly different from 0 given P-values $(0.002 < 0.05)$ and $(0.019 < 0.05)$ for H2S and Lactic respectively. Given the QQ-plots and plot of residuals below, we can affirm that the residuals are normally distributed with this linear model.



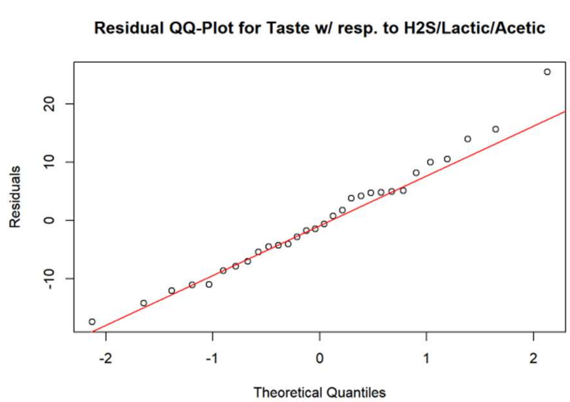


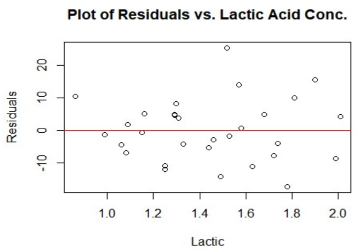


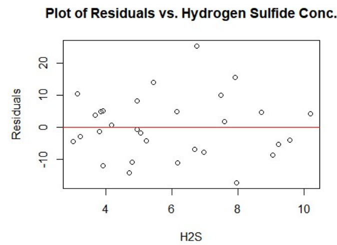
Although Acetic proved to be insignificant in the regression model with Hydrogen Sulfide, that does not conclude that it would not be a good predictor for Taste when accompanied with Hydrogen Sulfide and Lactic. It is to be assumed in multiple linear regression that the significance tests for individual regression coefficients assess the significance of each predictor variable assuming that all other predictors are included in the regression equation. Hence, the third linear model including all explanatory variables, is shown below.

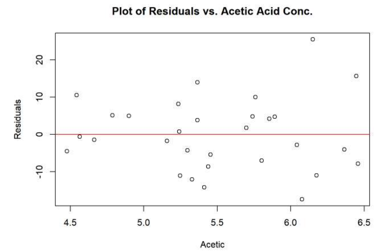
```
## Call:
## lm(formula = chz$taste ~ chz$lactic + chz$h2s + chz$acetic, data = chz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## chz$lactic   19.6705     8.6291   2.280  0.03108 *
## chz$h2s       3.9118     1.2484   3.133  0.00425 **
## chz$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
```

Like simple linear regression, it is assumed that the residuals are Normal, and this can be affirmed with a normal quantile plot and plot of residuals versus other explanatory variables.









The regression equation is: $\widehat{Taste} = -28.877 + 19.6705\ Lactic + 3.9118\ H2S + 0.3277\ Acetic$. The coefficient of determination R^2 remains 0.652. The residuals appear to be Normally distributed, given the approximately linear QQ-plot and show no particular pattern or curve in the scatterplot with the explanatory variables. There is no gain in adding Acetic to the regression model given the hypothesis t-test on the parameter, where $t = 0.07$ and the P-value $= \ 0.942 \ \gg \ 0.05$.

Concluding Remarks

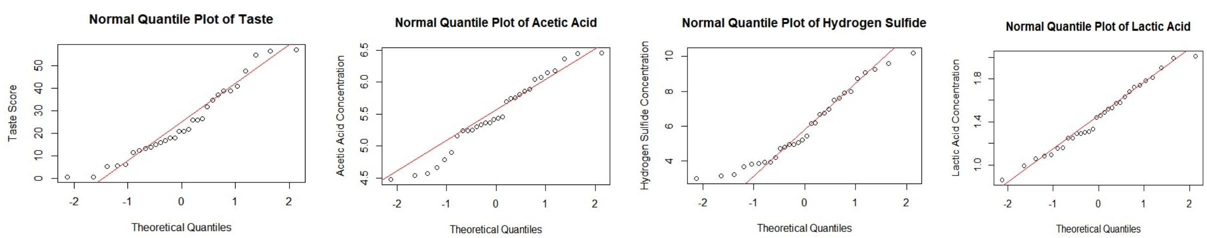
The best model would have to be the Taste with respect to H2S and Lactic, where $\widehat{Taste} = -27.592 + 19.887\ Lactic + 3.946\ H2S$. As explained previously, adding Acetic to the multiple linear regression model proves to be insignificant, given the large P-value in the RStudio computation. The insignificance of Acetic can also be noted given that model explains roughly 65.2% of the variation in Taste with or without the coefficient of Acetic. Despite this, about 34.8% percent of the variation in Taste is not explained by the concentrations of Lactic Acid and Hydrogen Sulfide. To explain this remaining variation, a larger dataset would be sufficient or other chemical compounds as explanatory variables. We also see a decrease in the standard deviation of the model, ($9.942 \ < \ 10.13$), which indicated the approximate magnitude of the errors you can expect when attempting to quantify Taste with the regression equation. I believe that this is the strongest model that will accurately quantify the taste of matured cheese when the concentrations of Hydrogen Sulfide and Lactic Acid are considered.

