# HW2_Sol

Manish Mehta

January 2, 2019

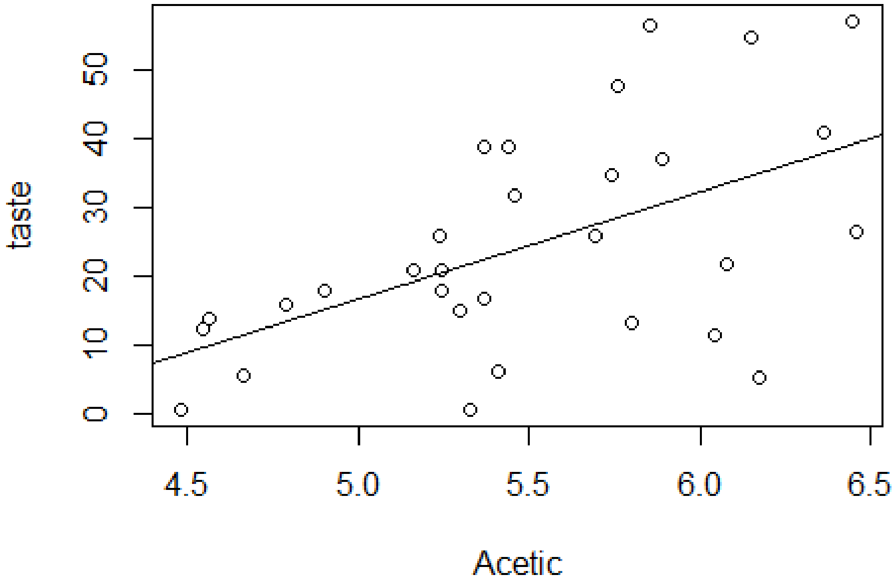
```
# install.packages("faraway")
library(faraway)
data(cheddar)
```

## Question 1

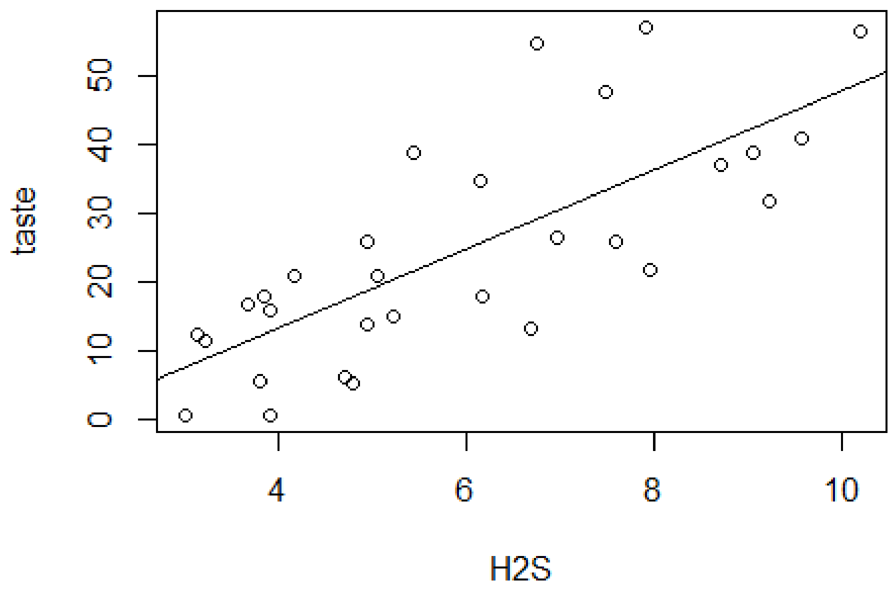
### (a) Plot the data (scatterplot) to observe and report the relationship between the response and each of the three predictors (there should be 3 plots reported). Comment on the general trend (direction and form).

Using the code below, we get the scatterplots as shown:

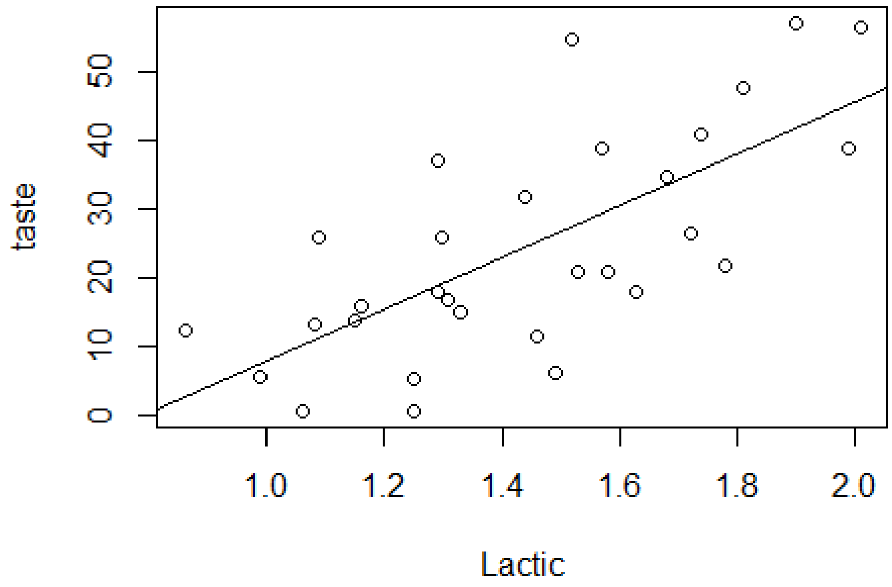
```
# Plots
plot(taste~Acetic, data=cheddar)
abline(lm(taste~Acetic, data=cheddar)) # not necessary, just for better visua
lization
```



```r
plot(taste~H2S, data=cheddar)
abline(lm(taste~H2S, data=cheddar)) # not necessary, just for better visualization
```



```r
plot(taste~Lactic, data=cheddar)
abline(lm(taste~Lactic, data=cheddar)) # not necessary, just for better visualization
```



