MA 331 / Final Project Report / Theodore Jagodits / 5/7/20

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

Introduction:

This study looks at the effects of acetic acid content, hydrogen sulfide content and lactic acid content in cheddar cheese from LaTrobe Austria. Thirty samples were taken and sampled for the chemicals mentioned above. Then they were given a score on a taste test. The score is made from several different tasters. This study is used to determines which variables and combination of them to predict the score value as best as possible. The data is given in an .xls file and it was used in R.

Tools:

We use the .xls file given to us and use statistical analysis in R and some graphs in excel. The rest of the details are in the R code file.

Summary:

The variables are first given a statistical summary to learn a little bit more about them. I looked at mean, median, standard deviation and Interquartile Range. I also check for any outliers, which there are not, and the data seems to normally distributed. Then I performed a correlation test on the different variables on each other. H2S and Acetic are strongly correlated, since the zero-point correlation test gave me a very low p-value, so we reject the null hypothesis that the correlations are zero.

Using the equation y i = β0 +β1 i +εi for to model (x1 … xN) for taste (y1 … yN). I propose that the null hypothesis for each case is slope 0 and the alternative hypothesis is that the slope has some significant correlation model for taste. The first model is y = -61.5 + 15.55(acetic content) and the p-value is 0.0017, which is very low. Which means acetic content has a significant impact on the taste of the cheese. On the second model, y = -9.78 - +5.77(H2S content) for which the p-value was almost zero, so H2S also had a significant impact on taste. For Lactic, the equation y = -29.85 + 37.72(Lactic Content) and the p-value was negligible as well. All the cases had approximately normally distributed residuals.

Next, I tried a generic model that used a linear combination of the variables. The first model I tried was y = -26.94 + 3.5(Acetic content) + 5.15(H2S content). The p-value for the Acetic content was .4 and the H2S was 0.0002 so we could not reject the null hypothesis for this model because of the Acetic content. The second model was Lactic and H2S, y = -27.59 + 19.89(Lactic content) + 3.95(H2S content). The p-values where both very low (0.018 and 0.0017 respectively), so these two coefficients were the best in this study so we select them as our best model.

Finally I tried a model with all variables that ended up y = -28.877 + 0.32(Acetic Content) + 3.9(H2S) + 19.671(Lactic). The p-values respectively were .94, 0.0042 and 0.031. Acetic was deemed not significant so I removed it from the model and so the model was based on the previous exercise.

11.53

Descriptive Statistics

	Taste	Acetic	H2S	Lactic
Mean	24.5	5.498	5.942	1.44
Median	21.0	5.425	5.329	1.45
Std. Dev.	16.25538	0.570878	2.126879	0.30349
IQR	23.15			

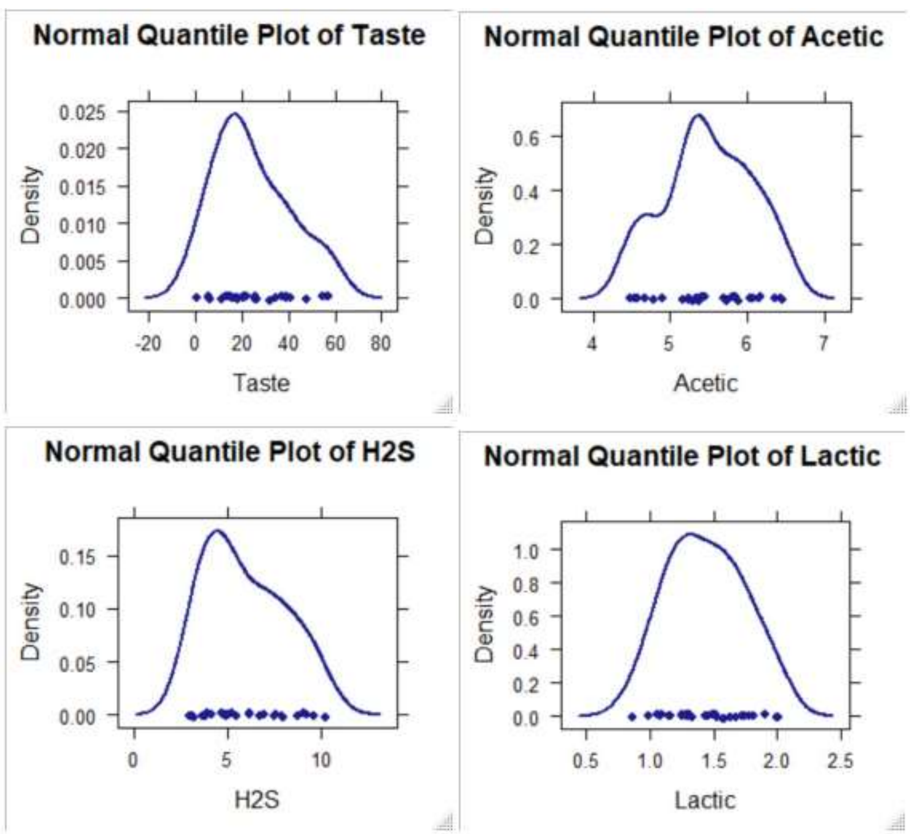
Stem plots of the Response and Explanatory Variables

