Final Project

Ian Baker, Loughlin Claus, Zack Schieberl

12/7/2019

Pledge

I pledge my honor that I have abided by the Stevens Honor System - Ian Baker, Loughlin Claus, Zack Schieberl

11.53

```
cheese <- as.matrix(read.csv2("cheese.csv", header = TRUE, sep = ","))

printRegEq <- function(funcSum, name) {
  cat("Taste vs", name, funcSum$coefficients[, 3][1], "+", name, "*", funcSum$coefficients[, 3][2], "\n")
}

cheeseCols <- colnames(cheese)
for (col in cheeseCols) {
  cur <- as.numeric(cheese[, col])
  # mean, median, sd, iqr
  out <- c(paste("Type:", col), paste("Mean:", round(mean(cur), 2)),
           paste("Median:", round(median(cur), 2)), paste("SD:", round(sd(cur), 2)),
           paste("IQR:", round(IQR(cur), 2)))
  print(format(out, justify = "left", trim = TRUE))
  # stemplot
  stem(cur)
  # normal quantile plot
  qqnorm(cur, main = col)
  qqline(cur)
}
```

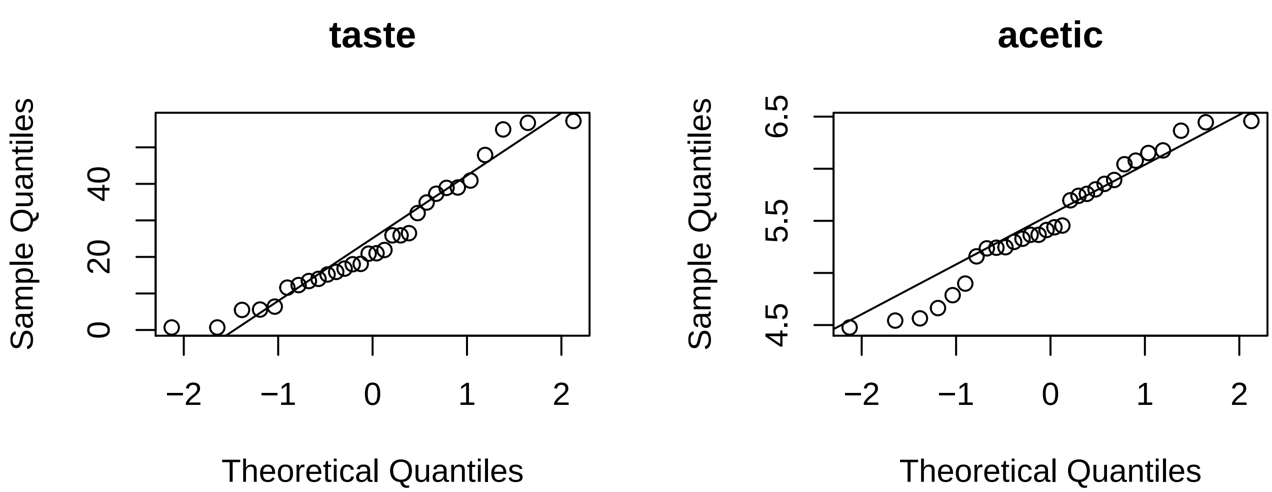
```
## [1] "Type: taste  " "Mean: 24.53  " "Median: 20.95" "SD: 16.26    "
## [5] "IQR: 23.15   "
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 11666
##   1 | 223456788
##   2 | 112667
##   3 | 25799
##   4 | 18
##   5 | 577
```

