Homework 2 Peer Assessment

Question 1: Exploratory Data Analysis [12 points]

(a) Plot the data (scatterplot) to observe and report the relationship between the response and each of the three predictors (there should be 3 plots reported). Comment on the general trend (direction and form).

Answer:

```
rm(list=ls())
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 3.5.2
```

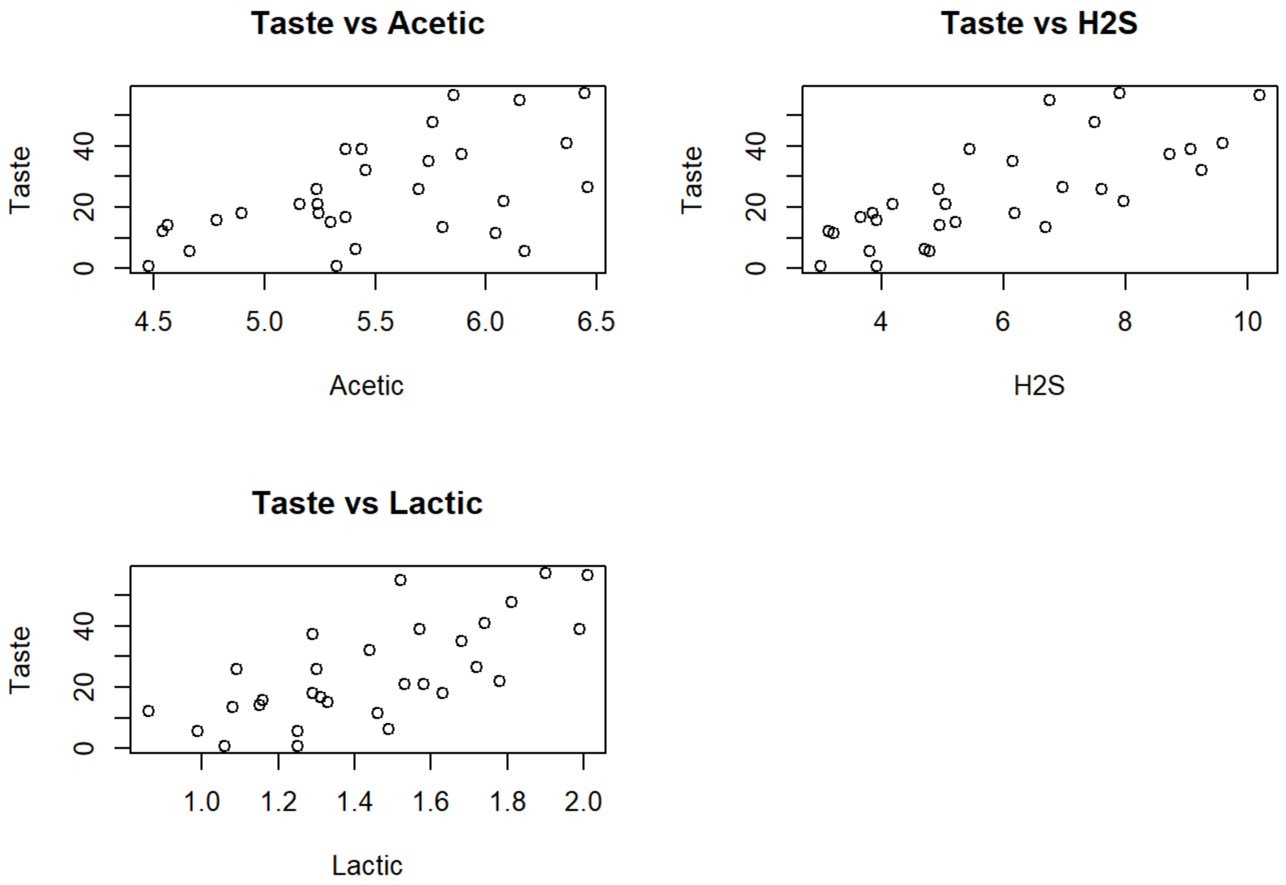
```
head(cheddar)
```

```
##   taste Acetic   H2S Lactic
## 1  12.3  4.543 3.135   0.86
## 2  20.9  5.159 5.043   1.53
## 3  39.0  5.366 5.438   1.57
## 4  47.9  5.759 7.496   1.81
## 5   5.6  4.663 3.807   0.99
## 6  25.9  5.697 7.601   1.09
```

```
dim(cheddar)
```

```
## [1] 30  4
```

```
par(mfrow = c(2,2))
plot(taste ~ Acetic, data = cheddar, xlab = 'Acetic', ylab='Taste', main = 'Taste vs
Acetic')
plot(taste ~ H2S, data = cheddar, xlab= 'H2S', ylab = 'Taste', main = 'Taste vs H2S')
plot(taste ~ Lactic, xlab = 'Lactic', ylab = 'Taste', data = cheddar, main = 'Taste v
s Lactic')
```



