Final Project

Simon Gao

5/14/2020

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

Section 1: A Statistical Report

Executive Summary

This report summarizes my findings on the potential differences in using different explanatory variables for simple linear regression models and multiple regression models. Specifically, I will discuss my findings in studying how 3 variables—Acetic, H2S, and Lactic—affect the Taste of cheese. I chose this topic because I was assigned it as a final project for my Intermdiate Statistics course. This report will contain the following: an explanation of the overall study, an explanation of the statistical methods used, a preliminary analysis of the data set, an anaylsis on the regression models, and a summary of the results.

The Study

As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The variable "Case" is used to number the observations from 1 to 30. "Taste" is the response variable of interest. The taste scores were obtained by combining the scores from several tasters. Three of the chemicals whose concentrations were measured were acetic acid, hydrogen sulfide, and lactic acid. For acetic acid and hydrogen sulfide (natural) log transformations were taken. Thus, the explanatory variables are the transformed concentrations of acetic acid ("Acetic") and hydrogen sulfide ("H2S") and the untransformed concentration of lactic acid ("Lactic"). We will analysis a variety of linear regression models to determine which is the best predictory for Taste.

Statistical Methodology

My goal is to compare simple and multiple linear regression models using 3 explanatory variables—Acetic, H2S, and Lactic—and Taste as a reponse variable. In this report, I will create simple linear regression models (total of 3 models) for each explanatory variable and then pair them for multiple regression models (total of 3 models). There are a lot of statistical calculations for regression analysis but we will utilize the software "R" to perform these calculations for us.

Preliminary Analysis of the Data

Before we construct linear models or perform regression analysis, it is essential for us to graphically check the relationships between variables and determine any deviations from normality for the variables. Preliminary analysis, using QQ plots, did not show any significant deviations from normality. However, the scatterplots of each pair of variables, 6 in total, all displayed a positive association betwn the pairs.

Regression Analysis

First, we created linear models using the 3 explanatory variables separately as predictors for taste. The table below summarizes the resulting linear models and their respective equations.

Linear Model	F	$P - value$	R^2	s
Taste-Acetic	12.11	0.002	30.2%	13.82
Taste-H2S	37.29	$1.37 * 10^-6$	57.1%	10.83
Taste-Lactic	27.55	$1.41 * 10^-5$	49.6%	11.75

$\hat{Taste}$ = -61.5 + 15.65 * *Acetic*
$\hat{Taste}$ = -9.787 + 5.776 * *H2S*
$\hat{Taste}$ = -29.86 + 37.72 * *Lactic*

Out of all the simple linear regression models, the one with H2S seems to be the best one as its adjusted R^2 value is the highest at 57.1%. This implies that the linear model accounted for 57.1% of variation in Taste. Moreover, the model with Acetic as a predictor seemed to be the worst model out of the 3 simple linear regression models as it had the lowest adjusted R^2 value at 30.2%. In all 3 models, their residual plots did not have any striking deviations from a normal distribution. Additionally, the p-values for coefficients H2S and Lactic are significantly less than that of coefficient Acetic. Although the Acetic model shows a statistically significant relationship to Taste, it is not as statistically significant as the relationship of H2S and Lactic to Taste.

As for the remaining 3 regression models, we created multiple regression models (see table below).

Linear Model	F	$P - value$	R^2	s
Taste-Acetic-H2S	18.81	$7.65 * 10^-6$	58.2%	10.89
Taste-H2S-Lactic	25.26	$6.55 * 10^-7$	65.2%	9.94
Taste-Acetic-H2S-Lactic	16.22	$3.81 * 10^-6$	65.2%$	10.13

Out of the three multiple regression models, the H2S/Lactic model performed the best with an adjusted R^2 value of 62.6%. The second best model was the model with all 3 explanatory variables which cover an adjusted 61.2% of total variation in Taste. In the two models which included Acetic as a predictor, the p-value for coefficient Acetic was not satistically significant. This result may suggest that the variable Acetic does not add any significant information to the model when paired with H2S or Lactic.

Conclusion

The regression analysis showed that the H2S/Lactic model was the best in predicting Taste. This is because the model has the highest adjusted R^2 value of 0.6259 which indicates that the model covers 62.59% of variation in Taste. The resulting regression equation from this model is: $\hat{Taste}$ = -27.6 + 3.95 * *H2S* + 19.89 * *Lactic*. In addition, the this study suggests that the variable Acetic may not bring as much information to predicting Taste as H2S and Lactic would as exemplified by the last regression model which included all 3 explanatory variables.

Section 2: The Details of the Cheese Study

Problem 11.53

```
## [1] "Taste"
```

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

```
## [1] "Acetic"
```

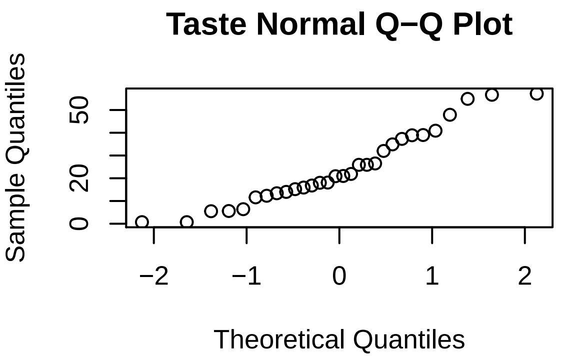
```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   44 | 846
##   46 | 69
##   48 | 0
##   50 | 6
##   52 | 4450377
##   54 | 146
##   56 | 046
##   58 | 069
##   60 | 4858
##   62 | 7
##   64 | 56
```