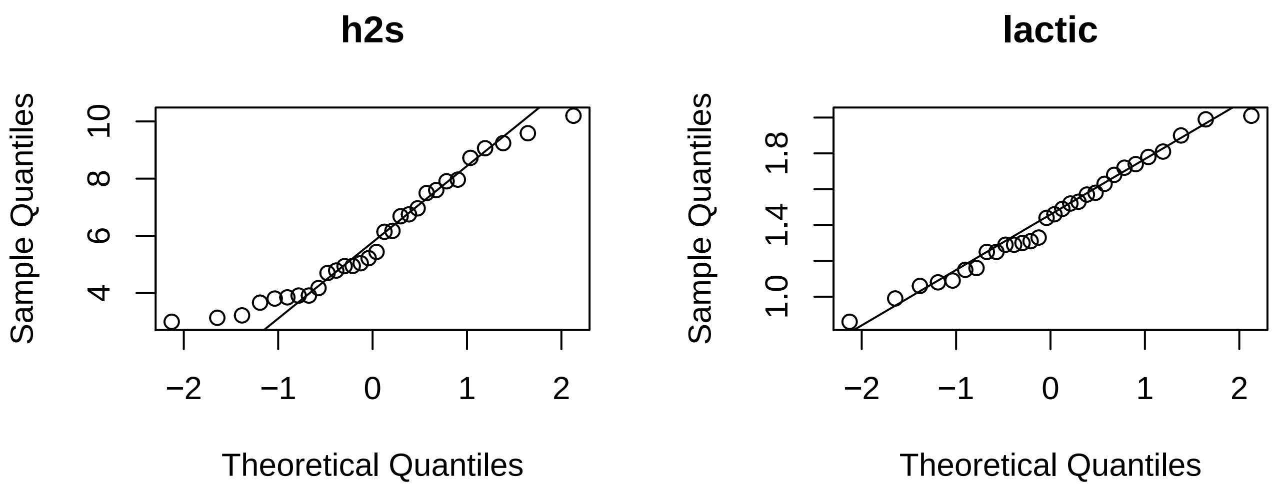
```
## [1] "Type: acetic" "Mean: 5.5   " "Median: 5.42" "SD: 0.57    " "IQR: 0.65   "
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   44 | 846
##   46 | 69
##   48 | 0
##   50 | 6
##   52 | 4450377
##   54 | 146
##   56 | 046
```

```
##   58 | 069
##   60 | 4858
##   62 | 7
##   64 | 56
```

```
## [1] "Type: h2s   " "Mean: 5.94  " "Median: 5.33" "SD: 2.13    " "IQR: 3.6    "
##
##   The decimal point is at the |
##
##    2 |
##    3 | 01278999
##    4 | 27899
##    5 | 024
##    6 | 1278
##    7 | 0569
##    8 | 07
##    9 | 126
##   10 | 2
```

```
## [1] "Type: lactic" "Mean: 1.44  " "Median: 1.45" "SD: 0.3     " "IQR: 0.42   "
##
##   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
```



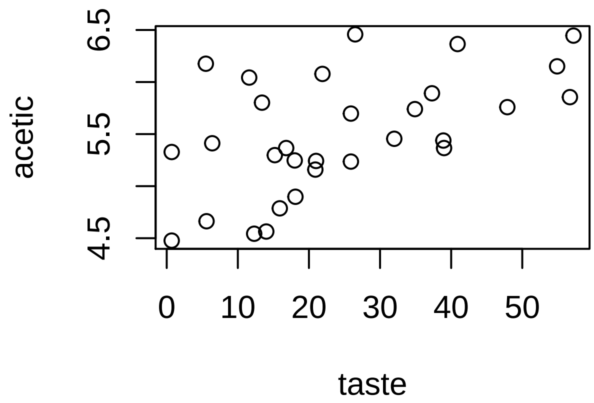


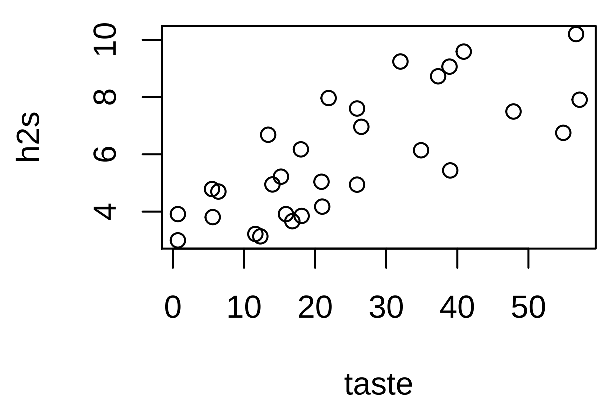
While H2S and Taste have some right skew, and Acetic has two peaks, the data all appears to be relatively normal. There are no outliers in the data.