Answer: These three plots show the positive relationship between the response and each of the three predictors individually, that is, if each of three predictors increase, the scores of taste will increase too. In addition, there are probably linear relationship.

(b) What is the value of the correlation coefficient for each of the above pair of response and predictor variables? What does it tell you about your comments in part (a).

```
cor(cheddar$Acetic, cheddar$taste)
```

```
## [1] 0.5495393
```

```
cor(cheddar$H2S, cheddar$taste)
```

```
## [1] 0.7557523
```

```
cor(cheddar$Lactic, cheddar$taste)
```

```
## [1] 0.7042362
```

Answer: The results show that these three corrlation coefficient are strong, especially for H2S-taste and Lactic-taste, that is, the linear relationship between response and predicting variables are strong.

(c) Based on this exploratory analysis, is it reasonable to assume a multiple linear regression model for the relationship between taste and all the predictor variables (Acetic, H2S and Lactic)? Did you note anything unusual?

Answer: Yes, it is reasonable to assume a multipe linear regerssion model. The unusual is that the distribution of dots seem to be scattered towards lower right corner in the first plot between taste and Acetic. In addtion, the concentrations of acetic acid and lactic acid use log scale.

(d) Based on the analysis above, would you pursue a transformation of the data?

Answer: We don’t need any more for plots of taste-H2S and taste-Lactic, because the concentration lactic acid has used log scale. But for plot of taste-Lactic, we maybe need some improvement by using transformation.

Please work on non-transformed data for all of the following questions.

Question 2: Fitting the Multiple Linear Regression Model [8 points]

Build a multiple linear regression model using the response and all the three predictors and then answer the questions that follow:

(a) Report the R2 for the model and give a single line interpretation of the same.

```
model_lm <- lm(taste ~., data = cheddar)
summary(model_lm)
```

```
##
## Call:
## lm(formula = taste ~ ., data = cheddar)
##
## 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
## 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
```

Answer: The R squared is 0.6518, which means the percentage of total variability in Y (taste) can be explained by the linear regression model.

(b) Identify the predictors that are statistically significant at the 5% and 10% level. Which extra predictor(s) become significant at the 10% level, as compared to the 5% level?