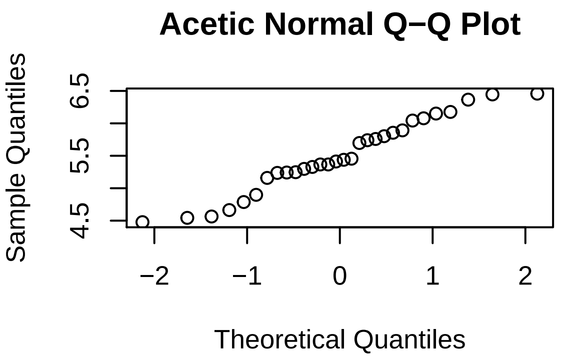
```
## [1] "H2S"
```

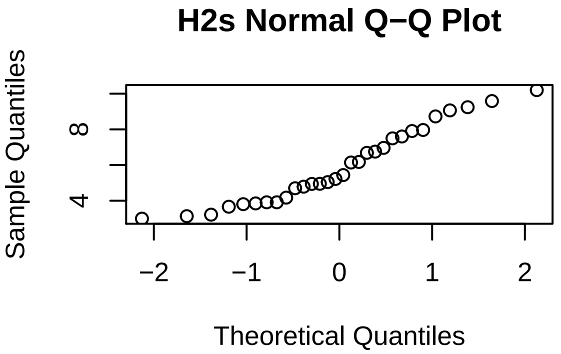
```
##
##   The decimal point is at the |
##
##    2 |
##    3 | 01278999
##    4 | 27899
##    5 | 024
##    6 | 1278
##    7 | 0569
##    8 | 07
##    9 | 126
##   10 | 2
```

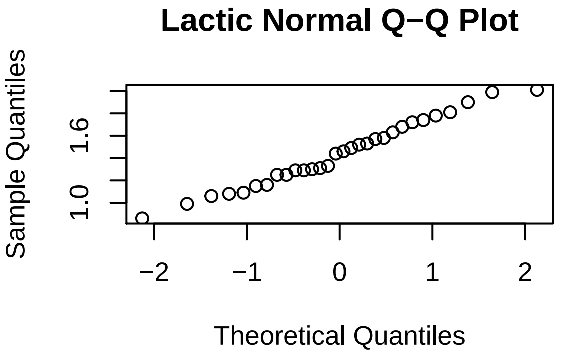
```
## [1] "Lactic"
```

```
##
##   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
```





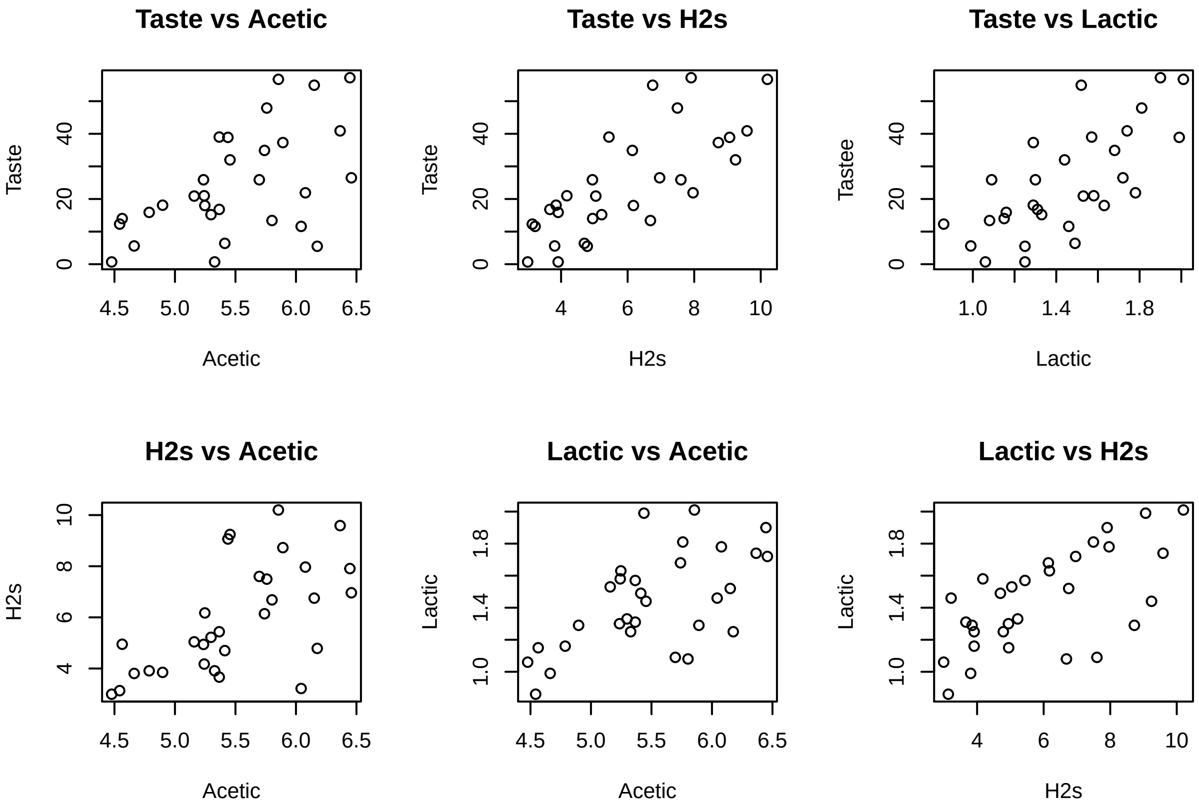




Variable	Taste	Acetic	H2s	Lactic
Mean	24.5333333	5.4980333	5.9417667	1.442
Median	20.95	5.425	5.329	1.45
Standard Deviation	16.2553828	0.5708784	2.1268792	0.30349
IQR	23.15	0.64525	3.59725	0.4175

The variables do not have any alarming deviations from normality. Taste and H2s are slightly right-skewed while Acetic seems to have 2 peaks.

