MA 331 Final Project Report

By: Erika McCarthy & April Ullrich

Introduction

In this study, we have observed the effect of acetic acid content (Acetic), hydrogen sulfide content (H2S), and lactic acid content (Lactic) on the taste of cheddar cheese from the LaTrobe Valley of Victoria, Austria. 30 cheese samples were taken and analyzed for the content of each of the three chemicals, and each sample was given a score based on a taste test. There were 30 samples taken, and each taste score is a composite score from several tasters. This study has sought to build regression models to determine which variable, or combination of variables, best predicts the taste of this cheese. The data was given as an excel file (titled cheesy.xls in the R code) and it was imported into R.

Software

The statistical analysis was done using R, and some graphs were made in excel. The "Mosaic" package was installed. The call "summary" was used to obtain general statics, "with(data, cor(subset))" was used for bivariate correlations, "lm" was used for linear models, "anova" was used to obtain anova tables, and plot(x, which=2) was used to obtain normal qq plots. See accompanying R code for further details.

Summary

Each variable was first analyzed for mean, median, standard deviation, outliers, normality, and correlation to each other. There are no outliers for any of the variables, and each variable is approximately normally distributed. H2S and Acetic are very strongly correlated, which is reflected in multiple linear regression using these variables. The zero-point correlation test for each variable yielded a low p-value, so the null hypothesis that the correlations are zero was rejected.

The simple linear regression models were done using the generic equation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for the set of explanatory variables $(x_1 ... x_n)$ for taste $(y_1 ... y_n)$. The null hypothesis in each case is that the slope, β_1, is zero, and the alternative is that the slope is significant the model to explain the taste. The first model was done between Taste and Acetic and the equation was Taste = -61.499 + 15.548 (Acetic) and the p-value was 0.0017, which means Acetic has a significant impact on Taste. For taste based on H2S, the equation was Taste = -9.7868 + 5.7761 (H2S) and the p-value was 1.37*10^-6 , so H2S has a significant influence on taste. For taste based on Lactic, the equation was Taste = -29.859 + 37.72 (Lactic) and the p-value was 1.41*10^-5, which means that Lactic has a significant influence on Taste. In all three cases, the residuals were approximately of normal distribution.

Three multiple linear regression models were conducted to explain taste using the generic formula $y_i = \beta_0 + \beta_1 x_i + ... + \beta_n x_n + \varepsilon_i$. The first model was taste explained by Acetic and H2S for which the equation was Taste = 3.80(Acetic) + 5.15(H2S) - 26.94. The p-value for Acetic was 0.406, and the p-value for H2S was 0.00022, so the null hypothesis could not be rejected for Acetic in this model. For taste explained by Lactic and H2S, the equation was Taste=19.89(Lactic)+3.95(H2S)-27.59. The p-value for lactic was 0.0188 and the p-value for H2S was 0.0017, so both coefficients were significant, which made this the best model in this study. The final model included all three explanatory variables with the equation

Taste=0.328(Acetic)+3.912(H2S)+19.671(Lactic)-28.877 and the p-values were 0.9420, 0.0042, and 0.0311 respectively. In this model, Acetic was not significant, so the next step is to remove it from the model, and this model was found previously. The next step in this study could be to find the multiple linear regression of taste based on Acetic and Lactic, however this was not required in the study.