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

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

```
H2S
  2 |
  3 | 01278999
  4 | 27899
  5 | 024
  6 | 1278
  7 | 0569
  8 | 07
  9 | 126
 10 | 2
```

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

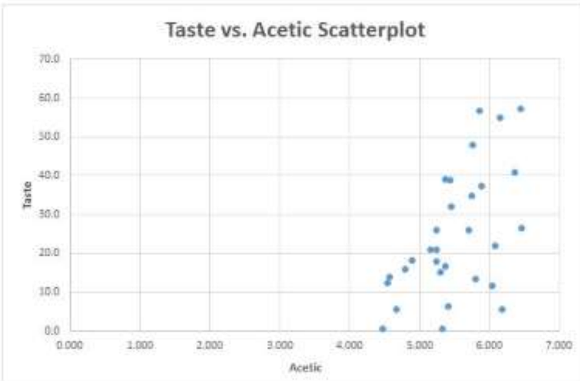


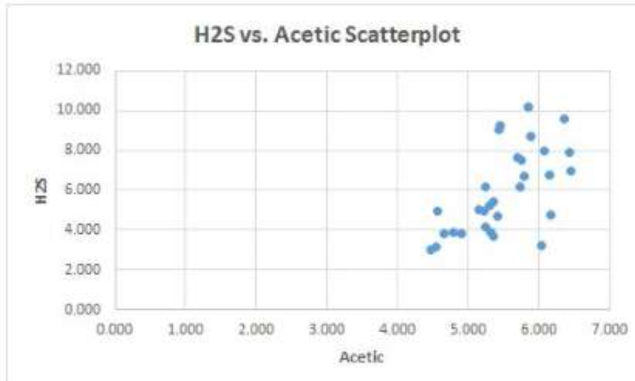
Normal Quantile Plots

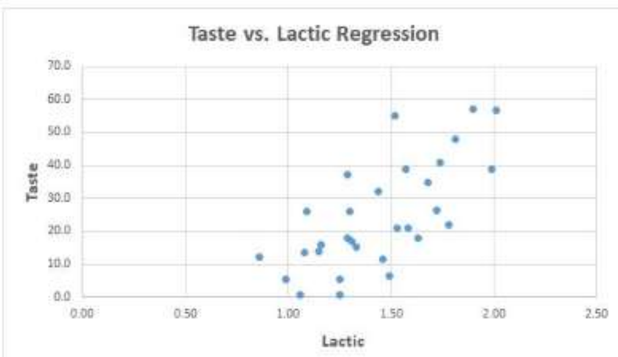
All the variables look approximately normal and have no obvious outliers. Lactic looks the most normal, then Taste with a slight skew. Acetic and H2S both have more of a skew but nothing out of the ordinary. We can do linear regression.