Problem 11.54



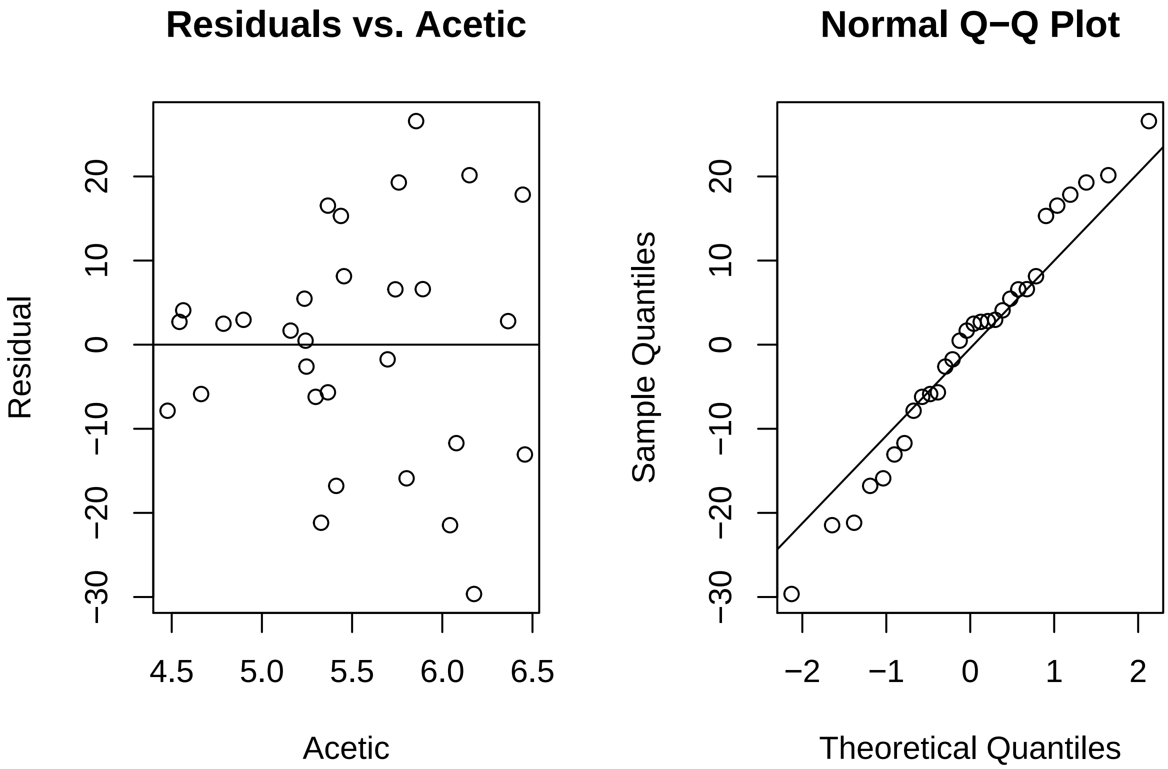
Pair	Correlation Coefficient (r)	P-value
Taste-Acetic	0.5495393	0.0016582
Taste-H2s	0.7557523	1.3737834×10^{-6}
Taste-Lactic	0.7042362	1.4051168×10^{-5}
Acetic-H2s	0.6179559	2.7391729×10^{-4}
Acetic-Lactic	0.6037826	4.1136568×10^{-4}
H2s-Lactic	0.6448123	1.1984014×10^{-4}