Exhibit 1: Stemplots of 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
```

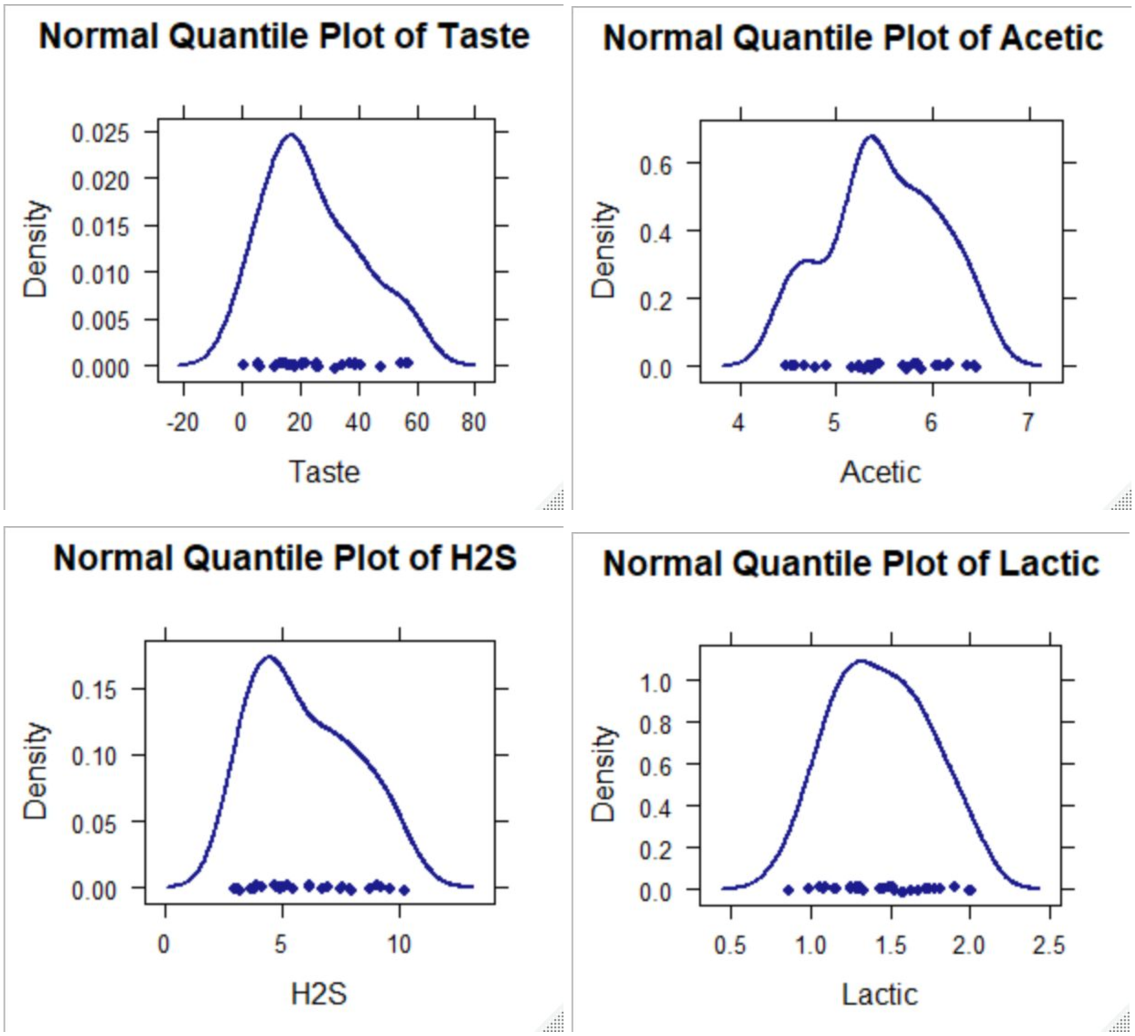
```
H2S
  2 |
  3 | 01278999
  4 | 27899
  5 | 024
  6 | 1278
  7 | 0569
  8 | 07
  9 | 126
 10 | 2
```

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

Exhibit 2: Descriptive Statistics for Response and Explanatory Variables

	Taste	**Acetic**	**h2s**	**Lactic**
Mean	24.53	5.50	5.94	1.44
Median	20.95	5.43	5.33	0.43
Std. Dev.	16.26	0.57	2.13	0.30
IQR	23.9	0.66	3.69	1.45

Exhibits 3-6: Normal Quantile Plots for Variables



Each of the variables have approximately normal distribution with no outliers. H2S and and Taste show right slight right skewness, and Acetic has a second minor peak. Based on these observations, it is safe to perform linear regression under the assumption of normal distribution.

Problem 11.54

Correlation Matrix

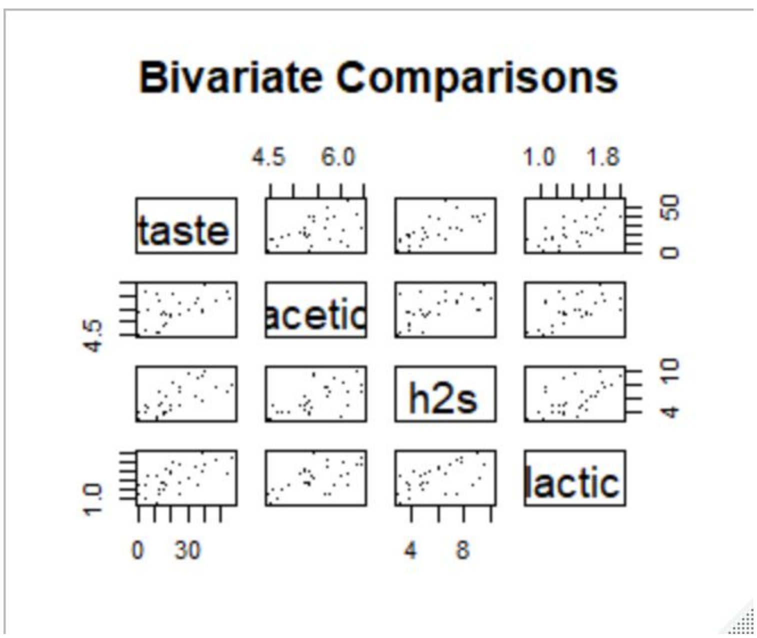
	taste	acetic	h2s	lactic
taste	1.000	0.550	0.756	0.704
acetic	0.550	1.000	0.618	0.604
h2s	0.756	0.618	1.000	0.645
lactic	0.704	0.604	0.645	1.000

Exhibit 7: Results of Zero Population Correlation Test