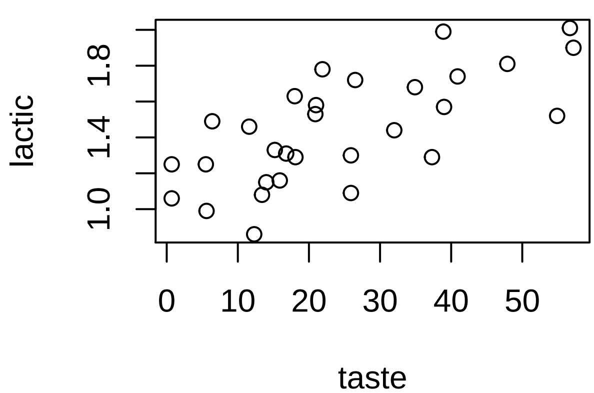
11.54

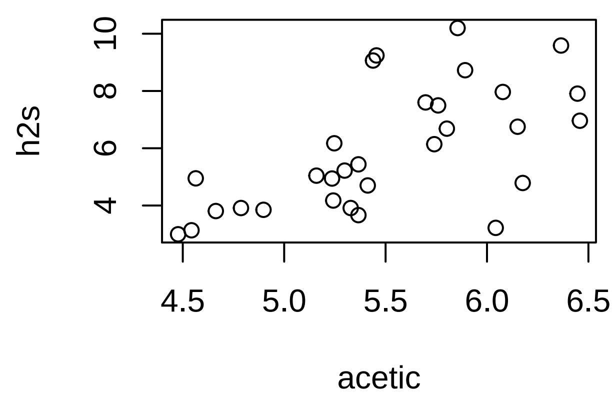
```
for (col in cheeseCols) {
  colIdx <- grep(col, cheeseCols)
  col1Data <- as.numeric(cheese[, col])
  for (col2 in cheeseCols) {
    if (colIdx < grep(col2, cheeseCols)) {
      col2Data <- as.numeric(cheese[, col2])
      plot(col1Data, col2Data, xlab = col, ylab = col2)
      correl <- cor.test(col1Data, col2Data)
      cat("Correlation between", col, "and", col2, "is:", correl$estimate,
          "with a p-value of", correl$p.value, "\n")
    }
  }
}
```

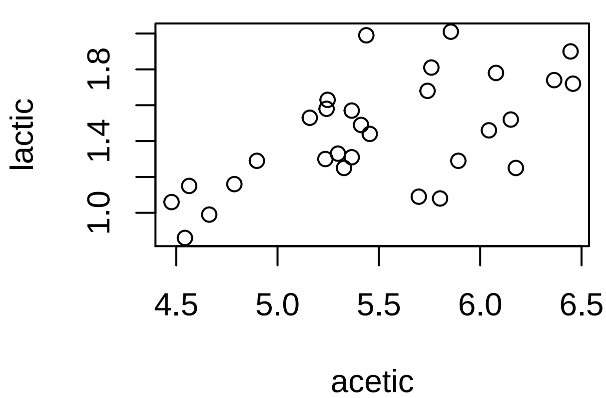
```
## Correlation between taste and acetic is: 0.5495393 with a p-value of 0.001658192

## Correlation between taste and h2s is: 0.7557523 with a p-value of 1.373783e-06

## Correlation between taste and lactic is: 0.7042362 with a p-value of 1.405117e-05

## Correlation between acetic and h2s is: 0.6179559 with a p-value of 0.0002739173

## Correlation between acetic and lactic is: 0.6037826 with a p-value of 0.0004113657

## Correlation between h2s and lactic is: 0.6448123 with a p-value of 0.0001198401
```