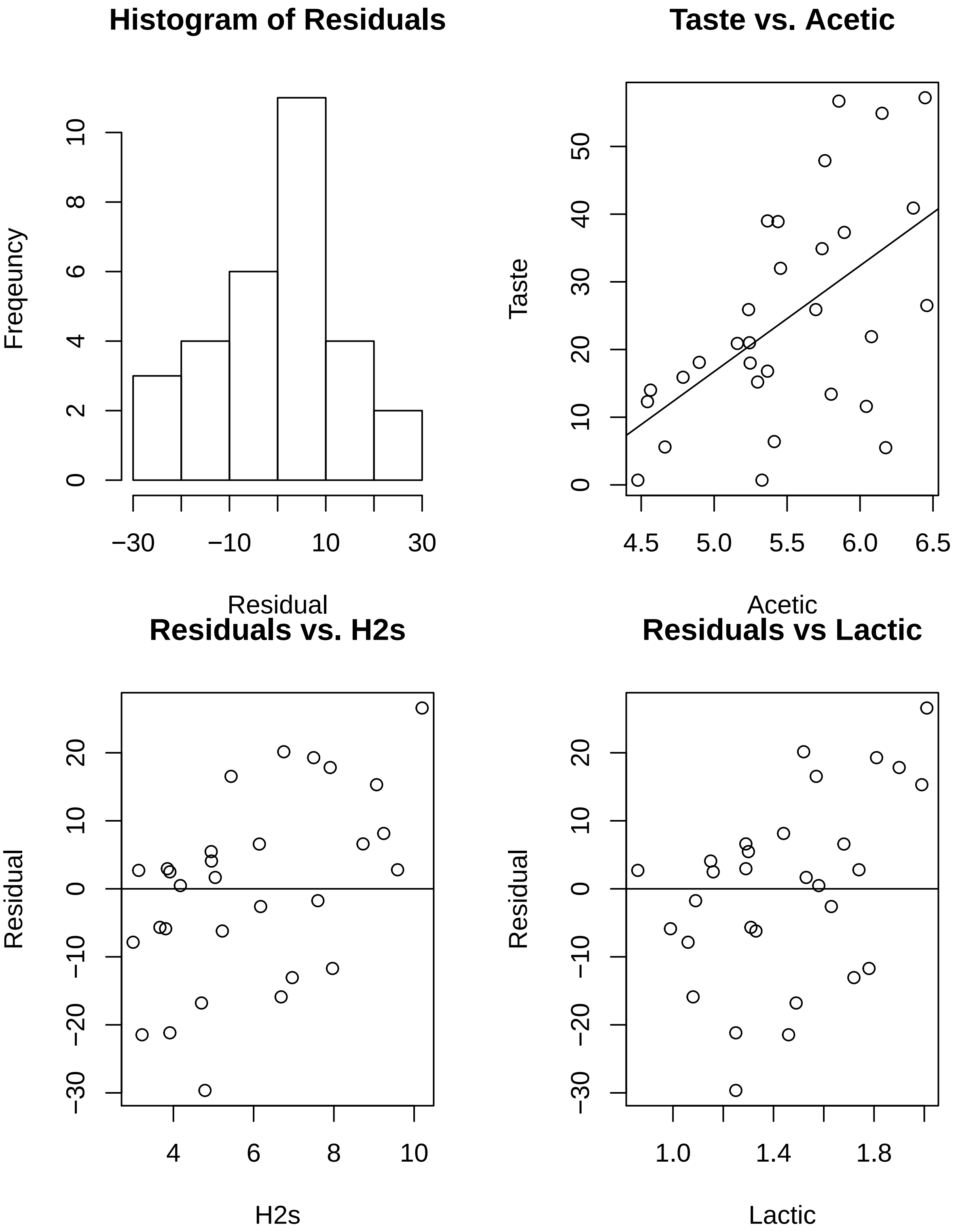
All 6 plots show a positive association between each pair of variables.

Problem 11.55

```
##
## Call:
## lm(formula = dt$taste ~ dt$acetic)
##
## Coefficients:
## (Intercept)    dt$acetic
##      -61.50        15.65
```

```
##
## Call:
## lm(formula = dt$taste ~ dt$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 *
## dt$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
```





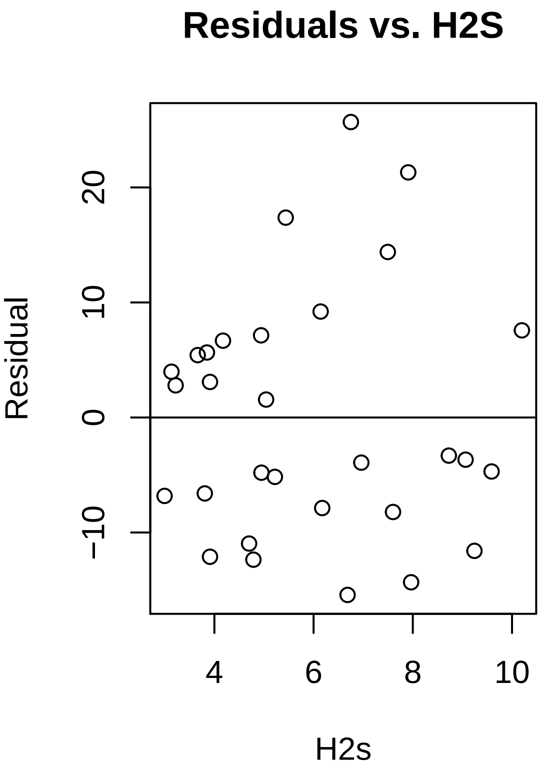
$\hat{Taste}$ = -61.5 + 15.65 * $Acetic$

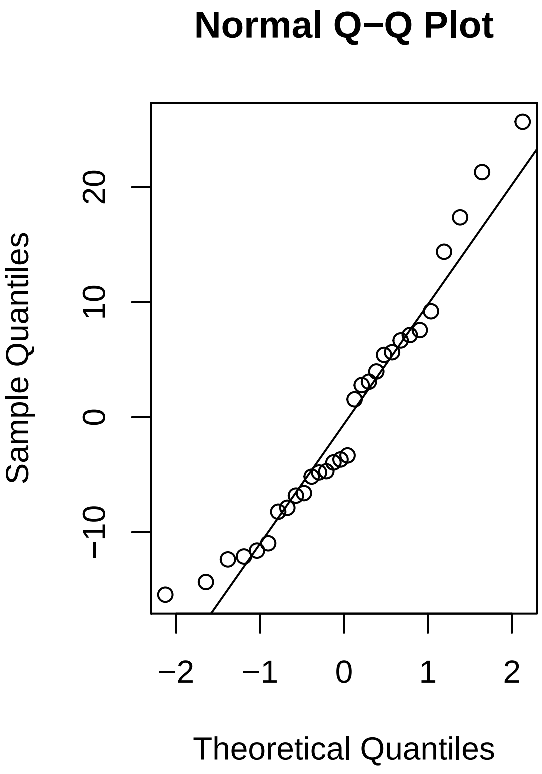
The linear model has a p-value of 0.002 which signifies that there is a statistically significant relationship between Taste and Acetic. In addition, the residuals of the taste-acetic linear model seem to have a normal distribution. When the residuals are plotted against H2S and Lactic, they seem to have a positive association.