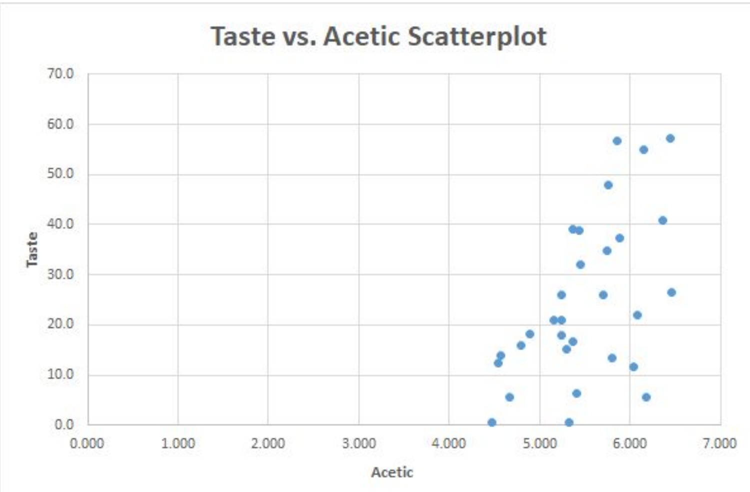
	Taste vs Acetic	Taste vs H2S	Taste vs Lactic
P-value for test of zero population correlation	0.002	1e-06	1e-05

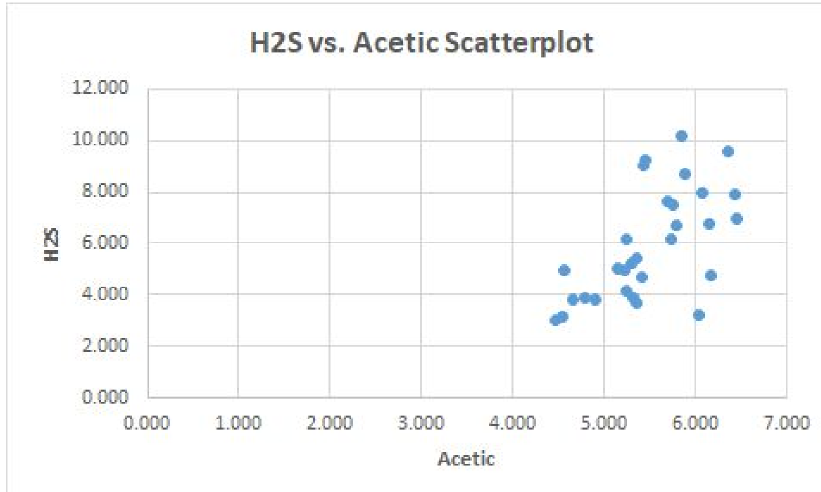
The correlation between Taste and Acetic is 0.550, taste and H2S is 0.756, and taste and lactic is 0.704. The strongest correlation is between taste and H2S, and the lowest p-value is for Taste and H2S. Based on the p-values, we can reject the null hypothesis that the correlation for each variable to Taste is zero. The graph of bivariate comparisons below shows the positive correlation of taste with each of the three variables. Individual scatter plots have also been provided for a clearer view of the relationships between the variables.

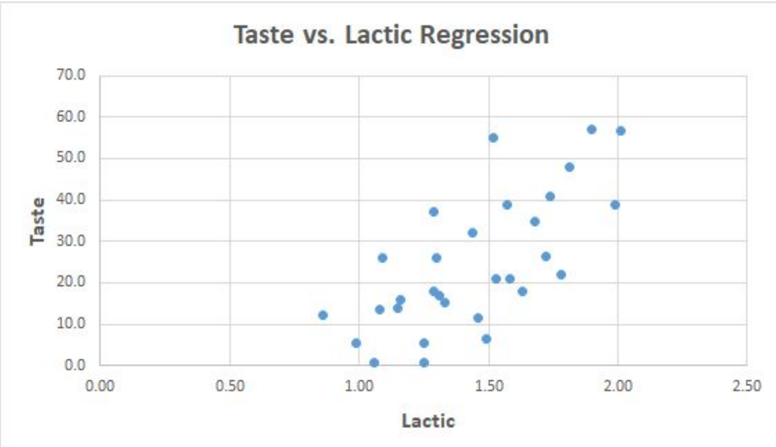
Exhibit 8: Bivariate Comparison Chart

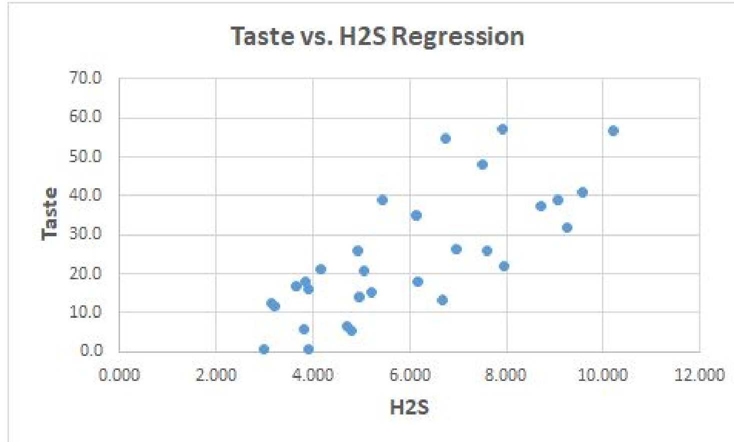


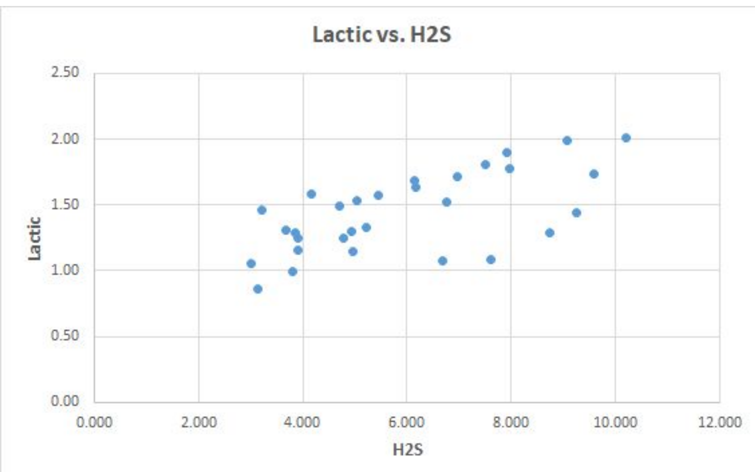
Exhibits 9-12: Scatterplots of Each Variable Pairing

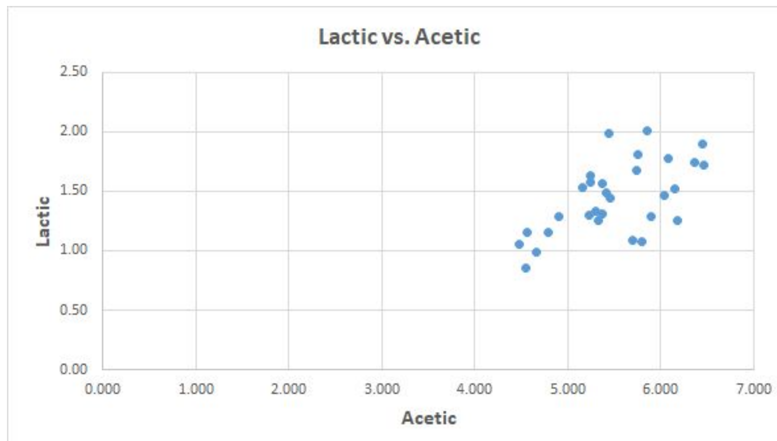






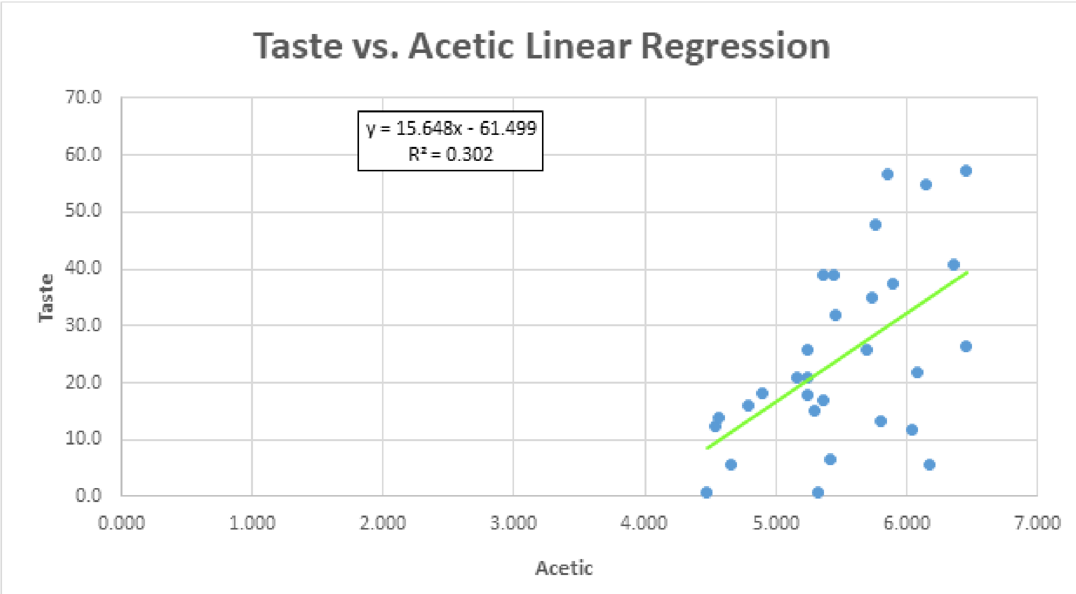






<u>Problem 11.55</u>

