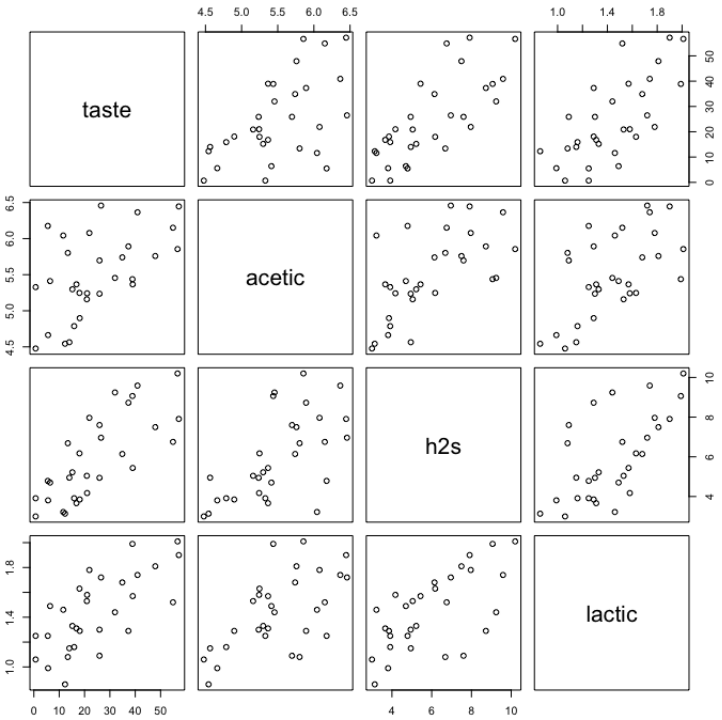
ST512 – Lab 03 *Solutions*

1. (a) Get pairwise correlations and scatterplots for all four variables. Is the response variable linearly related to the predictor variables individually?

   We can evaluate whether the response is linearly related to the three predictors individually by looking at a scatterplots, shown below.



   None of the individual predictor variables are strongly related to `taste`, though both `h2s` and `lactic` have a relatively weak linear relationship with `taste`. This is supported by the correlations reported below, which shows that `taste` and `h2s` have correlation 0.756 and `taste` and `lactice` have correlation 0.704.

```
           taste    acetic       h2s    lactic
taste  1.0000000 0.5495393 0.7557523 0.7042362
acetic 0.5495393 1.0000000 0.6179559 0.6037826
h2s    0.7557523 0.6179559 1.0000000 0.6448123
lactic 0.7042362 0.6037826 0.6448123 1.0000000
```

   (b) Justify this claim: variable among these three predictor variables, `h2s` has the strongest relationship with the response.

   This claim can be evaluated using the sample correlation coefficients. From the correlation coefficient table, we see that the the correlation between `taste` and `acetic` is 0.550, the correlation between `taste` and `h2s` is 0.756, and the correlation between `taste` and `lactic` is 0.7046. The strength of the linear relationship between two variables is measured by the absolute value of the correlation coef-

ficient. Of those three correlation coefficients, the highest absolute value is the correlation coefficient between `taste` and `h2s` which supports the claim.

(c) Fit a simple linear regression model of `taste` versus that `h2s`. Find the estimated slope parameter and, based on the plots produced, comment on the quality of the model fit and on appropriateness of the model assumptions.

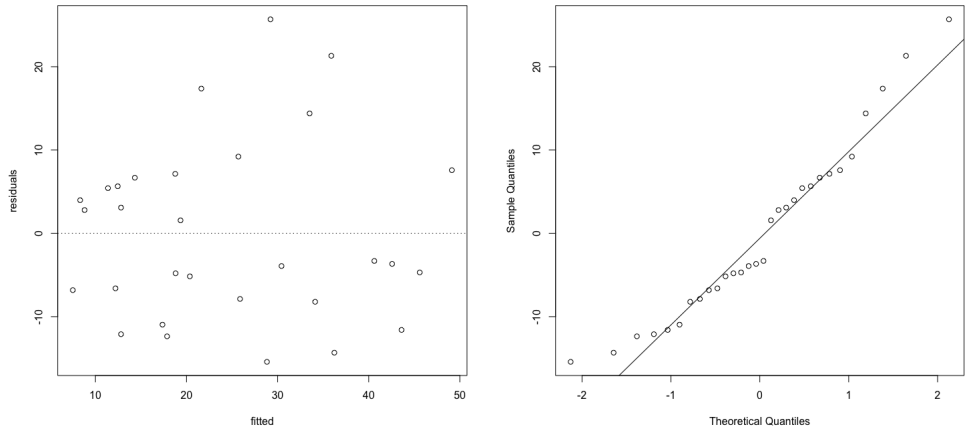
The `lm` output is below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.7868     5.9579  -1.643    0.112
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
```

From the `lm` output below, we can see that the estimated slope parameter is 5.776. One measure of how well the model fits the data is R^2. For this model, $R^2 = 0.5712$. This means that 57.12% of the observed variation in `taste` can be attributed to the linear relationship between `taste` and `h2s`. This indicates that the model fits the data moderately well.

The following residual plot can be used to assess some of the model assumptions. Indeed, there is no clear trend suggesting a possible non-linear relationship. Also, since this is a simple linear regression, the residuals versus `h2s` would look identical,[1] just with a different scale on the x-axis. There's no clear trend (i.e., no obvious megaphone shape) that suggests inconsistencies with the model assumptions. The QQ plot of the residuals looks great, points tightly hug the line.



[1]This is because fitted values are linear in the single predictor variable.

2. Since we have three predictor variables available, we investigate their joint effect on the response. Fit a multiple linear regression model with all three predictors.

   (a) Find the equation of the fitted least-squares regression line. Compare the new estimated slope for `h2s` to that from Problem 1(c) above. Should the two estimates be the same? Explain.

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
acetic        0.3277     4.4598   0.073  0.94198
h2s           3.9118     1.2484   3.133  0.00425 **
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
```

   From the output, the fitted least squares regression line is

$$\hat{y}(x_1, x_2, x_3) = -28.877 + 0.328\, x_1 + 3.912\, x_2 + 19.671\, x_3,$$

   where x_1, x_2, and x_3 represent generic values of the `acetic`, `h2s`, and `lactic` variables, respectively.