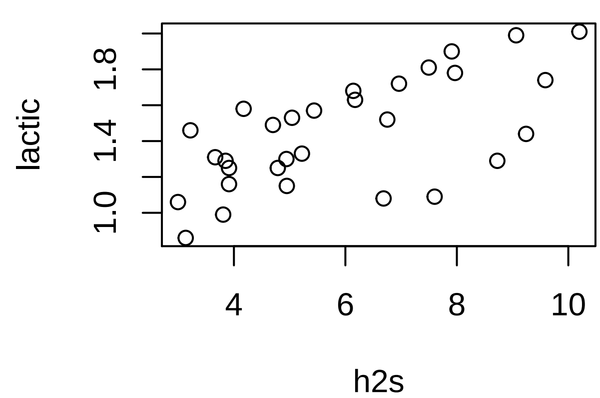

acetic
4.5
5.5
6.5
0
10
20
30
40
50
taste


h2s
4
6
8
10
0
10
20
30
40
50
taste


lactic
1.0
1.4
1.8
0
10
20
30
40
50
taste

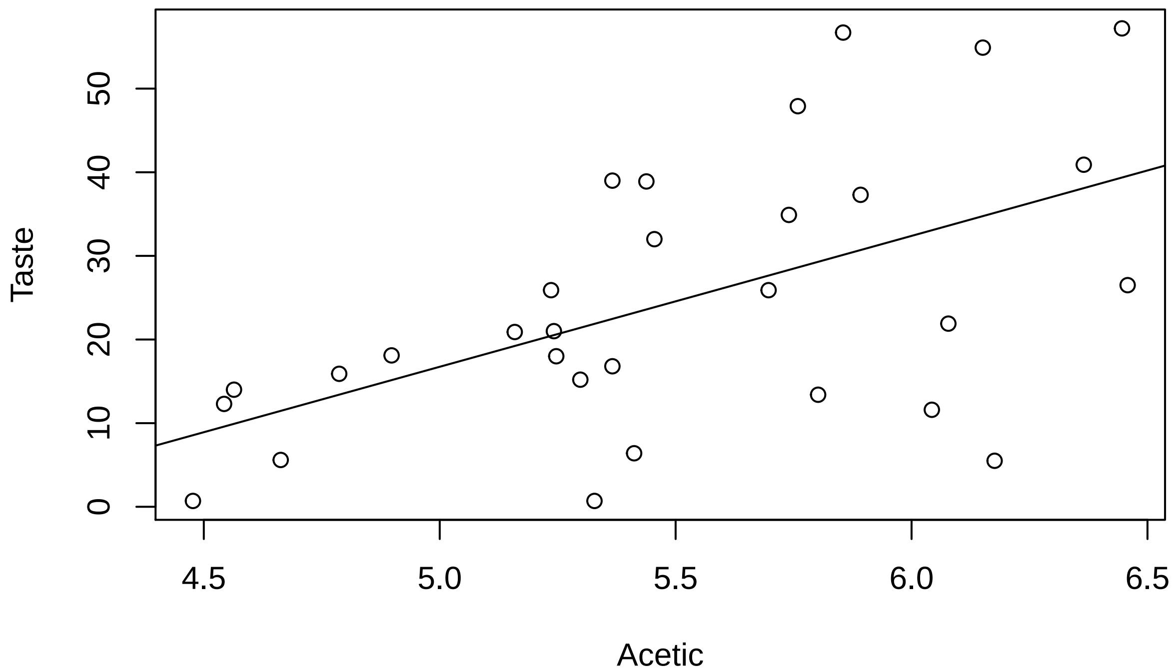

h2s
4
6
8
10
4.5
5.0
5.5
6.0
6.5
acetic


lactic
1.0
1.4
1.8
4.5
5.0
5.5
6.0
6.5
acetic


lactic
1.0
1.4
1.8
4
6
8
10
h2s

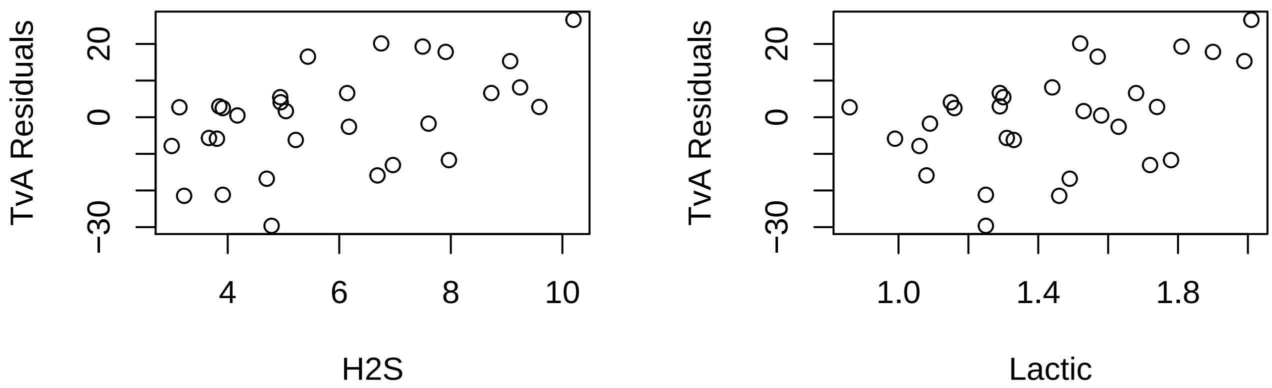
11.55

```r
tasteCol <- as.numeric(cheese[, "taste"])
aceticCol <- as.numeric(cheese[, "acetic"])
tasteVsAcetic <- lm(tasteCol ~ aceticCol, data.frame(cheese))