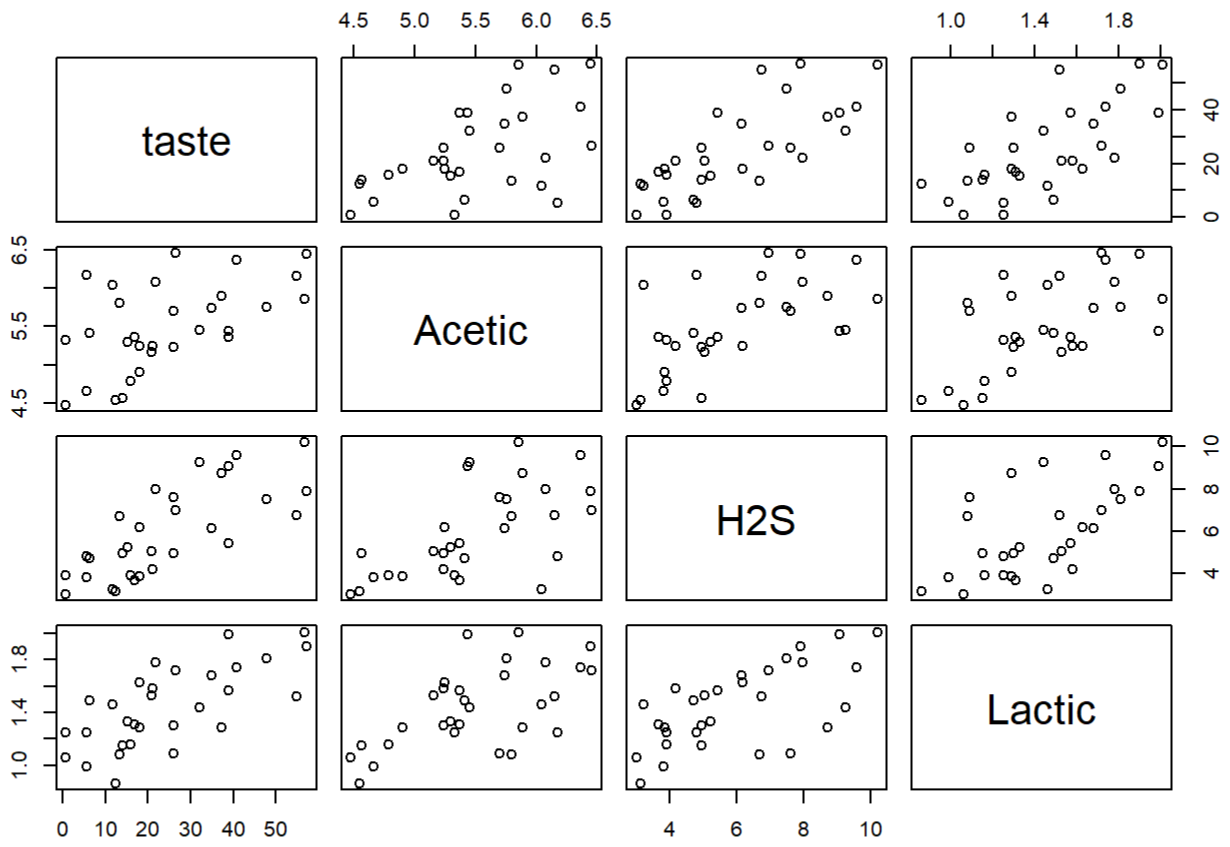
Answer: The predictor that are statistically significant at the 5% and 10% is H2S and Lactic. There are no extra predictors are significant at the 10% level compared to the 5% level.

Question 3: Checking Assumptions of Model and Coefficient Interpretation [14 points]

(a) Provide plots to check for Linearity, Constant Variance and Normality assumptions of the model (use your knowledge from Homework 1 Peer Assessment). Provide your interpretations (i.e. whether the assumptions hold) for each plot.

Answer: (1). To check the linearity, we use scatterplot for the relationship between taste and Acetic, taste and H2S, tast and Lactic. We can see from the graphical display below, which show the linearity assumption between response and predictor virables.