   The estimated slope for `h2s` in the multiple linear regression model is 3.912. In Problem 1(c), we found that the estimated slope for `h2s` in the simple linear regression model was 5.776. We should not expect these two values to be the same because the interpretation of the slope for `h2s` is not quite the same in both models. In the simple linear regression model, the slope for `h2s` is a measure of the marginal association between `h2s` and `taste` and is interpreted as the estimated change in `taste` associated with a 1-unit increase in `h2s`. This ignores any possible effect or association between `taste` and the other two predictors we considered in this exercise, `acetic` and `lactic`. In the multiple linear regression model, the slope for `h2s` is instead the expected change in `taste` associated with a 1-unit increase in `h2s`, when `acetic` and `lactic` are both held constant.

   (b) Find the F test p-value given in the output. State the null and alternative hypotheses being tested with that p-value, and your conclusion.

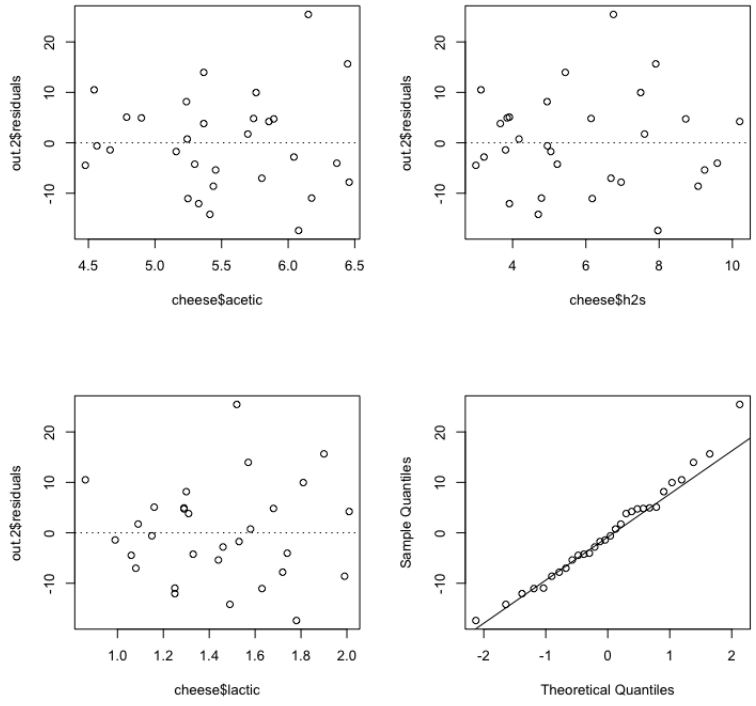
   Let β_1, β_2, and β_3 be the slopes of `acetic`, `h2s`, and `lactic`, respectively. The null hypothesis associated with the test performed is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. The alternative hypothesis is H_1 : at least one of β_1, β_2, or β_3 is non-zero. The p-value from this hypothesis test is 3.8×10^{-6}, incredibly small. Therefore, we'd be inclined to reject the null hypothesis and conclude that at least one of the three variables `acetic`, `h2s`, or `lactic` is important for explaining variation in `taste`.

(c) Give a 95% confidence interval for the slope attached to `h2s` and use this to carry out a test of the null hypothesis that `h2s` and `taste` are not related.

The 95% confidence interval for the slope attached to `h2s` is given in the output to be (1.3457, 6.4780). To test whether or not `h2s` and `taste` are related, the null hypothesis is $H_0 : \beta_2 = 0$ and the alternative hypothesis is $H_1 : \beta_2 \neq 0$. Because 0 does not fall in the 95% confidence interval, we would reject the null hypothesis and instead conclude that `h2s` and `taste` are related.

(d) Look at the plot of the residuals versus the three predictor variables. Any indication that the assumption of constant error variance is violated?

Under the assumption of constant error variance, we would expect the range of the residuals to be the same across different values of the predictor variables. In all three plots (excluding bottom right—see below), it appears that the variation of the residuals remains constant across the different values of the predictor variables. This supports the assumption of constant error variance.



(e) Find the normal QQ plot of the residuals. What does this plot tell you?

The plot is shown in the bottom right corner of the previous figure. The QQ plot is a tool to use to check the assumption of normality for the residuals. If the residuals are normally distributed, then the points would fall "close to" the line. This QQ plot looks very good in the sense that all points fall very close to the line; even the one possible exceptional point isn't that far off. Therefore, I find that this plot supports the assumption of normality for the errors.

3. Get a 90% prediction interval for the actual `taste` value for a new cheese with the following predictor variable values: `acetic` = 10, `h2s` = 5.5, and `lactic` = 4.

   After creating a new dataset with an observation corresponding to the desired values for the predictors, a 90% prediction interval can be found. In this case, the 90% prediciton interval for `taste` associated with a new cheese having `acetic` = 10, `h2s` = 5.5, and `lactic` = 4 is (29.063, 120.133).