Section 2

(Homework Questions)

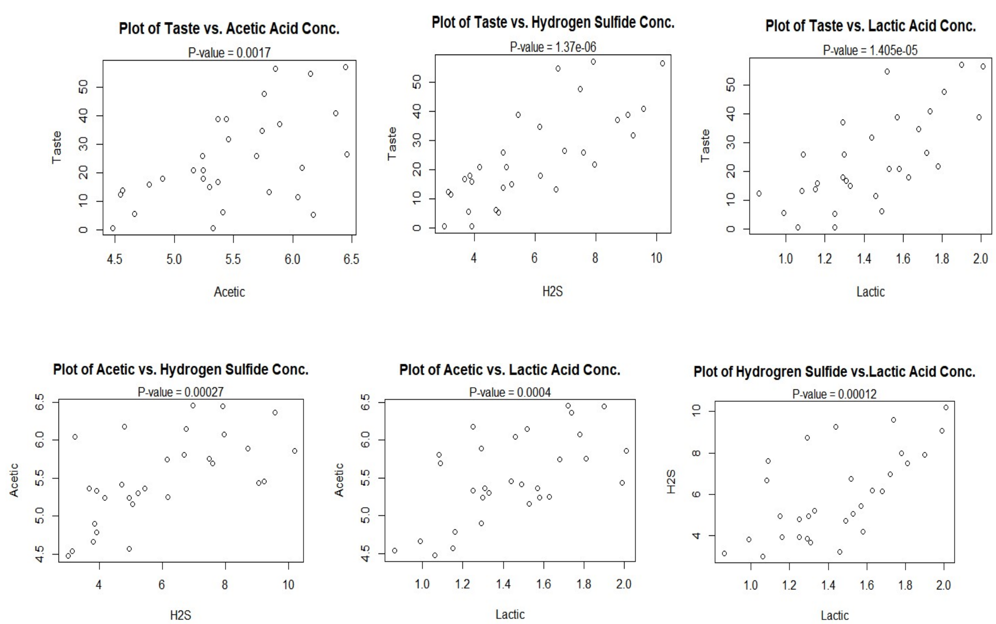
(11.53). For the variables Taste, Acetic, H2S and Lactic, the sample means are: 24.533, 5.498, 5.941 and 1.422 respectively. The medians are: 20.950, 5.425, 5.329 and 1.450 respectively. The standard deviations are: 16.255, 0.570, 2.127 and 0.303 respectively. The interquartile ranges are: 23.150, 0.645, 3.597 and 0.417 respectively.

Taste	Acetic	H2S	Lactic
## 0 | 11666 ## 1 | 223456788 ## 2 | 112667 ## 3 | 25799 ## 4 | 18 ## 5 | 577	## 44 | 846 ## 46 | 69 ## 48 | 0 ## 50 | 6 ## 52 | 4450377 ## 54 | 146 ## 56 | 046 ## 58 | 069 ## 60 | 4858 ## 62 | 7 ## 64 | 56	## 2 | ## 3 | 01278999 ## 4 | 27899 ## 5 | 024 ## 6 | 1278 ## 7 | 0569 ## 8 | 07 ## 9 | 126 ## 10 | 2	## 8 | 69 ## 10 | 68956 ## 12 | 5599013 ## 14 | 4692378 ## 16 | 38248 ## 18 | 109 ## 20 | 1



The variables exhibit minimal deviations from the Normality and the QQ-plot appears to be linear. The variables, Taste and H2S show a slight right skew and Acetic has two peaks, as shown in the stem plot. There do not appear to be any extreme outliers in any of the variables analyzed.

