HW 2

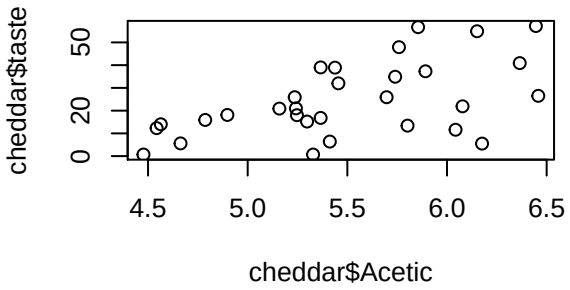
Jeff Tilton

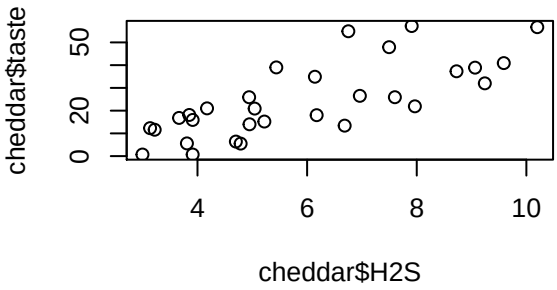
2/4/2019

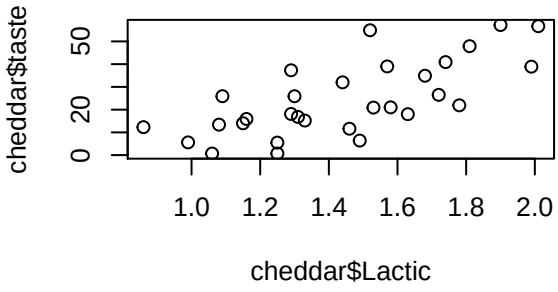
```
library(faraway)
data(cheddar)
```

```
par(mfrow=c(2,2))
p1 = plot(cheddar$Acetic,cheddar$taste)
p2 = plot(cheddar$H2S, cheddar$taste)
p3 = plot(cheddar$Lactic,cheddar$taste)

cor_acetic = round(cor(cheddar$Acetic,cheddar$taste),3)
cor_H2S = round(cor(cheddar$H2S, cheddar$taste),3)
cor_lactic = round(cor(cheddar$Lactic,cheddar$taste),3)
```







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).

The above plots all exhibit a positive, linear correlation between the given predictor and taste. The log concentration of hydrogen sulfice (H2S) has the strongest relationship followed by lactic acid and finally acetic acid.

(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).

Table 1: Taste Correlation table. The below values are the correlation coefficients for each predictor and taste

Lactic Acid	H2S	Acetic Acid
0.704	0.756	0.55

The above table confirms the visual inspection in part a. The values are positive and (H2S) has the strongest relationship followed by lactic acid and finally acetic acid.

(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?

The results of (a) and (b) suggest that a multiple linear regression is a reasonable assumption. The only thing to note is that the relationship between taste and acetic acid is much weaker than the other predictors.

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

I had not intended to do any transformations on the above data, because none of the data looked like it would benefit from a transform. However, because the question was asked I tried to do a simple log transform and then a box-cox transform on the Acetic acid data. It did not change the plot or correlation enough to continue on with the transformed data.