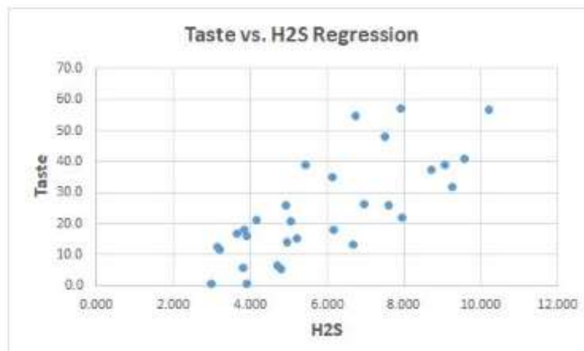
11.54

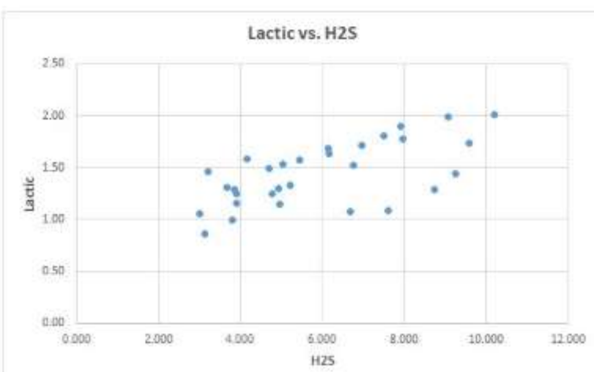
Scatterplot of Each Variable

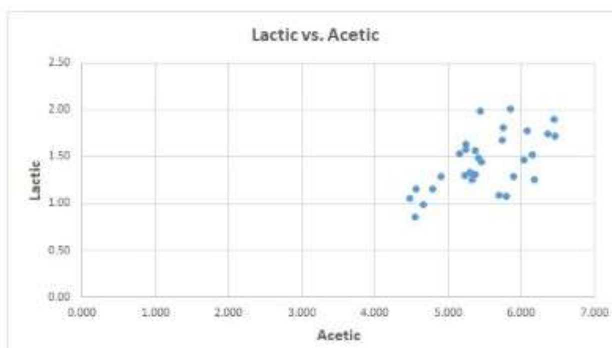












Correlation Matrix				
	taste	acetic	h2s	lactic
taste	1	0.55	0.756	0.704
acetic	0.55	1	0.618	0.604
h2s	0.756	0.618	1	0.645
lactic	0.704	0.604	0.645	1

<u>Pvalue Test of Zero Population Correlation Test</u>

Taste vs Acetic = 0.002

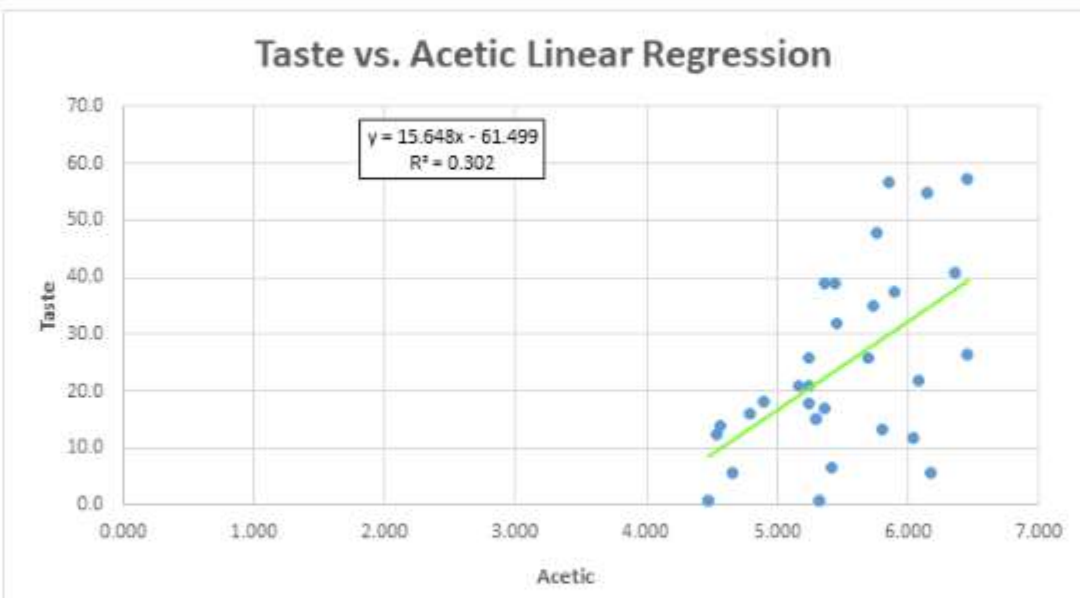
Taste vs H2s = 1e-06

Taste vs Lactic = 1e-05

From the correlation matrix above we can see that Taste vs. Acetic is 0.55, Taste vs. H2S is 0.756 and Taste vs. Lactic is 0.704. The strongest relationship of them is Taste vs. H2S. The p-values above indicate that we can reject the null hypothesis for all the variables vs. Taste is 0.