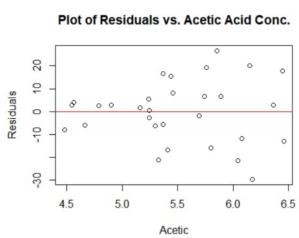
(11.54). Using the t-test where H_o : $\rho = 0$ and H_a : $\rho \neq 0$, with the testing statistic $t = \frac{r(X,y)\sqrt{n-2}}{t\sqrt{1-r^2}}$, the respective P-values are shown above each scatterplot of the two variables. All the plots show a positive correlation between the variables. All correlations are thus positive, and the corresponding P-values are significantly different from 0. The p values are correct given that the two variables under investigation are normally distributed

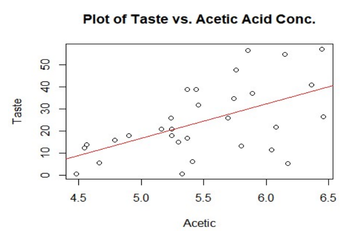


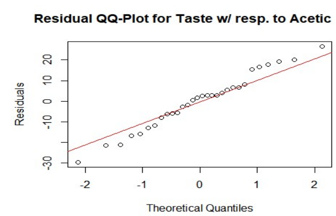
Correlation Table	Taste	Acetic	H2S	Lactic
Taste	1.000000	0.549539	0.755752	0.704236
Acetic	0.549539	1.000000	0.617956	0.603783
H2S	0.755752	0.617956	1.000000	0.644812
Lactic	0.704236	0.603782	0.644812	1.000000

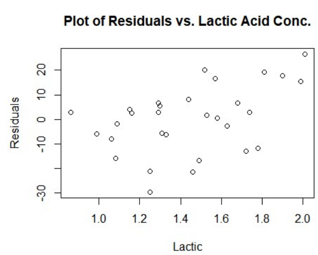
(11.55). The regression equation is $\widehat{Taste} = -61.50 + 15.648\ Acetic$, with a standard deviation of s = 13.82 and a coefficient of determination $R^2 = 0.302$. For the significance testing on regression parameter β_1 the testing statistic and corresponding P-value are given, respectively $(\mathrm{t} = 3.48, \mathrm{P} = 0.00166)$. Based on this, the slope is significantly different from 0. Based on the stemplot and the QQ-plot, the residuals appear to have a Normal distribution. The scatterplot of residuals exhibit a positive association when in respect to H2S and Lactic Acid. The plot of residuals show a greater scatter in the residuals for the large Acetic values.

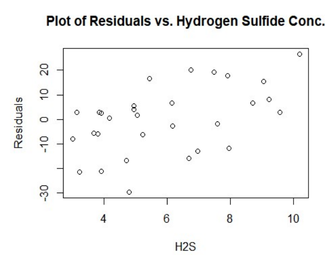
Stem Plot of Residuals
-3 | 0
-2 | 11
-1 | 7632
-0 | 866632
0 | 02233345778
1 | 5789
2 | 07







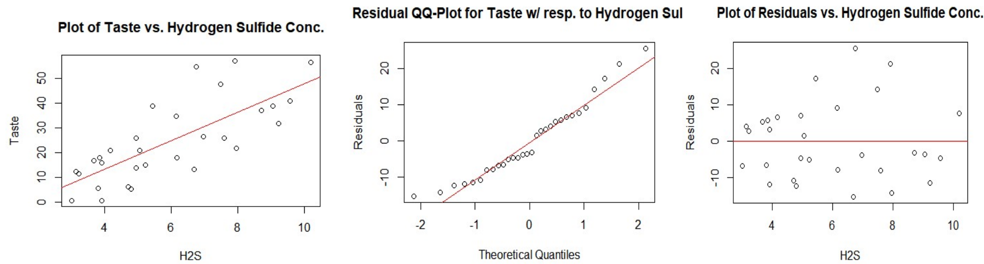


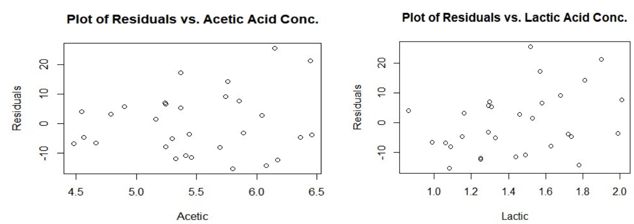


(11.56). The regression equation is $\widehat{Taste} = -9.787 + 5.7761\ H2S$, with a standard deviation of s = 10.83 and coefficient of determination $R^2 = 0.571$. For the significance testing on regression parameter β_1, the testing statistic and corresponding P-value are given, respectively $(\mathrm{t} = 6.11, \mathrm{P} = 1.37 \times 10^{-6} \ll 0.05)$. Based on this, the slope is significantly different from 0. Based on the stemplot and the QQ- plot, the residuals may be slightly skewed right, but can still be considered normal distribution. The scatterplot of the residuals show weak positive associations between the residuals and both Acetic and Lactic Acid. The plot of the residuals against H2S shows greater scatter in the residuals for large H2S values.