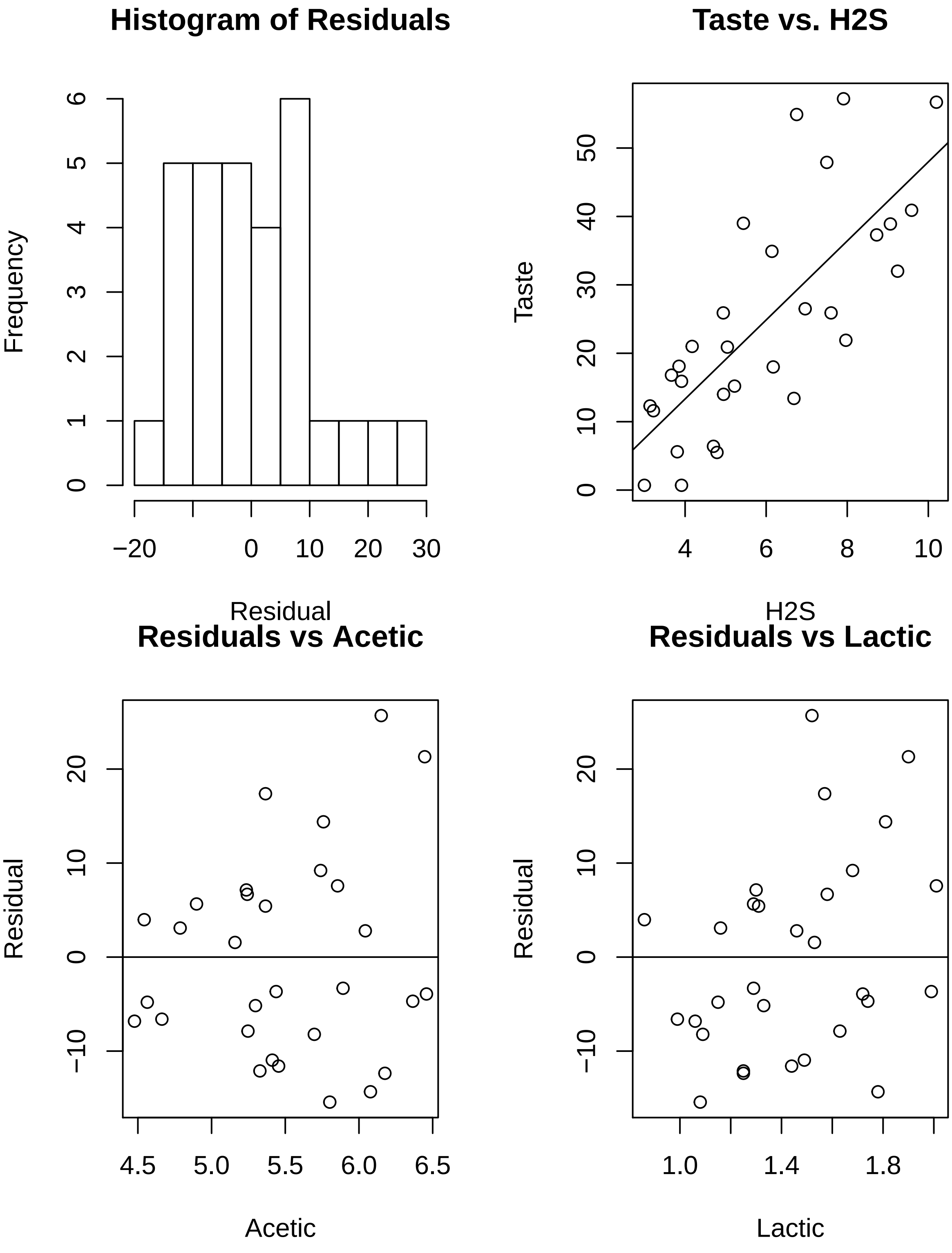
Problem 11.56

```
##
## Call:
## lm(formula = dt$taste ~ dt$h2s)
##
## Coefficients:
## (Intercept)       dt$h2s
##      -9.787        5.776
```

```
##
## Call:
## lm(formula = dt$taste ~ dt$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
## dt$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
```