<u>11.55</u>

<u>Linear Regression on Taste vs. Acetic</u>



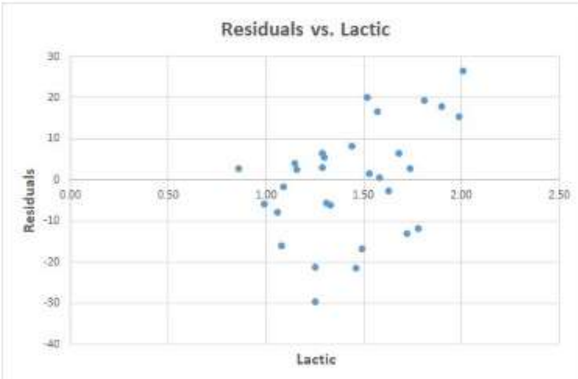
The model shows Taste = -61.5 + 15.55(Acetic Content)

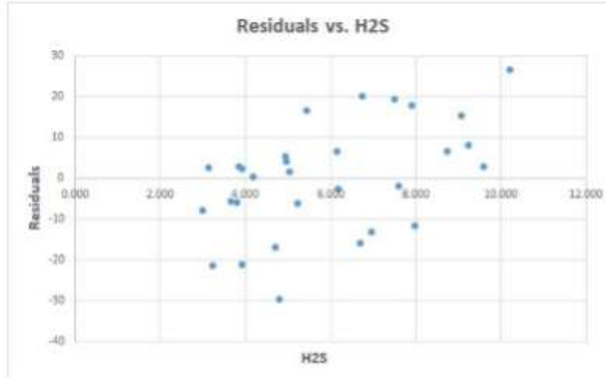
T-Stat: 3.48

P-Value: 0.0017

Confidence Interval (2.5% - 97.5%): Intercept [-112.39, -10.6]; Acetic [6.44,24.9]

<u>Residuals vs. Explanatory Variables</u>



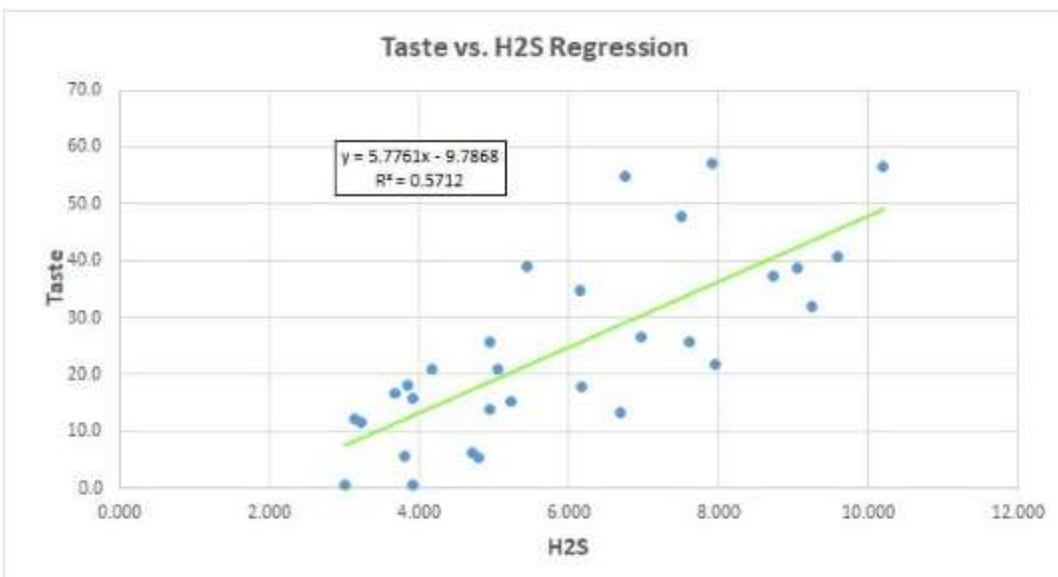


Analysis:

Both residuals appear to be normally distributed and look like they are positively associated with H2S and Lactic Content. The P-Value was very low so there is a relationship between Taste and Acetic Content.