plot(aceticCol, tasteCol, xlab = "Acetic", ylab = "Taste")
abline(tasteVsAcetic)
```




```r
tVsAResiduals <- residuals(tasteVsAcetic)
plot(cheese[, "h2s"], tVsAResiduals, xlab = "H2S", ylab = "TvA Residuals")
plot(cheese[, "lactic"], tVsAResiduals, xlab = "Lactic", ylab = "TvA Residuals")

printRegEq(summary(tasteVsAcetic), "Acetic")
```

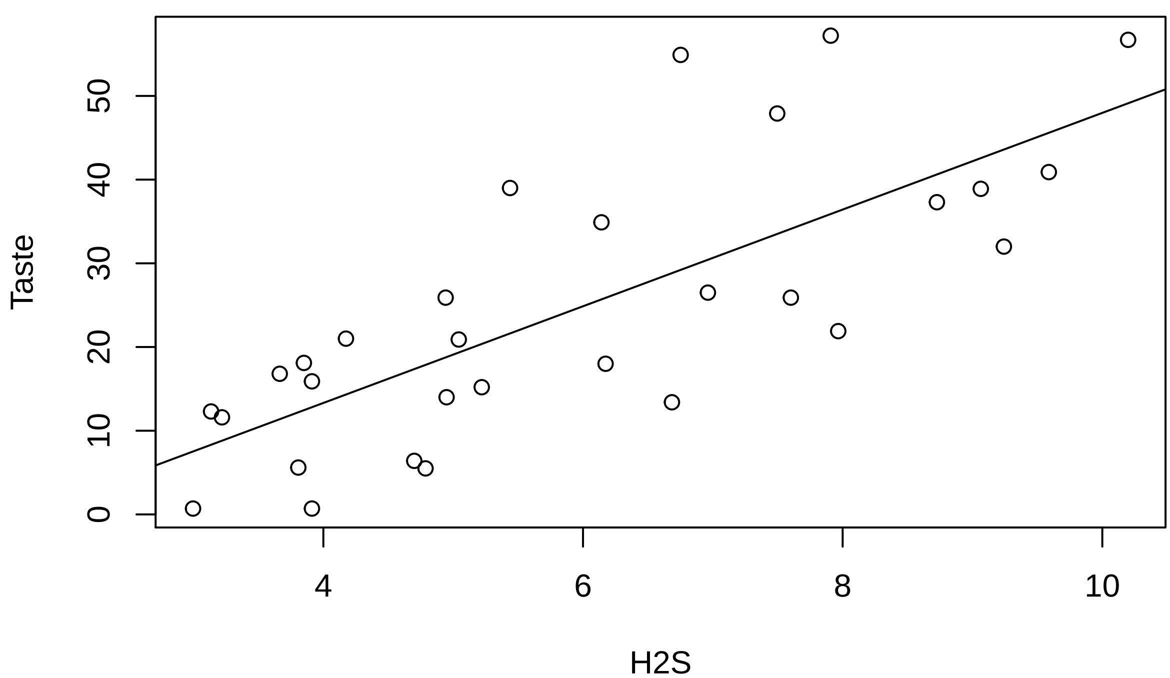
```
## Taste vs Acetic -2.475154 + Acetic * 3.480551
```



The residuals both have a normal distribution and seem to be positively associated with Lactic and H2S.