Stem Plot of Residuals
-1 | 5
-1 | 42221
-0 | 8877555
-0 | 443
0 | 2334
0 | 567789
1 | 4
1 | 7

```
## 2 | 1
## 2 | 6
```

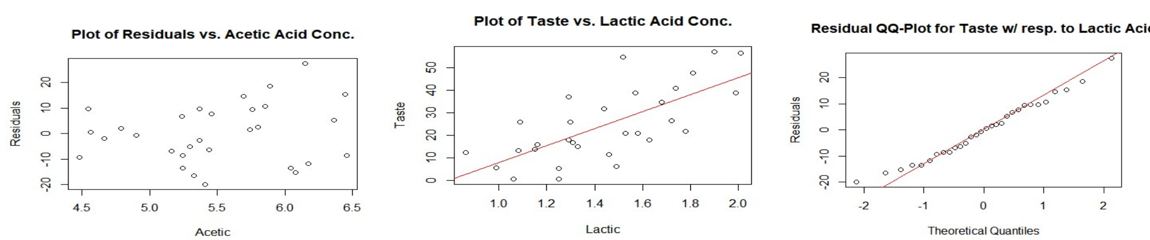


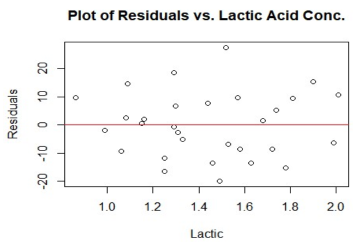


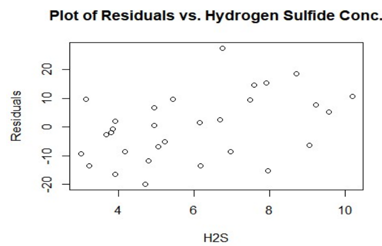
(11.57). The regression equation is $\widehat{Taste} = -29.86 + 37.720\ Lactic$ with a standard deviation s = 11.75 and a coefficient of determination $R^2 = 0.496$. For the significance testing on regression parameter β_1 , the testing statistic and corresponding P-value are given, respectively $(t = 5.25, P = 1.41 \times 10^{-5} \ll 0.05)$. Based on this, the slope is significantly different from 0. Based on the stemplot and the QQ-plot, the residuals appear to be roughly Normal. The residual scatterplots do not reveal any striking patterns or associations.

Stem Plot of Residuals

```
## -2 | 0
## -1 | 75
## -1 | 442
## -0 | 999765
## -0 | 321
##  0 | 0123
##  0 | 5789
##  1 | 001
##  1 | 559
##  2 |
##  2 | 7
```







(11.58).

	Taste	F	P	R^2	s
Acetic	-61.50 + 15.648x	12.11	0.00166	30.2%	13.82
H2S	-9.787 + 5.7761x	37.29	1.37×10^{-6}	57.1%	10.83
Lactic	-29.86 + 37.720x	27.55	1.41×10^{-5}	49.6%	11.75

The intercepts differ from model to model given that they represent different explanatory variables, representing a number where the specific explanatory being analyzed equals 0 with respect to the response variable Taste.

(11.59). The regression equation is: $\widehat{Taste} = -26.94 + 3.801\ Acetic + 5.146\ H2S$ with standard deviation s = 10.89 and coefficient of determination of $R^2 = 0.582$. For the hypothesis testing on the regression parameters, the testing static for the coefficient of Acetic and corresponding P-value respectively are ($\mathrm{t} = 0.84, \mathrm{P} = 0.406 \gg 0.05$) thus indicating the covariate Acetic is not significant in predicting the response variable Taste. This can also be noted given that Acetic Acid and H2S are highly correlated (r = 0.6179559). Although this model does better than any of the three simple linear regression models, but is not significantly better than H2S model alone, as 58.2% is not much larger than 57.1%, as expected given the t-test on Acetic.

(11.60). The regression equation is: $\widehat{Taste} = -27.592 + 19.887\ Lactic + 3.946\ H2S$. The coefficient of determination $R^2 = 0.652$ is much larger than the R^2 of each respective explanatory variable alone. Furthermore, both coefficients are significantly different from 0 given P-values (0.002 < 0.05) and (0.019 < 0.05) for H2S and Lactic respectively.

(11.61). The regression equation is: $\widehat{Taste} = -28.877 + 19.6705\ Lactic + 3.9118\ H2S + 0.3277\ Acetic$. The coefficient of determination R^2 remains 0.652. The residuals suggest normal distribution and show no curve in the scatterplot with the explanatory variables. Acetic is insignificant to the regression model given the hypothesis t-test on the parameter, where t = 0.07 and the P-value = 0.942 >> 0.05. Therefore, the best model would have to be the Taste with respect to H2S **and** Lactic, where $\widehat{Taste} = -27.592 + 19.887\ Lactic + 3.946\ H2S$.