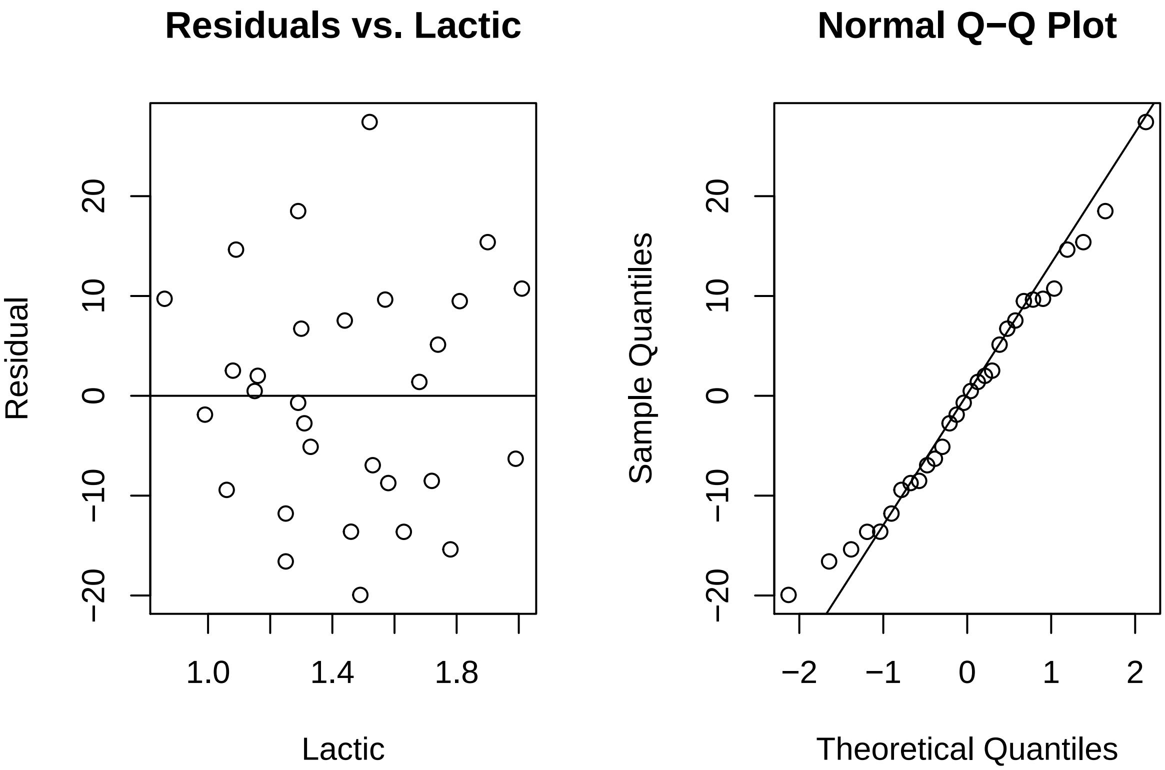


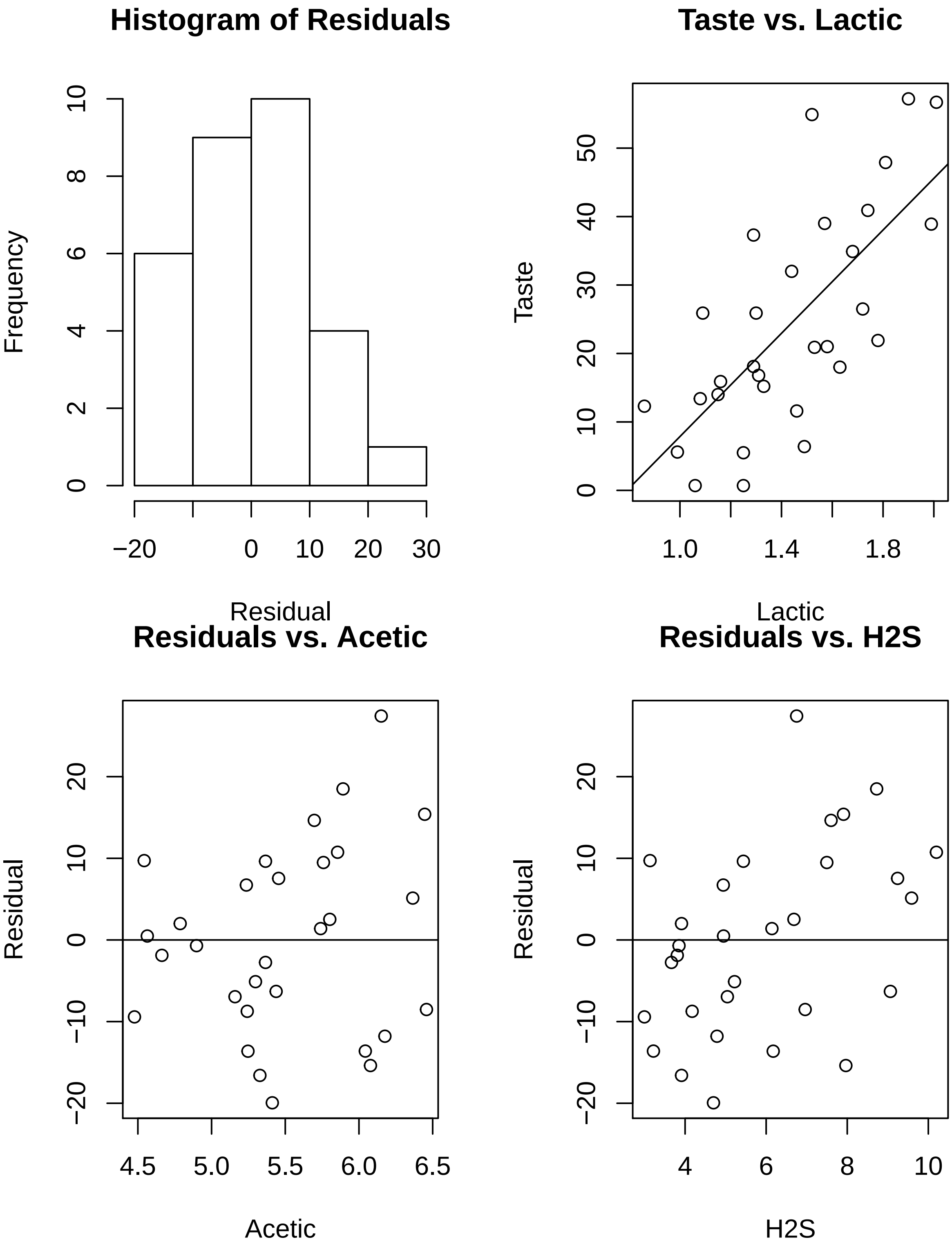
$\hat{Taste}$ = -9.787 + 5.776 * $H2S$
The linear model has a p-value of $1.37 * 10^-6$ which signifies that there is a statistically significant relationship between Taste and H2S. Based on the residual normal Q-Q plot and histogram, the residuals seem to have a normal distribution. When residuals are plotted against Acetic and Lactic, there seems to be larger scatter in the Lactic plot for higher values.