<u>11.56</u>

<u>Linear Regression on Taste vs. H2S</u>



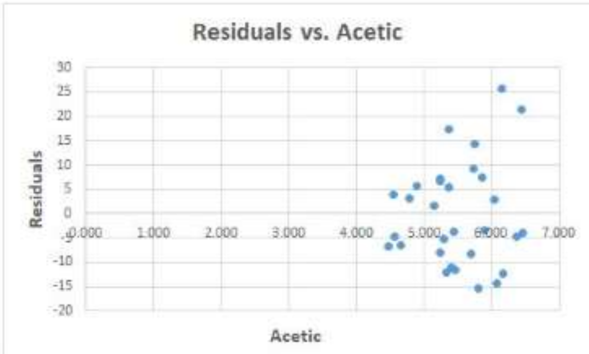
Taste = -9.78 + 5.77(H2S Content)

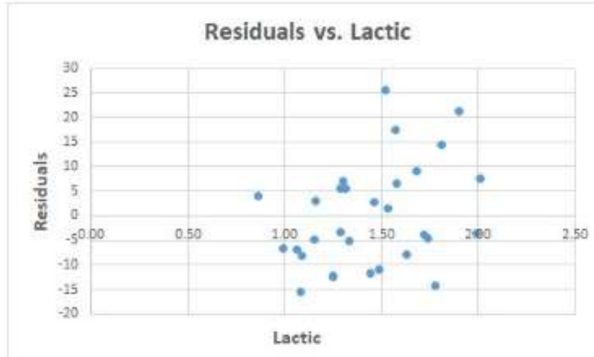
T-Stat: 6.107

P-Value: approx. 0

Confidence Interval 97.5: Intercept [-21.99,2.42]; H2S [3.84,7.71]

<u>Residuals Graph</u>



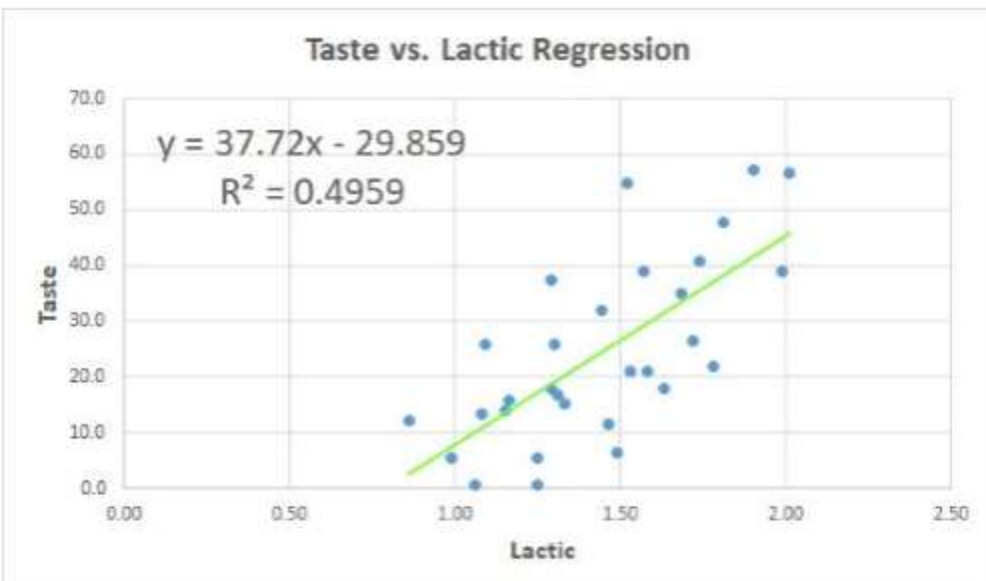


Taste vs. H2S Linear Regression Analysis

The residuals look to have an approximately normal distribution and there no clear patterns so random, between the residuals of the two other comparisons. The P-Value appears to be close to zero, so we can confidently say that there is a relationship between the variables taste and H2S.