<u>Exhibit 13: Linear Regression Model of Dependence of Taste on Acetic</u>



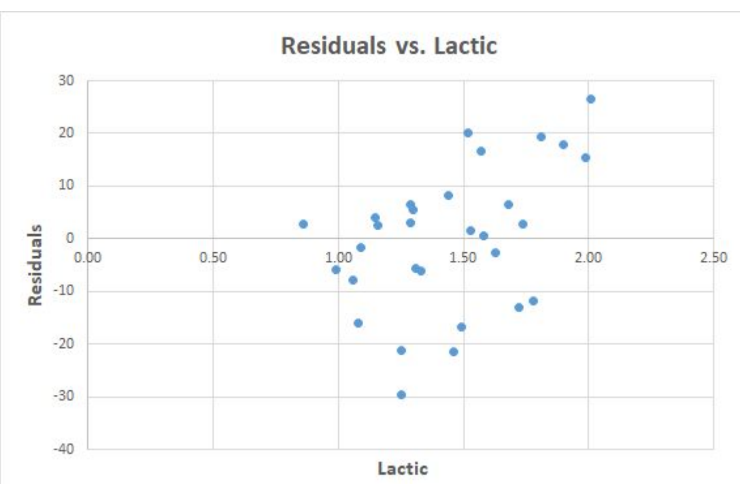
Linear Regression Model: Taste = -61.499 + 15.548 (Acetic)

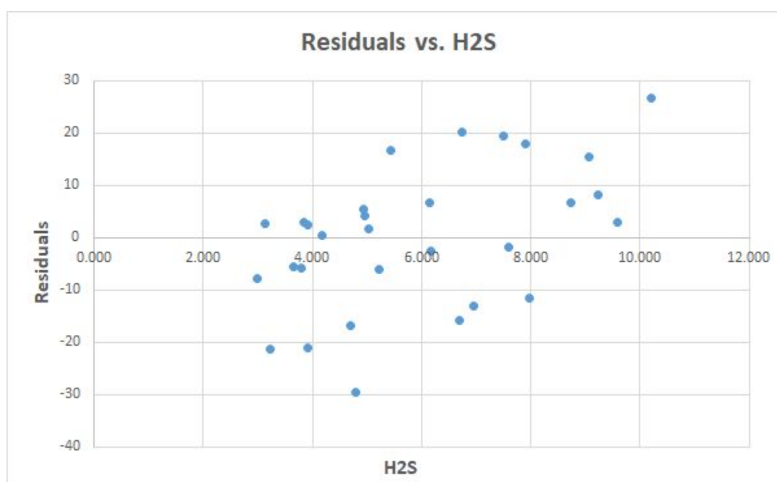
t-Stat	P-Value
3.4806	0.0017

Confidence Interval

	2.5 %	97.5 %
(Intercept)	-112.39	-10.6
Acetic	6.44	24.9

Exhibits 14-15: Graph of Regression Residuals vs. Remaining Explanatory Variables



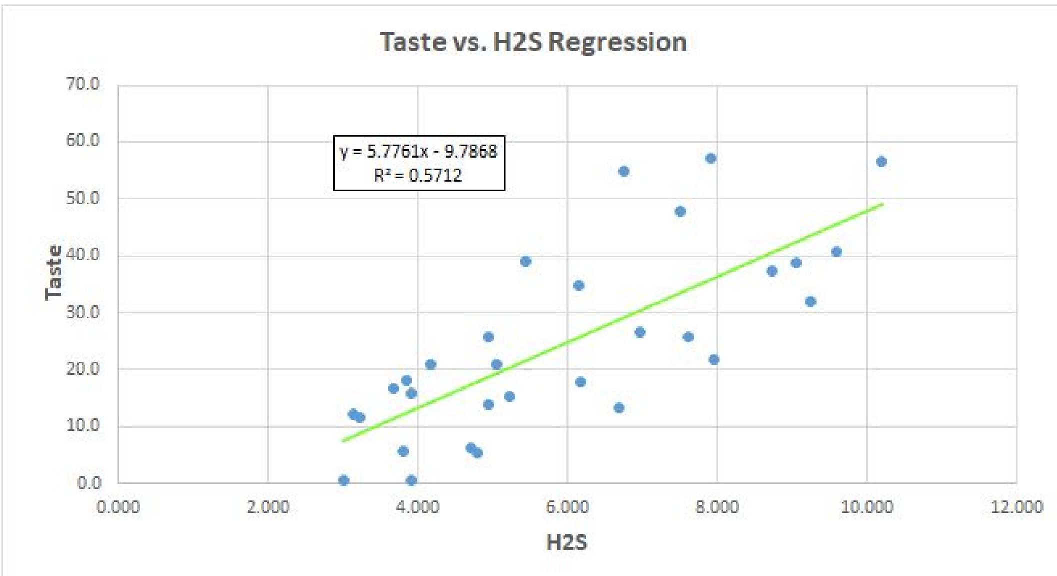


Taste vs. Acetic Regression Analysis

- The residuals appear to have a Normal Distribution but are positively associated with both H2S and Lactic.
- The P-value of the test statistic is low at 0.002, thus there appears to be an association between Taste and Acetic.

Problem 11.56

Exhibit 16: Linear Regression Model of Dependence of Taste on Acetic



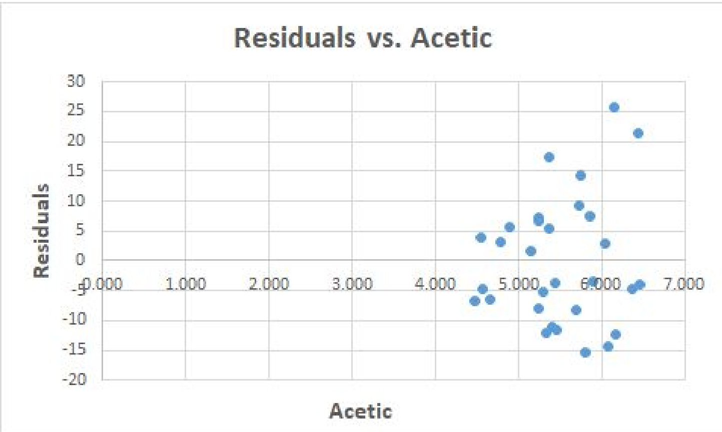
Linear Regression Model: Taste = -9.7868 + 5.7761 (H2S)

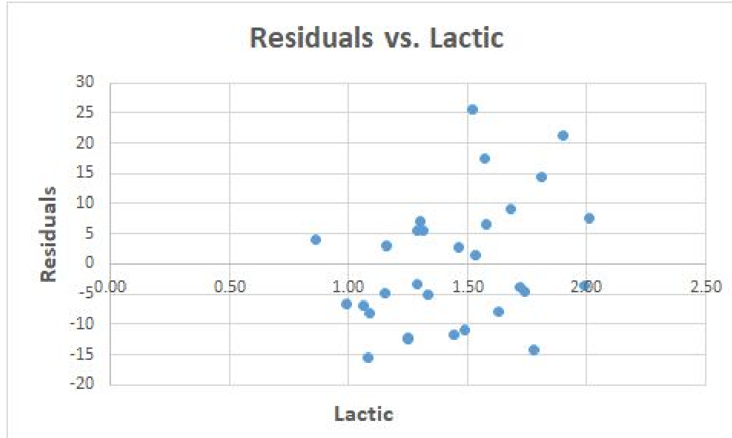
t-Stat	P-Value
6.107	1.37*10^-6

Confidence Interval

2.5 % 97.5 %
(Intercept) -21.99 2.42
H2S 3.84 7.71

Exhibits 17-18: Graph of Regression Residuals vs. Remaining Explanatory Variables



