Project #1

Dustin Leatherman

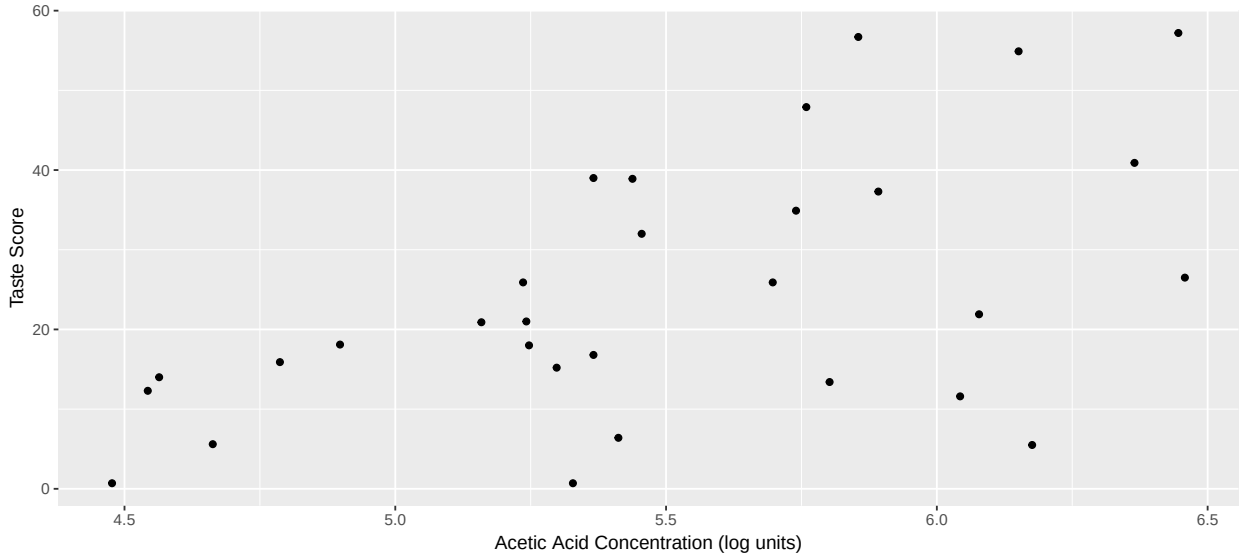
10/5/2019

1

Goal: Determine the relationship between chemical presence in cheddar cheese and its tatse.

a

Construct a scatter plot of taste (y-axis) and acetic acid (x-axis). Interpret the relationship between the variables using the plot.



There appears to be a positive correlation between Log Acetic Acid Concentration and Taste Score. As concentration of Acetic Acid increases, Taste Score tends to increases.

b

Find the estimated regression model with acetic acid as the predictor variable and taste as the response variable.

```
cheese.model1 <- lm(taste ~ Acetic, data = cheese)
```