General trend : In each plot, there seems to be a positive and linear relationship between the response (taste) and each of the predictor variables (Acetic, H2S and Lactic for each plot respectively)

**(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).**

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

## [1] 0.5495393

cor(cheddar$taste, cheddar$H2S)

## [1] 0.7557523

cor(cheddar$taste, cheddar$Lactic)

## [1] 0.7042362
```

The correlations between:

taste and Acetic = 0.5495393 taste and H2S = 0.7557523 taste and Lactic = 0.7042362

This shows that there is a positive correlation between response and each of the predictor variables. This shows that our comments about the general trend for each of the plots were correct and it aligns with our hypothesis that the response is positively correlated with each of the predictor variables.

### (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?

Yes, that is a resonable assumption. There is nothing unusual to be seen.

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

No, there will not be any transformation because there is a linear relationship between the response and each of the predictors.

*Grading Scheme*

| Response Quality | Description | Points (out of 16) |
|---|---|---|
| Poor | The student does not answer any parts of the question correctly | 0 |
| OK | The student does 1 of 4 parts correctly | 3 |
| Good | The student does 2 of 4 parts correctly | 6 |
| Great | The student does 3 of 4 parts correctly | 9 |
| Perfect | The student answers all parts of the question correctly | 12 |

## Question 2

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

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

## 
## 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
```

#### (a) Report the R-sq for the model and give a single line interpretation of the same.

R-sq is 65.18% and interpretation is : 65.18% of the variation in the response is explained by the predictors.

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

- Predictors significant at 5% level : H2S and Lactic
- Predictors significant at 10% level : H2S and Lactic

No extra predictors became significant at 10% level as compared to 5% level

*Grading Scheme*

| Response Quality | Description | Points (out of 16) |
|---|---|---|
| Poor | The student does not answer any parts of the question correctly | 0 |
| OK | The student does only part (b) correctly | 3 |
| Good | The student does only part (a) correctly | 5 |
| Perfect | The student answers all parts of the question correctly | 8 |

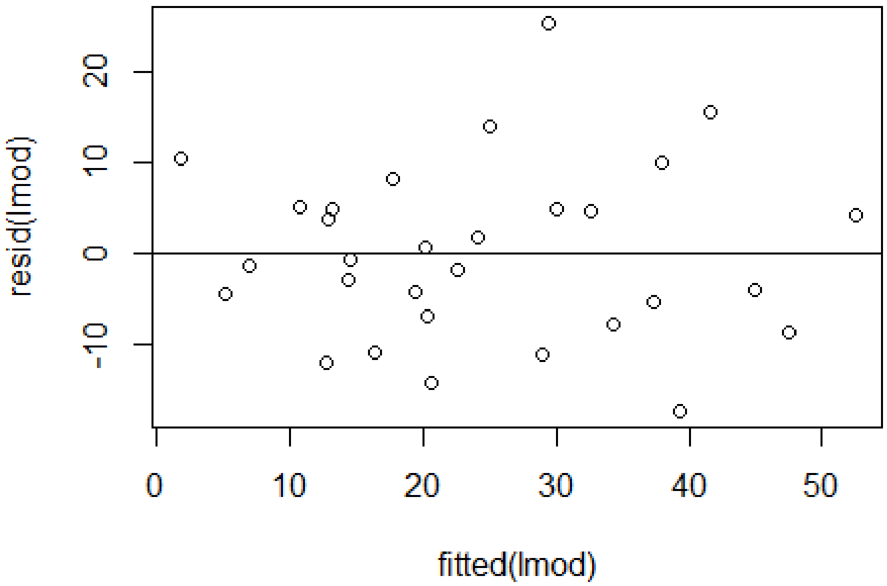
## Question 3

### (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.

For Linearity assumption, the plots should be the same 3 scatterplots as shown in part (a) of Question 1, and the code should be the same as part (a) of Question 1 too. Interpretation: The Linearity assumption holds.