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:

```
model = lm(taste~.,data=cheddar)

summary(model)
```

```
##
## 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 R2 for the model and give a single line interpretation of the same.

The R^2 as seen above is 0.65 and, as defined in the book, is a quantitative measure of how well the fitted model predicts the dependent variable, taste.

(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?

H2S is significant below the 0.01 level (p-value$<$ .01), lactic acid is significant below the 0.05 level (p-value $<$ 0.05), acetic acid is not significant (p-value$>$ 0.1). I do not see an extra predictors significant at the 10% 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.

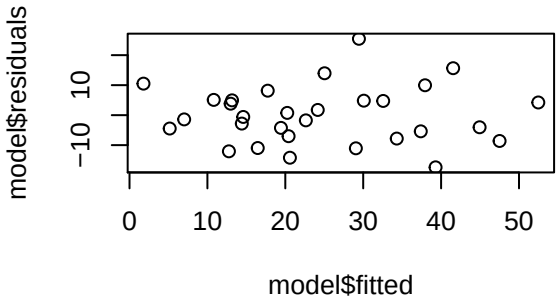
```
library(car)
```

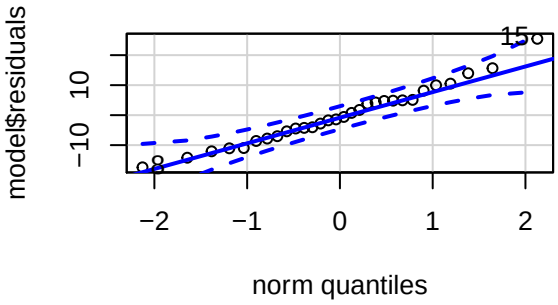
```
## Loading required package: carData
```

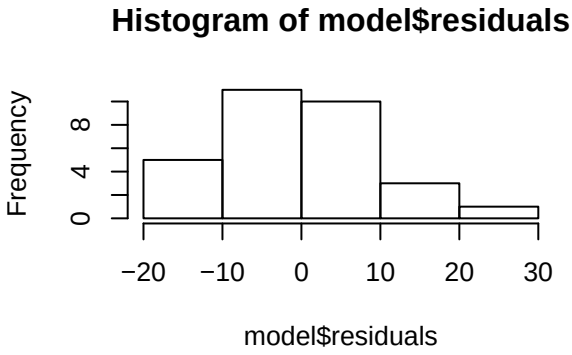
```
##
## Attaching package: 'car'
```

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

```
par(mfrow=c(2,2))
p2 = plot(model$fitted, model$residuals)
p3 = qqPlot(model$residuals)
p4 = hist(model$residuals)
```







Linearity

The first three plots from the exploratory data analysis in question 1 show each predictor vs the response, taste. The results indicate a linear relationship between each of the predictors and the response. This assumption holds.

Constant variance

The residual plot above demonstrates homoscedasticity with the residuals scattered around 0. Therefore the constant variance assumption holds.

Independence

The residual plot does not show any pattern or clustering, which can indicate a lack of independence in the data. Although not a true test of independence this is a good proxy and I say the assumption holds.

Normality

The last two plots, q-q and histogram, demonstrate normally distributed errors, with one possible outlier. The q-q plot has a point on the extreme high end that is almost out of the confidence interval and the histogram is right skewed. Although not perfect I would say this assumption holds.

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

The Acetic Acid coefficient is 0.327. This represents the estimated expected change in taste with one unit of change in acetic acid, holding H2S and lactic acid constant.

(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?

If there is a 0.01 increase in H2S there would be an expected change in taste of 0.039118, while holding other model predictors constant.

Question 4: Log Scale

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?

A percentage change in a value is roughly equal to the log difference of that value. So $\frac{x_2-x_1}{x_1} \approx ln(x_2) - ln(x_1)$. So in our example, an increase in .01 of the natural log would require a .01*100 = 1% increase in the coefficient of H2S, if it was not in log form.

Question 5: Confidence Intervals and Interpretation

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?

```
confint(model, level = .9)
```

```
##                     5 %      95 %
## (Intercept) -62.537853  4.784314
## Acetic       -7.278899  7.934382
## H2S           1.782496  6.041186
## Lactic        4.952673 34.388414
```

```
confint(model, level = .95)
```

```
##                    2.5 %    97.5 %
## (Intercept) -69.443503 11.689964
## Acetic       -8.839420  9.494902
## H2S           1.345656  6.478026
## Lactic        1.933267 37.407820
```

I found the below online and it helped me understand the relationship between the p-value and confidence intervals.

> You can use either P values or confidence intervals to determine whether your results are statistically significant. If a hypothesis test produces > both, these results will agree.
>
> The confidence level is equivalent to 1 – the alpha level. So, if your significance level is 0.05, the corresponding confidence level is 95%.
>
> - If the P value is less than your significance (alpha) level, the hypothesis test is statistically significant.
> - If the confidence interval does not contain the null hypothesis value, the results are statistically significant.

- If the P value is less than alpha, the confidence interval will not contain the null hypothesis value.

Looking at our example, the p-value from the question 2 summary for H2S is 0.00425. This is less than both the 0.1 and 0.05 significance level (90% and 95% confidence intervals). Neither confidence interval contains the null hypothesis value 0 ($H_0 : \beta_{H2S} = 0$) so the confidence intervals and p-value agree as expected.