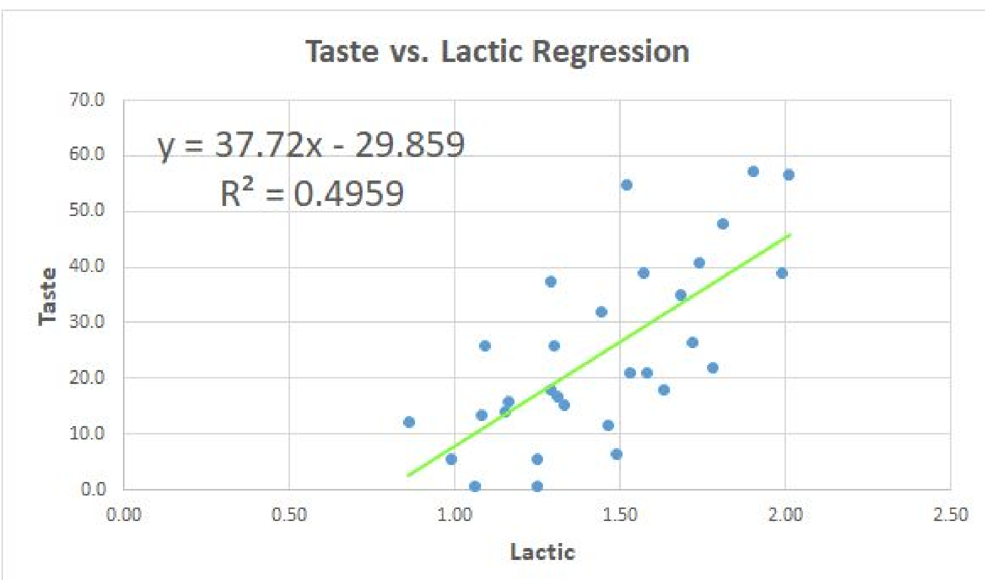


Taste vs. H2S Regression Analysis

- The residuals appear to have a Normal distribution and there are no clear patterns evident between the residuals and the other variables of Acetic and Lactic
- The P-value of the test statistic is low at 1.37*10^-6, thus there appears to be an association between Taste and H2S.

Problem 11.57

Exhibit 19: Linear Regression Model of Dependence of Taste on Lactic



Linear Regression Model: Taste = -29.859 + 37.72 (Lactic)

t-Stat	P-Value
5.249	1.41*10^-5

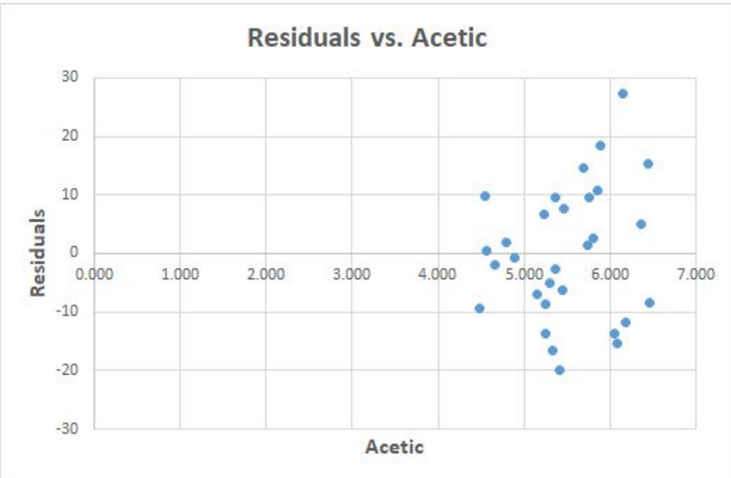
Confidence Interval

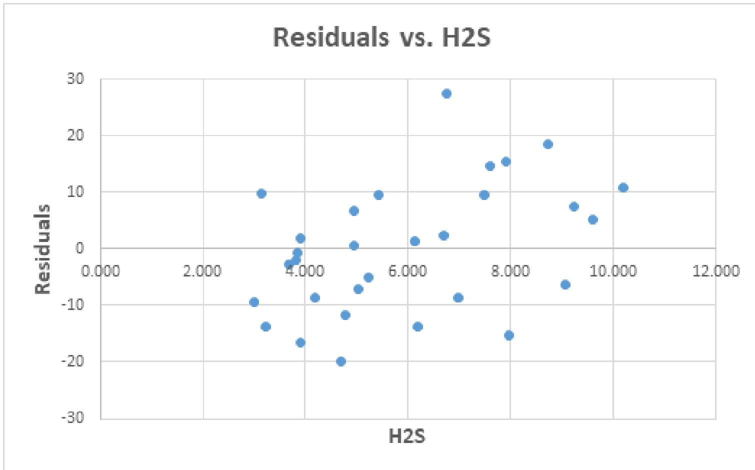
2.5 % 97.5 %

(Intercept) -51.5 -8.18

Lactic 23.0 52.44

Exhibits 20-21: Graphs of Regression Residuals vs. Remaining Explanatory Variables





Taste vs. Lactic Regression Analysis

- The residuals appear to have a Normal distribution and there are no clear patterns evident between the residuals and the other variables of Acetic and H2S.
- The P-value of the test statistic is low at 1.41*10^-5, thus there appears to be an association between Taste and Lactic.

Problem 11.58

Exhibit 22: Table of Key Metrics from Regressions of Taste vs. Explanatory Variables

Statistic	Taste vs Acetic	Taste vs H2S	Taste vs Lactic
F	12.1	37.3	27.5
P-value	0.0017	1.4e-06	1.4e-05

R^2	0.302	0.571	0.496
s	10.32	10.83	11.75
Equation	Taste=Acetic*15.6-61.5	Taste=H2S*5.776-9.787	Taste=Lactic*37.72-29.86

From simple linear regression, the most significant variable was H2S, and the least significant was Acetic. The intercept values represent the minimum taste score given to each cheese if there each chemical was absent. Each simple linear model assumes that each chemical is the only one that affects taste, however this is not realistically the case, and there could be many other factors not examined in this study that could contribute to the intercept. Given that the experimental study took log transformed concentrations of acetic acid and hydrogen sulfide, the intercepts may also be difficult to compare.

Problem 11.59

Using Acetic and H2S as explanatory variables for Taste yields the following multiple regression model with the equation Taste = 3.80(Acetic) + 5.15(H2S) - 26.94.

Residuals:

```
  Min     1Q Median     3Q    Max
-16.11  -6.89  -1.67   6.59  23.71
```

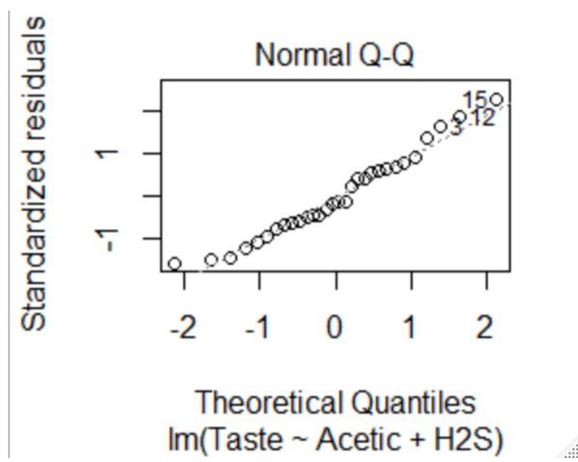
Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -26.94      21.19   -1.27  0.21454
Acetic          3.80       4.51    0.84  0.40625
H2S             5.15       1.21    4.26  0.00022
```

Residual standard error: 10.9 on 27 degrees of freedom

Multiple R-squared: 0.582,   Adjusted R-squared: 0.551

F-statistic: 18.8 on 2 and 27 DF, p-value: 7.65e-06



Confidence Interval:

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

Analysis of Variance Table

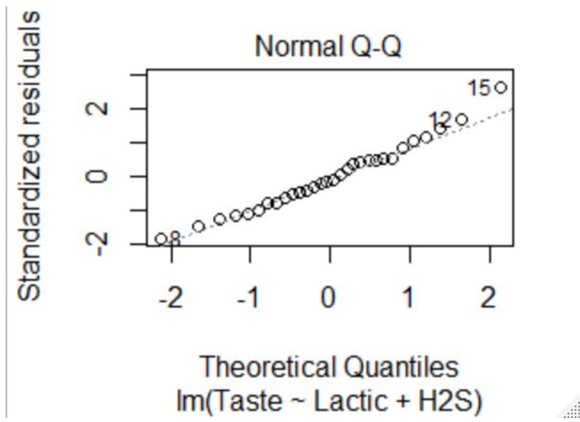
Response: Taste

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

The p-value for Acetic is 0.406, and the p-value for H2S is 0.00022, which means that Acetic is not significant in this model. From the correlation matrix discussed previously, Acetic and H2S have are correlated (0.618), which explains why Acetic is not significant in this model. In the simple linear regression model of Taste using Acetic, the p-value for the slope is 0.0017, so Acetic is a better predictor variable when it is not combined with H2S. The residual plot is approximately normal with slight left skew in particular for higher quantile values.