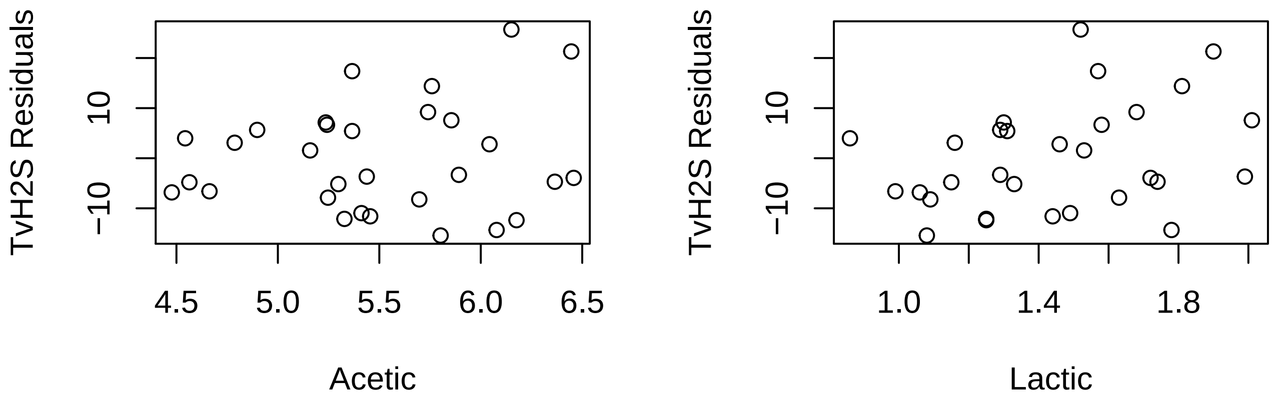
11.56

```r
h2sCol <- as.numeric(cheese[, "h2s"])
tasteVsH2S <- lm(tasteCol ~ h2sCol, data.frame(cheese))
plot(h2sCol, tasteCol, xlab = "H2S", ylab = "Taste")
abline(tasteVsH2S)
```



```
tVsHResiduals <- residuals(tasteVsH2S)
plot(cheese[, "acetic"], tVsHResiduals, xlab = "Acetic", ylab = "TvH2S Residuals")
plot(cheese[, "lactic"], tVsHResiduals, xlab = "Lactic", ylab = "TvH2S Residuals")

printRegEq(summary(tasteVsH2S), "H2S")
```