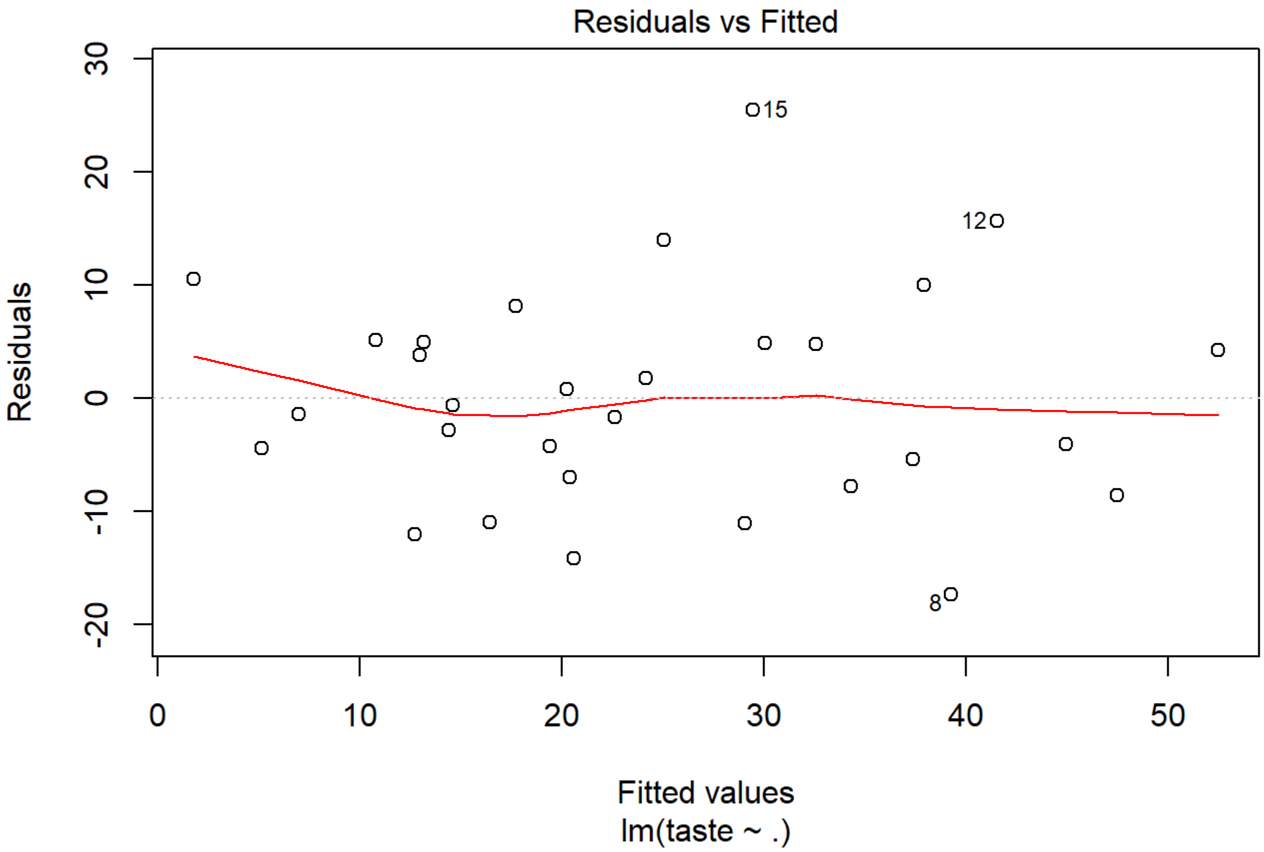
```
pairs(taste ~., data = cheddar)
```



```
#plot(model_lm)
# plot(cheddar)
```

(2). In order to evaluate the constant variance and independence assumption, we can use a scatter plot of the fitted values against the residuals. From the figure below, most of residuals are scatted around the 0 line, which indicates that we do have constant variance.

```
#plot(model_lm$fitted, model_lm$residuals, main = "Fitted vs Residuals"
plot(model_lm, 1)
```



(3). To evaluate the normality, we use the normal probability plot. Under the ideal condiction, the plot of residuals should appropriately follow a straight line. From the figure below, we can see most the points follow a straight line very well.

```
library(car)
```