Problem 11.60

The following multiple regression model uses Lactic and H2S to predict Taste. The equation of the regression line is Taste=19.89(Lactic)+3.95(H2S)-27.59.



Residuals:

Min	1Q	Median	3Q	Max
-17.34	-6.53	-1.16	4.84	25.62

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.59	8.98	-3.07	0.0048
Lactic	19.89	7.96	2.50	**0.0188**
H2S	3.95	1.14	3.47	**0.0017**

Residual standard error: 9.94 on 27 degrees of freedom
Multiple R-squared: 0.652, Adjusted R-squared: 0.626
F-statistic: 25.3 on 2 and 27 DF, p-value: 6.55e-07

Confidence Interval:

	2.5 %	97.5 %
(Intercept)	-46.02	-9.16
Lactic	3.56	36.22
H2S	1.62	6.28

Analysis of Variance Table

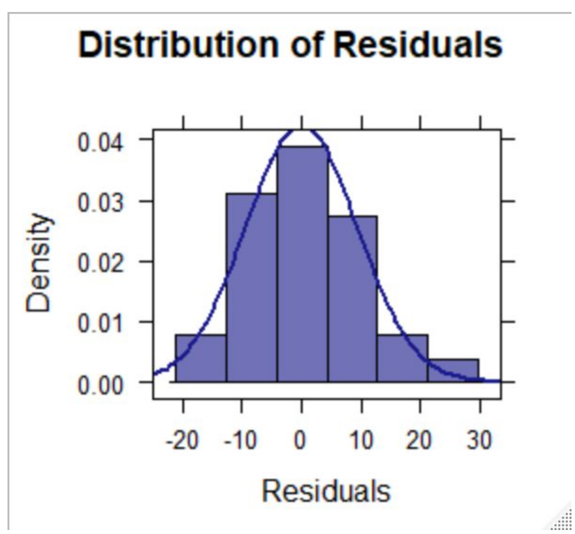
Response: Taste

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lactic	1	3800	3800	38.5	1.2e-06
H2S	1	1194	1194	12.1	0.0017
Residuals	27	2669	99		

Based on the p-values, both Lactic and H2S are significant variables in this model. These variables have a correlation of 0.645, which is strong, however this does not render either of the variables insignificant like in the previous problem. The residuals are approximately normal with a slight left skew.

Problem 11.61

The final multiple regression model uses Acetic, H2S, and Lactic to explain Taste. The equation is Taste=0.328(Acetic)+3.912(H2S)+19.671(Lactic)-28.877.



Residuals:

Min	1Q	Median	3Q	Max
-17.39	-6.61	-1.01	4.91	25.45

Coefficients:

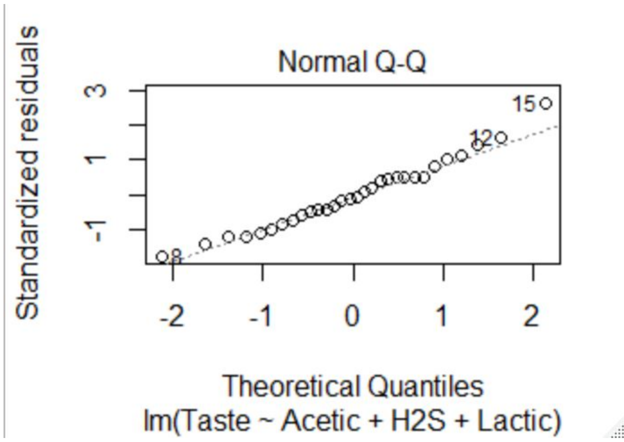
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28.877	19.735	-1.46	0.1554
Acetic	0.328	4.460	0.07	**0.9420**

H2S	3.912	1.248	3.13	**0.0042**
Lactic	19.671	8.629	2.28	**0.0311**

Residual standard error: 10.1 on 26 degrees of freedom

Multiple R-squared: 0.652, Adjusted R-squared: 0.612

F-statistic: 16.2 on 3 and 26 DF, p-value: 3.81e-06



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 Table

Response: Taste

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Acetic	1	2314	2314	22.6	6.5e-05
H2S	1	2147	2147	20.9	0.0001
Lactic	1	533	533	5.2	0.0311
Residuals	26	2668	103		

In this model, the p-value for Acetic is 0.942, H2S is 0.0042 and Lactic is 0.0311. The residuals are approximately normally distributed with a slight right sket. The p-value for acidic is high and means that this variable is not significant in the model. The next step would be to eliminate the Acetic variable and do multiple regression with H2S and Lactic, which we did in the previous question. The model Taste=19.89(Lactic)+3.95(H2S)-27.59 gives the best explanation of Taste. The 2 variable model has lower residual standard error (9.94) than the 3 variable model (10.10) while having the same multiple R-squared value (0.652). The normal nature of the residual plots for all of these models rules out the need for any form of polynomial regression, meaning linear regression is sufficient.

I pledge my honor that I have abided by the Stevens Honor system

Erika McCarthy, April Ullrich