Problem 11.57

```
##
## Call:
## lm(formula = dt$taste ~ dt$lactic)
##
## Coefficients:
## (Intercept)    dt$lactic
##      -29.86        37.72
```

```
##
## Call:
## lm(formula = dt$taste ~ dt$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 **
## dt$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
```





$\hat{Taste}$ = -29.86 + 37.72 * $Lactic$

The p-value of the coefficient is significantly close to 0 which implies that there is a statistically significant relationship between Taste and Lactic. The residuals seem to be normally distributed based on its histogram and normal Q-Q plot. When plotted against the other two chemical variables, the plot seems to be randomly scattered with no real distinguishable pattern.

Problem 11.58

Linear Model	F	$P - value$	R^2	s
Taste-Acetic	12.11	0.002	30.2%	13.82
Taste-H2S	37.29	$1.37 * 10^-6$	57.1%	10.83
Taste-Lactic	27.55	$1.41 * 10^-5$	49.6%	11.75

$\hat{Taste}$ = -61.5 + 15.65 * *Acetic*
$\hat{Taste}$ = -9.787 + 5.776 * *H2S*
$\hat{Taste}$ = -29.86 + 37.72 * *Lactic*
The intercepts are different because the linear regression models use different explanatory variables to predict Taste.

Problem 11.59

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

For the Acetic coefficient, this linear model is not better than the model with only Acetic as an explanatory variable. The p-value in this model for the Acetic coefficient is 0.406 which does not signify a statistically significant relationship between the Acetic variable and Taste in this model. Acetic and H2S have a correlation coefficient of 0.618 and H2S in this model has a statistically significant relationship to Taste. Therefore, the Acetic variable does not add significant information to help us predict Taste if H2S is included due to the two explanatory variables' correlation.

Problem 11.60

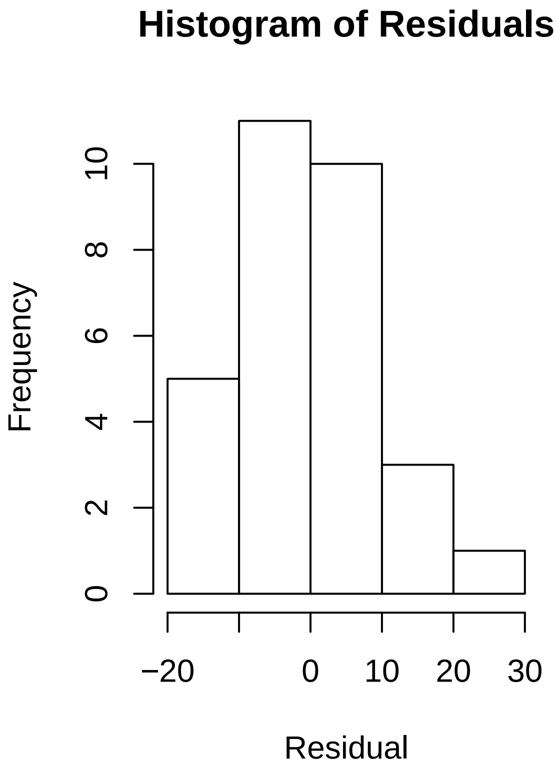
```
##
## Call:
## lm(formula = dt$taste ~ dt$h2s + dt$lactic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.343  -6.530  -1.164   4.844  25.618
##
## Coefficients:
```