For Constant variance assumption, the code is below:

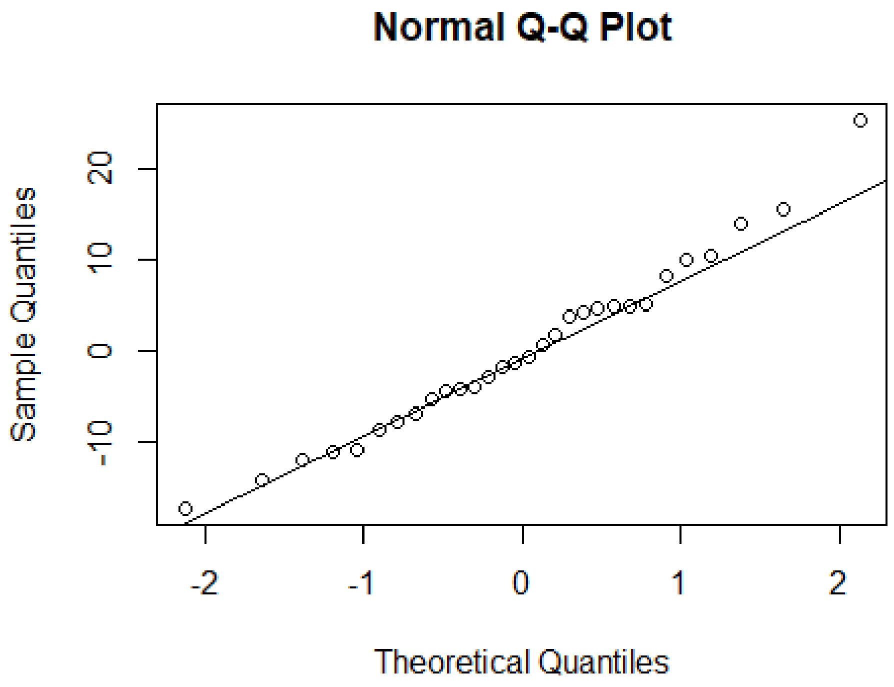
```r
plot(fitted(lmod),resid(lmod))
abline(h=0) # not necessary, just for better visualization
```



Since the residuals are randomly scattered across zero line, the Constant variance assumption holds.

For Normality assumption, the code is below:

```
qqnorm(resid(lmod))
qqline(resid(lmod))
```



From the plot, we can observe that (except for the rightmost point) the Normality assumption holds.

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

The coefficient of Acetic is = 0.3277 which means that with an increase in 1 unit of predictor Acetic (concentration of Acetic acid on log scale (should mention this)), there is an increase in taste score by 0.3277 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?**

Since the coefficient of H2S is 3.9118 in the regression equation, with an increase in 0.01, the value of taste will increase by = 3.9118 X 0.01 = 0.039 units

*Grading Scheme*

| Response Quality | Description | Points (out of 16) |
|---|---|---|
| Poor | The student does not answer any parts of the question correctly | 0 |
| OK | The student does only part (a) correctly | 6 |

| | | |
|---|---|---|
| Good | The student does part (a) and any one of (b) or (c) correctly | 10 |
| Perfect | The student answers all parts of the question correctly | 14 |

## Question 4

### In the given cheddar data, we see from Data Description section of the question that 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?

Since H2S was actually measured on a log scale, we have H2S = log(H2S_orig) let's say. Thus, for an increase of 0.01 on the (natural) log scale, we have new value = H2S + 0.01 = log(H2S_orig) + 0.01 = log(H2S_orig X exp(0.01)) = log(1.01005 H2S_orig)

So, H2S_new = exp(log(1.01005 H2S_orig)) = 1.01005 H2S_orig

Thus, there is an increase in H2S on the original scale by = ((H2S_new - H2S_orig)/H2S_orig) x 100% = 1.005%

*Grading Scheme*

| Response Quality | Description | Points (out of 16) |
|---|---|---|
| Poor | The student does not answer the question correctly | 0 |
| OK | The student has the right approach, which they demonstrate in the solution, but does not reach the correct answer | 3 |
| Perfect | The student answers all parts of the question correctly | 6 |

## Question 5

### 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?

For the 90% CI:

```r
confint(lmod, "H2S", level = 0.9)
```

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

For the 95% CI:

```
confint(lmod, "H2S", level = 0.95)

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

Since none of the CIs include 0, it means that H2S is statistically significant at both levels. Hence, p-value for H2S is < 0.05. Thus, Upper bound is 0.05, Lower bound is 0 (which even if the student doesn't mention, it's fine)

*Grading Scheme*

| Response Quality | Description | Points (out of 16) |
|---|---|---|
| Poor | The student does not answer the question correctly | 0 |
| OK | The student finds just one CI correctly and does not find a bound for p-value | 4 |
| Good | The student finds both CIs correctly but not the bound for p-value | 7 |
| Perfect | The student answers all parts of the question correctly | 10 |