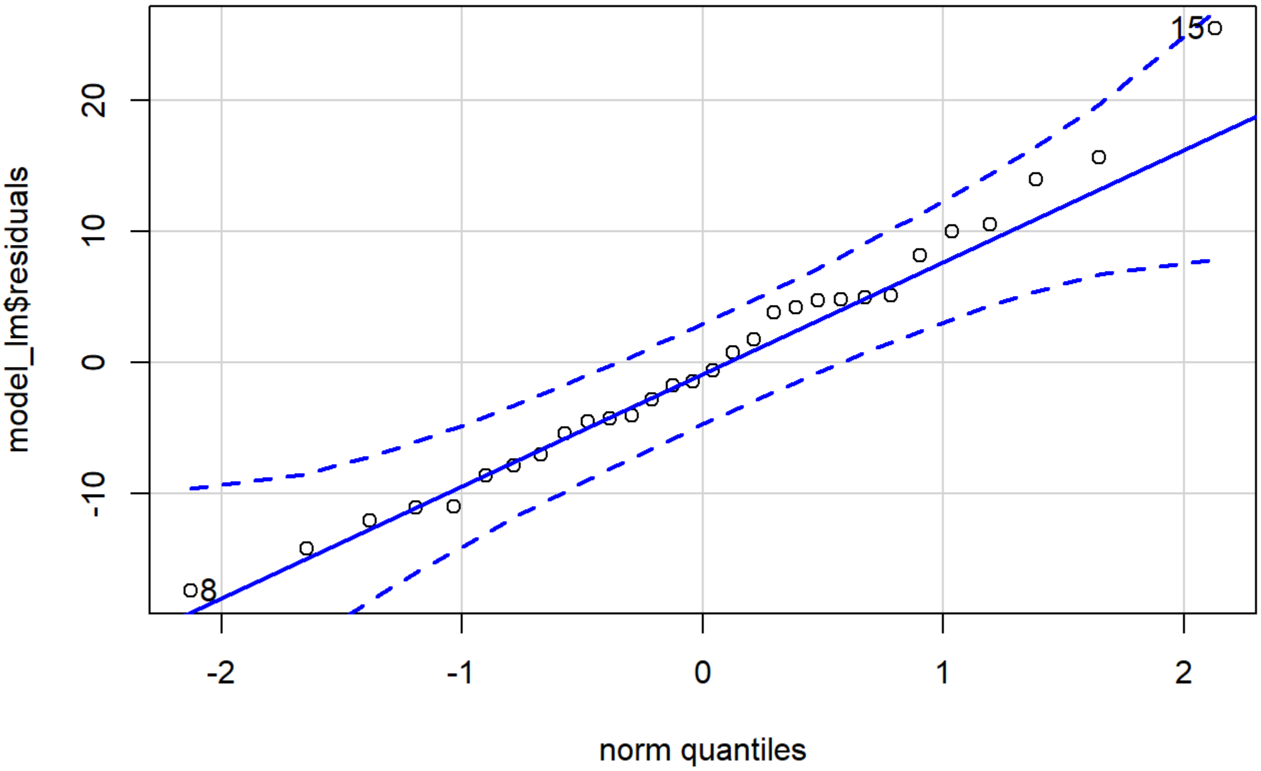
```
## Warning: package 'car' was built under R version 3.5.2
```

```
## Loading required package: carData
## Warning: package 'carData' was built under R version 3.5.2
```

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```
qqPlot(model_lm$residuals)
```



```
## [1] 15  8
```

```
#hist(model_lm$residuals)
```

(b) Interpret the coefficient of Acetic (mention any assumption you make about other predictors clearly when stating the interpretation).

Answer: The estimated coefficient of Acetic is 0.3277, which means the expected additional gain in tasted scores for each additional Acetic (log scale), while holding all other predictors fixed. In the linear-log model, the literal interpretation of the estimated coefficient hat(b) is that a one-unit increase in logX will produce an expected increase in Y of hat(b) units.

(c) If value of predictor H2S in the above model is increased by 0.01 keeping other predictors constant, what change in the response would be expected?

Answer: Because the estimated coefficient of H2S is 3.9118, so if increases by 0.01 in H2S will result in incresing taste scores by 0.0319118, while keeping other predictors constant.

Question 4: Log Scale [6 points]

In the given cheddar data, assume Acetic and H2S measured were actually on a log scale. What is the percentage change in H2S on the regular scale corresponding to an additive increase of 0.01 on the (natural) log scale?

Answer: Because ln(1+0.01) = 1% (approx.), a one percent increase in the independent variable increases the dependent variable by (coefficient/100) units, that is, a 1 percent increase in X increases (natural) log(X) by .01 and, therefore, changes the Y variable by .01 * coeffieient(H2S).

Question 5: Confidence Intervals and Interpretation [10 points]

Compute 90% and 95% confidence intervals (CIs) for the parameter H2S for the model in Question 2. Using just these intervals, what could you deduce about the range (Upper Bound or Lower Bound or both) of p-value for H2S in the regression summary for model in Question 2?

Answer: For the parameter H2S, the 90% confidence intervals range (1.782496, 6.041186) for regression coefficients, which is positive result of regression coefficient when given all other predicting variables in the model. The 95% confidence intervals range (1.345656, 6.478026) for regression coefficients, which are positive result of regression coefficient when given all other predicting variables in the model.

```
confint(model_lm, 'H2S', level = 0.90)
```

```
##          5 %     95 %
## H2S 1.782496 6.041186
```

```
confint(model_lm, 'H2S', level = 0.95)
```

```
##        2.5 %   97.5 %
## H2S 1.345656 6.478026
```