11.57

Linear Regression of Taste vs. Lactic



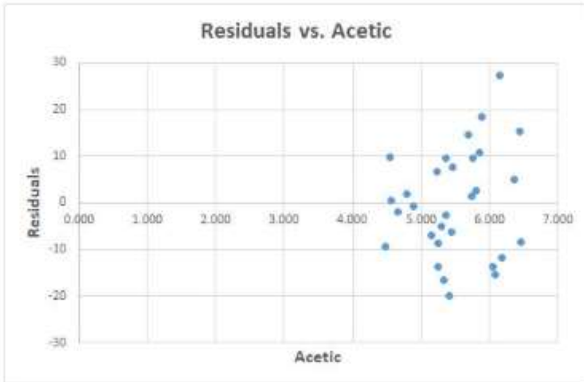
Taste = -29.86 + 37.72(Lactic Content)

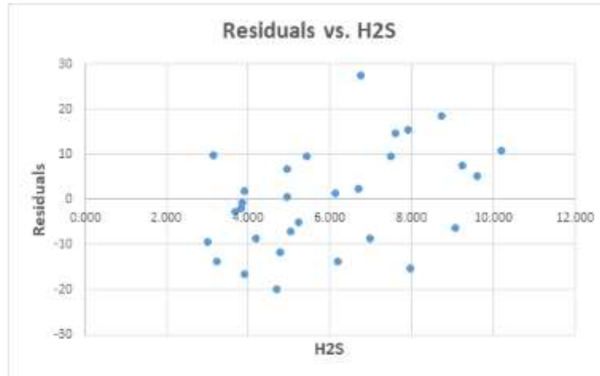
T-Stat: 5.249

P-Value: Approx. 0

Confidence Interval 97.5%: Intercept [-51.5,-8.18]; Lactic Content [23, 52.44]

Regression of residuals vs. remaining Variables





Taste vs. Lactic Analysis

The residuals look to have a normal distribution and there is no clear pattern between the residuals and the other variables. The p-value for the test is almost zero, so we can conclude there is a relationship between Taste and Lactic.

11.58

	Taste vs. Acetic	Taste vs. H2S	Taste vs. Lactic
F	12.1	37.3	27.5
P-Value	0.0017	Approx. 0	Approx. 0
R-squared	10.32	10.83	11.75
Formula	Taste = -61.5 + 15.55(Acetic Content)	Taste = -9.78 + 5.77(H2S Content)	Taste = -29.86 + 37.72(Lactic Content)

Looking at the values in the table above we can see that H2S was the most likely to predict taste and Acetic was the least likely. Each model assumes that only that chemical was involved in making it taste well. However, that is usually not the case as different chemicals blend with each other to make us taste. I found that the study took log transformations acetic and H2S content, it might be difficult to compare them.

11.59

Using Acetic and H2S content as the variables to describe taste, it gives the following formula.

Taste = 3.8(Acetic Content) + 5.15(H2S Content) – 26.94

Residuals

Min	1st Quartile	Median	3rd Quartile	Max
-16.11	-6.89	-1.67	6.59	23.7

Coefficients

	Estimated	Std. Error	t-value	Pr(>|t|)
Intercept	-26.94	21.19	-1.27	0.21

Acetic	3.8	4.51	0.84	0.406
H2S	5.15	1.21	4.26	0.0002

Residual Standard Error: 10.9 with 27 degrees of freedom

Multiple R-squared: 0.582

Adjusted R-squared: 0.551

F-stat: 18.8 on 2 and 27 d.o.f.

p-value: approx. 0

Confidence Interval

	2.5 %	97.5 %
Intercept	-70.43	16.55
Acetic	-5.44	13.05
H2S	2.66	7.63

Analysis of Variance

	Df	Sum square	Mean Square	F value	Pr(>F)
Acetic	1	2314	2314	19.5	0.00015
H2S	1	2147	2147	18.1	0.00022
Residuals	27	3202	119	-	-

After looking at the P-values of acetic, H2s and lactic. I conclude that for multiple values acetic is not fit for the formula because of the very high p-value(.406) meaning it does not correlate well with the other variables. Acetic and Lactic together correlate very well on their own but not with Lactic.