$$\hat{Y_i} = -61.499 + 15.648 \text{ Acetic}$$

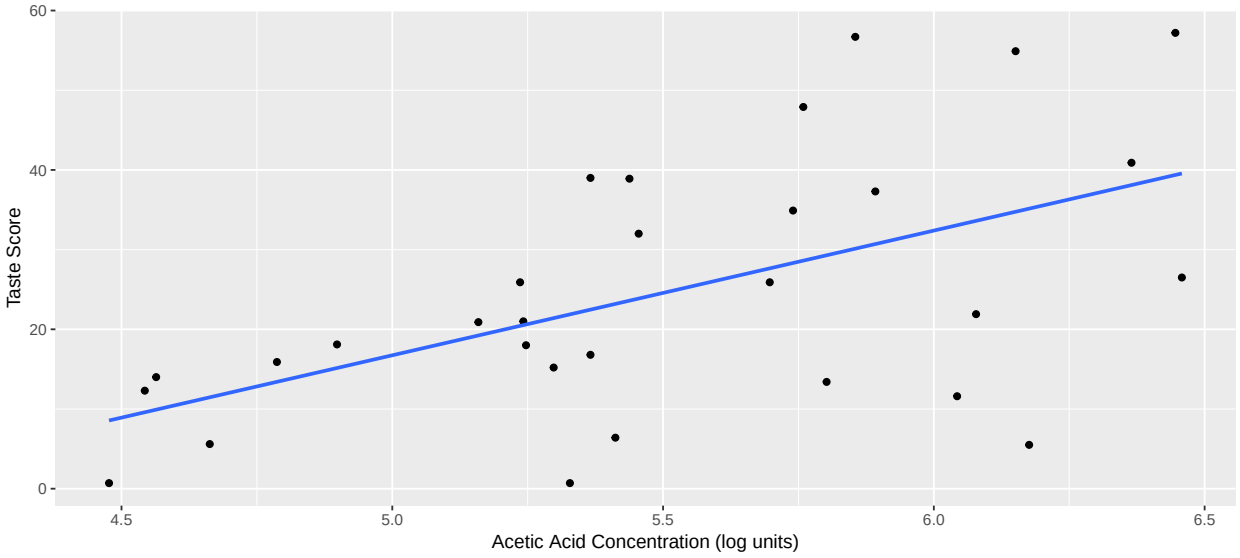
c

Interpret the relationship between the variables using the regression model from (b). Describe how your interpretation corresponds to what you found for (a).

It is estimated that a 1 unit increase in log Acetic Acid Concentration is associated with a 15.648 increase in average Taste Score for cheddar cheese. This is consistent with the previous interpretation since this model confirms that there is a positive correlation between these two variables.

d

Add an appropriate line to the plot in (a) for the sample regression model.



e

Is there sufficient evidence to indicate that acetic acid has a linear relationship with taste? Perform a hypothesis test with $\alpha = 0.05$ to help answer this question.

```
##
## Call:
## lm(formula = taste ~ Acetic, data = cheese)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.642  -7.443   2.082   6.597  26.581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -61.499     24.846  -2.475  0.01964 *
## Acetic        15.648      4.496   3.481  0.00166 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.82 on 28 degrees of freedom
## Multiple R-squared:  0.302,  Adjusted R-squared:  0.2771
## F-statistic: 12.11 on 1 and 28 DF,  p-value: 0.001658
```

```
##                   2.5 %    97.5 %
## (Intercept) -112.394113 -10.60311
## Acetic         6.438594  24.85694
```

There is convincing evidence that there is a relationship between Acetic Acid Concentration and Taste Score (Two-tailed T-Test. p-value = 0.0017). With 95% confidence, there is between a 6.4386 and 24.8569 increase in average Taste Score associated with a 1 unit increase in log Acetic Acid Concentration.

f

Suppose the acetic acid content was listed on the package for each individual block of cheddar cheese available for purchase at a grocery store. Also, an estimate of taste is given in the form of both a confidence interval and a prediction interval using the previously found regression model. If you were going to purchase a block of cheddar cheese, which would you be more important to you – a confidence interval or prediction interval? Explain.

A confidence interval indicates the range of values in which the "true" mean lies with a percentage of confidence. A prediction interval indicates the range of values in which a single value will reside with a percentage of confidence. If I were cheese shopping, I would be interested in assessing brands and types of cheddar cheese in which a **mean** would be more helpful than a single value. Thus, I would put more stock in the confidence interval than the prediction interval.

g

Calculate the 95% confidence and prediction intervals when acetic acid is 4.5. Interpret the intervals.

```
predict(cheese.model1, newdata = data.frame(Acetic=c(4.5)), interval = "confidence")
```

```
##       fit       lwr      upr
## 1 8.91634 -1.628502 19.46118
```

```
predict(cheese.model1, newdata = data.frame(Acetic=c(4.5)), interval = "prediction")
```

```
##       fit       lwr      upr
## 1 8.91634 -21.29518 39.12786
```

It is estimated that the average taste score given a log Acetic Acid concentration of 4.5 is 8.9163. With 95% confidence, the mean taste score for a log acetic acid concentration of 4.5 is between -1.6285 and 19.4612. With 95% confidence, the taste score for a random observation where log acetic acid concentrion is 4.5 is between -21.2952 and 39.1279.

h

If good tasting cheese is the ultimate goal, what would be preferred for acetic acid levels in cheese?

Given the range of values in our data, the best tasting cheese would the cheese containing highest concentration of Acetic acid. This would be the cheese with a log Acetic Acic Concentration of 6.458.