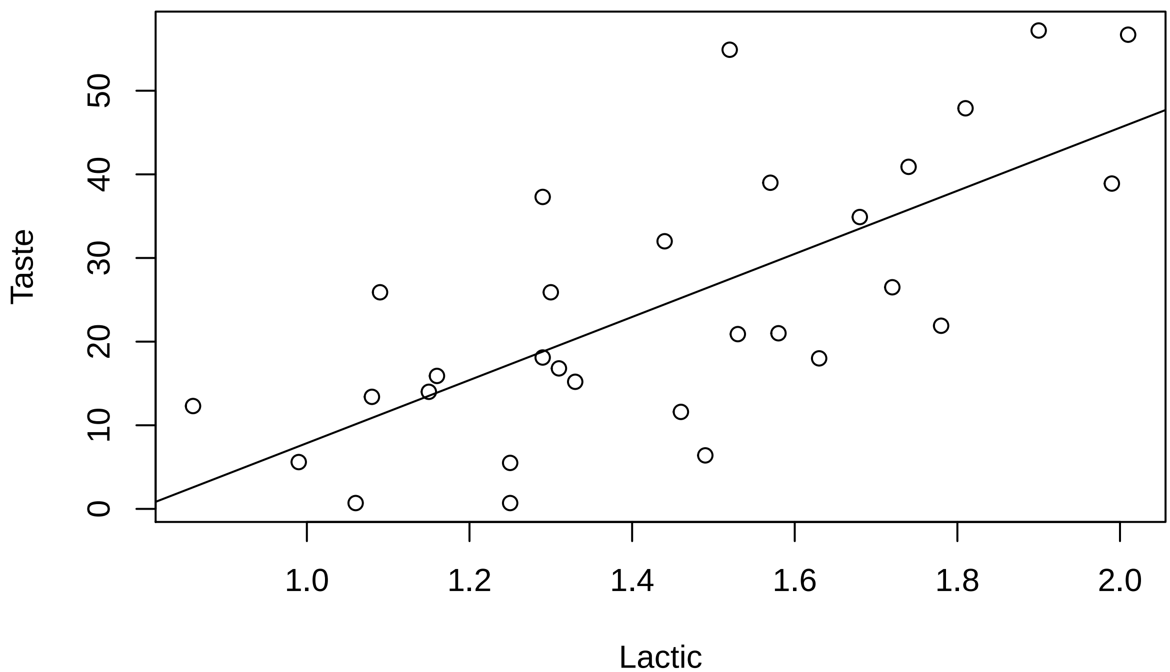
```
## Taste vs H2S -1.642663 + H2S * 6.10677
```



From the graphs there appears to be no correlation between the residuals and other variables.

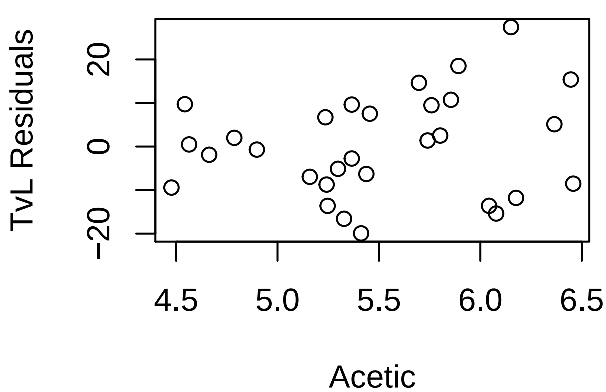
11.57

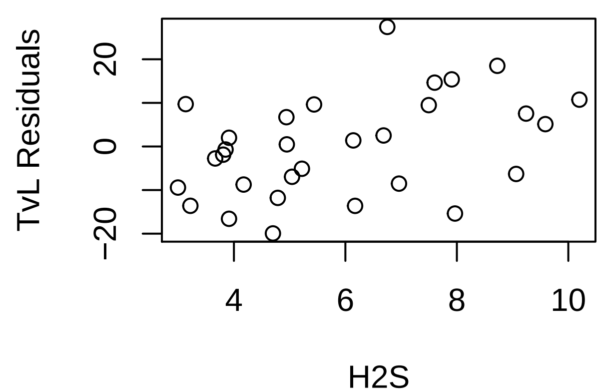
```
lacticCol <- as.numeric(cheese[, "lactic"])
tasteVsLactic <- lm(tasteCol ~ lacticCol, data.frame(cheese))
plot(lacticCol, tasteCol, xlab = "Lactic", ylab = "Taste")
abline(tasteVsLactic)
```




```
tVsLResiduals <- residuals(tasteVsLactic)
plot(cheese[, "acetic"], tVsLResiduals, xlab = "Acetic", ylab = "TvL Residuals")
plot(cheese[, "h2s"], tVsLResiduals, xlab = "H2S", ylab = "TvL Residuals")

printRegEq(summary(tasteVsLactic), "Lactic")
```

```
## Taste vs Lactic -2.821577 + Lactic * 5.248799
```





Again, there appears to be no correlation between the residuals and the other variables.