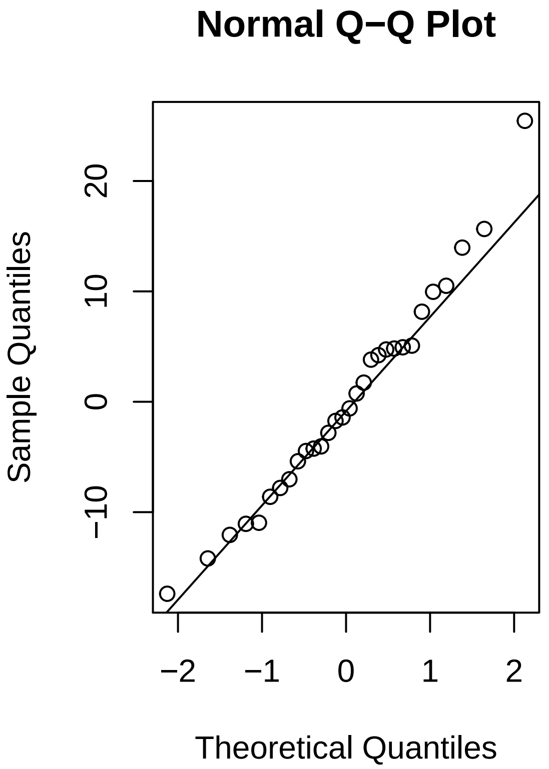
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -27.592      8.982  -3.072  0.00481 **
## dt$h2s         3.946      1.136   3.475  0.00174 **
## dt$lactic     19.887      7.959   2.499  0.01885 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual 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
```

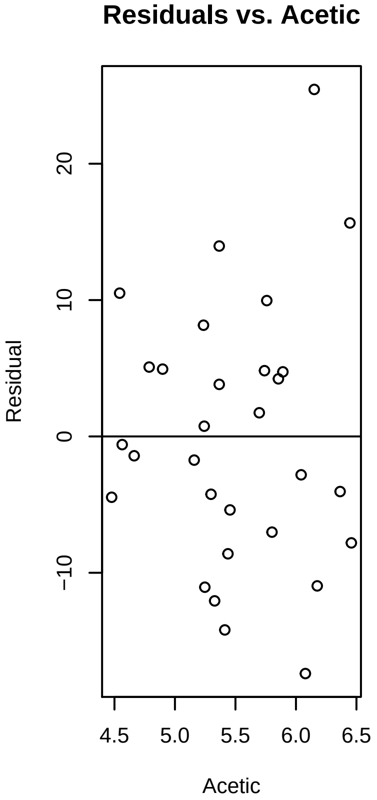
Our multiple regression model using H2S and Lactic to predict Taste is a better predictor for Taste and the model covers 65.2% of the variation in Taste, which is higher than the 57.1% and 49.6% coverage of variation for H2S and Lactic respectively. Additionally, both coefficients in this model has statistically significant p-values (0.002 for H2s and .02 for Lactic). These significant p-values suggest that, in this model, H2S and Lactic have a statistically significant relationship to Taste.

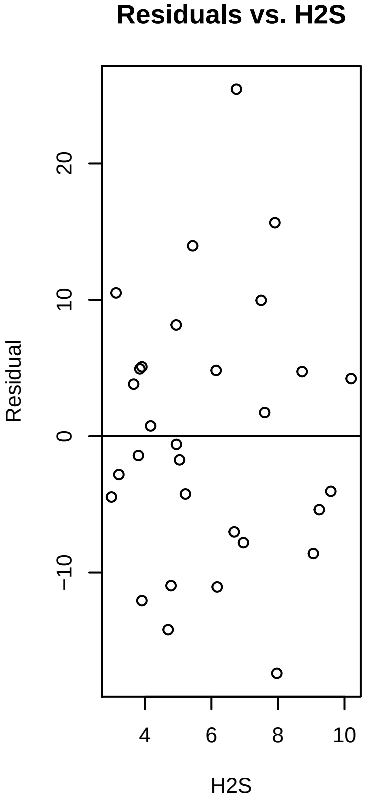
Problem 11.61

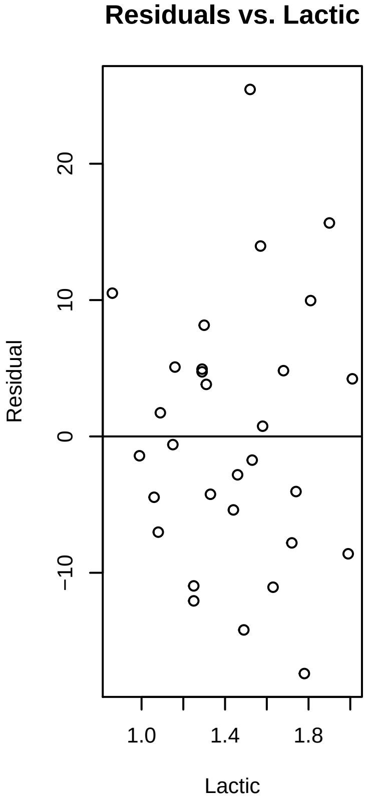
```
##
## Call:
## lm(formula = dt$taste ~ dt$acetic + dt$h2s + dt$lactic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## dt$acetic     0.3277     4.4598   0.073  0.94198
## dt$h2s        3.9118     1.2484   3.133  0.00425 **
## dt$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
```


Histogram of Residuals


Normal Q−Q Plot


Residuals vs. Acetic


Residuals vs. H2S


Residuals vs. Lactic

The coefficient of Acetic is not statistically different from 0 as its p-value is 0.94198 (significantly high). However, from the multiple regression model, H2S and Lactic have p-values of 0.004 and 0.031 respectively which signify a statistically signifcant relationship to Taste. Based on its histogram and QQ plot, the residuals seem to be normally distributed and show no significant patterns when plotted with the explanatory variables.

The model using H2S and Lactic as predictors for Taste is the best as it covers the most variation in Taste with its adjusteed R^2 value of 0.6259. In addition, all of its coefficients are statistically significant within the model.