11.60

The new formula to predict taste: Taste=19.89(Lactic Content) + 3.95(H2S) – 27.59

Residuals

Min	1st Quartile	Median	3rd Quartile	Max
-17.34	-6.53	-1.16	4.84	25.62

Coefficients

	Estimated	Std. Error	t-value	Pr(>|t|)
Intercept	-27.94	8.98	-3.07	0.0048
Lactic	19.89	7.96	2.5	0.0188
H2S	3.95	1.14	3.47	0.0017

Confidence Interval

	2.5 %	97.5 %
Intercept	-46	-9.16
Lactic	3.56	36.22
H2S	1.62	6.28

Analysis of Variance

	Df	Sum square	Mean Square	F value	Pr(>F)
Lactic	1	3800	3800	38.5	Approx. 0
H2S	1	1194	1194	12.1	0.0017
Residuals	27	2669	99	-	-

Residual Standard Error: 9.94 with 27 degrees of freedom

Multiple R-squared: 0.653

Adjusted R-squared: 0.626

F-stat: 25.3 with 2 and 27 d.f.

p-value: approx. 0

Analysis

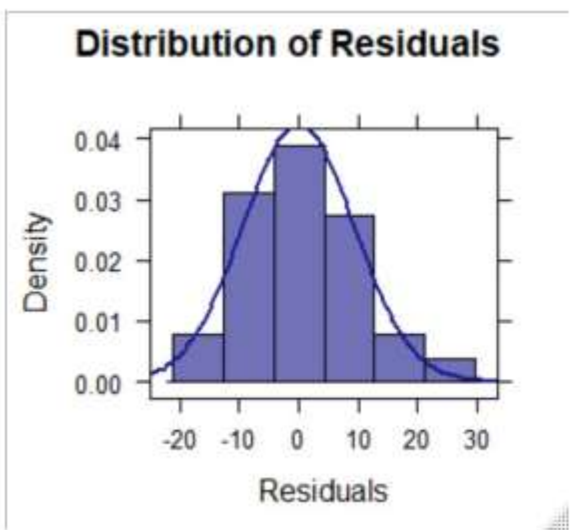
Looking the values in the tables, we can see that lactic and h2s both have a significant impact on the model. The correlation was 0.645 which is a very strong correlation. But also, the variables still have a statistical significance compared to the other problem. The residuals are approximately normal.

11.61

The final model uses all 3 variables in the formula. Taste = 0.328(acetic content) + 3.912(H2S content) + 19.671(Lactic content) – 28.877.

Residuals

Min	1st Quartile	Median	3rd Quartile	Max
-17.39	-6.61	-1.01	4.91	25.45



Coefficients

	Estimated	Std. Error	t-value	Pr(>|t|)

Intercept	-28.87	19.735	-1.46	0.155
Acetic	0.28	4.46	0.07	0.942
H2S	3.912	1.24	3.13	0.0042
Lactic	19.671	8.629	2.28	0.0311

Residual Standard Error: 10.1 with 26 degrees of freedom.

Multiple R-squared: 0.652

Adjusted R-squared: 0.612

F-stat: 16.2 with 3 and 26 d.f.

p-value: approx. 0

Confidence Interval

	2.5 %	97.5 %
Intercept	-69.44	11.69
Acetic	-8.84	9.49
H2S	1.35	6.48
Lactic	1.93	37.41

Analysis of Variance

	Df	Sum square	Mean Square	F value	Pr(>F)
Acetic	1	2314	2314	22.6	Approx. 0
Lactic	1	533	533	5.2	0.0311
H2S	1	2147	2147	20.9	0.0017
Residuals	26	2668	103	-	-

In the model, the p-values for acetic, h2s and lactic are 0.942, 0.0042 and 0.0311 respectively. The residuals are approximately normally distributed. The p-value for acetic is high so that means that the variable is not significant for the model. We would take out the acetic variable as the next thing and just to multiple regression on the remaining two variables. Which I already showed in the previous problem. So, the best model would be Taste = 19.89(Lactic)+3.95(H2S)-27.59 which best describes taste. This model has a lower error than the 3 variable model (9.94 vs 10.1) and also having the same multiple-R-squared value (0.652). The plots look normal, so we do not need to perform polynomial regression.