11.58

```r
tVsASum <- summary(tasteVsAcetic)
tVsHSum <- summary(tasteVsH2S)
tVsLSum <- summary(tasteVsLactic)

fStats  <- c(tVsASum$fstatistic[1], tVsHSum$fstatistic[1], tVsLSum$fstatistic[1])
pVals   <- c(tVsASum$coefficients[, 4][2], tVsHSum$coefficients[, 4][2],
             tVsLSum$coefficients[, 4][2])
rSqVals <- c(tVsASum$r.squared, tVsHSum$r.squared, tVsLSum$r.squared)
sdEst   <- c(tVsASum$sigma, tVsHSum$sigma, tVsLSum$sigma)

knitr::kable(data.frame(
  fStats, pVals, rSqVals, sdEst
))
```

	fStats	pVals	rSqVals	sdEst
aceticCol	12.11424	0.0016582	0.3019934	13.82124
h2sCol	37.29265	0.0000014	0.5711615	10.83338
lacticCol	27.54989	0.0000141	0.4959486	11.74504

```r
printRegEq(tVsASum, "Acetic")
```

```
## Taste vs Acetic -2.475154 + Acetic * 3.480551
```

```r
printRegEq(tVsHSum, "H2S")
```

```
## Taste vs H2S -1.642663 + H2S * 6.10677
```

```r
printRegEq(tVsLSum, "Lactic")
```

```
## Taste vs Lactic -2.821577 + Lactic * 5.248799
```

The intercepts in the three equations are different because the explanitory variables all have different values, leading to the points being plotted in different places. Since the linear equations are estimating the best fit line for these datapoints, it is only natural that the differing data produces different intercepts.

11.59

```r
tasteVsAnH <- lm(tasteCol ~ aceticCol+h2sCol, data.frame(cheese))
summary(tasteVsAnH)
```

```
##
## Call:
## lm(formula = tasteCol ~ aceticCol + h2sCol, data = data.frame(cheese))
##
## 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
## aceticCol      3.801      4.505   0.844 0.406245
## h2sCol         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
```

Taste^ = -26.94 + 3.801*acetic + 5.146*h2s

There is not much statistical significance of Acetic in this model, this loss in (Acetic) significance from the prior model, which used Acetic as the sole predictor, is most likely caused by the large positive correlation between H2S and Acetic. When Acetic is used to predict Taste in conjunction with H2S, there is not much significant information contributed by Acetic that H2S hasn't already done better.

Between using Acetic and H2S as a predictor model and using Acetic as the lone predictor model, I would prefer using the Acetic and H2S as a predictor model, due to its much larger overall significance level (7.635*10^-6)

11.60

```r
THLModel <- lm(tasteCol ~ (h2sCol + lacticCol), data.frame(cheese))
THLSum   <- summary(THLModel)

cat("Regression Equation for H2S and Lactic:",
    THLSum$coefficients[, 3][1], "+", "H2s+Lactic *", THLSum$coefficients[, 3][2], "\n")
```

```
## Regression Equation for H2S and Lactic: -3.071961 + H2s+Lactic * 3.474768
```

```r
lmp <- function(fStats) {
  return (pf(fStats[1], fStats[2], fStats[3], lower.tail = F))
}

cat("P-Value:", lmp(THLSum$fstatistic), "\n")
```

```
## P-Value: 6.551371e-07
```

Since the p-value of the model with both variables is far less than the p-values of the variables by themselves, it is a much better fit for predicting cheese taste.

11.61

```r
tasteVall <- lm(tasteCol ~ (h2sCol + lacticCol + aceticCol), data.frame(cheese))
summary(tasteVall)
```

```
##
## Call:
## lm(formula = tasteCol ~ (h2sCol + lacticCol + aceticCol), data = data.frame(cheese))
##
## 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
## h2sCol        3.9118     1.2484   3.133  0.00425 **
## lacticCol    19.6705     8.6291   2.280  0.03108 *
## aceticCol     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
```

```r
summary(THLModel)
```

```
##
## Call:
## lm(formula = tasteCol ~ (h2sCol + lacticCol), data = data.frame(cheese))
##
## 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 **
## h2sCol         3.946      1.136   3.475  0.00174 **
## lacticCol     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
```

Of all the Model's computed