2

Continue using taste as the response variable and acetic acid as the predictor variable.

a

Give the ANOVA table

```
anova(cheese.model1) %>% kable
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Acetic	1	2314.142	2314.1415	12.11424	0.0016582
Residuals	28	5348.745	191.0266	NA	NA

b

Using the relevant information from the ANOVA table, perform an F-test for

$$H_0 : \beta_1 = 0 H_A : \beta_1 \neq 0$$

Use $\alpha = 0.05$

There is convincing evidence that there is a relationship between Acetic Acid Concentration and Taste Score for Cheddar Cheese (Sum of Squares F-Test. p-value = 0.00166).

c

What is R^2 *for the estimated regression model? Fully interpret its value.*

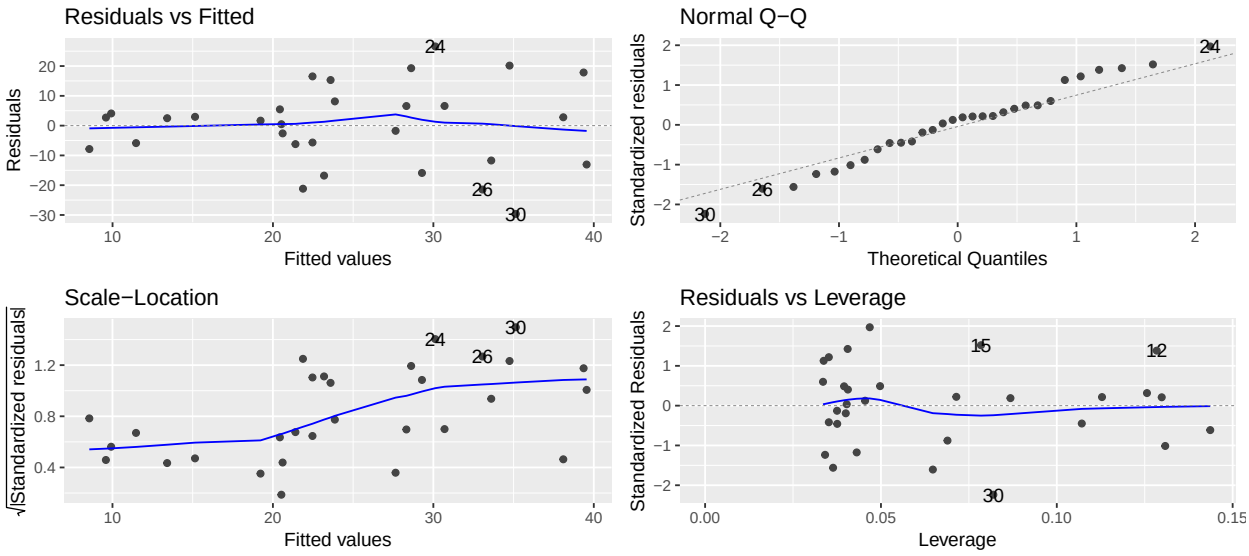
$R^2 = 0.302$

The Log Acetic Acid Concentration in Cheddar Cheese explains 30.2% of the variance in the mean taste score for cheddar cheese.

d

Comment on the following items with regards to the model: i. Linearity of the regression function ii. Constant error variance iii. Normality of ϵ_i *iv. Outliers Make sure to specifically refer to plots and numerical values in your comments.*

```
autoplot(cheese.model1)
```



Linearity of the Regression Function

There do not appear to be any noticable patterns for in the **Residuals vs Fitted** plot so the linearity assumption is likely a correct one.

Constant Error Variance

```
lmtest::bptest(cheese.model1, student = F)
```

```
##
##  Breusch-Pagan test
##
## data:  cheese.model1
## BP = 5.4974, df = 1, p-value = 0.01905
```

```
group <- cheese$Acetic <= median(cheese$Acetic)

levene.test(cheese.model1$residuals, group)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  cheese.model1$residuals
## Test Statistic = 4.5572, p-value = 0.04167
```

The residuals in the **Residuals vs Fitted** plot widen as the Taste Score increases so the constant error variance appears suspect. The Brown-Forsythe Test for Constant Variance indicates that there is moderate evidence that the variance is not constant (p-value = 0.0417). A follow-up Bruesch-Pagan test for Constant Variance also provides moderate evidence that the variance is not constant (p-value = 0.019). According to the results of the normality test below, it is safe to say that the variance within the data is likely not constant.

Normality of ϵ_i

```
shapiro.test(cheese.model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cheese.model1$residuals
## W = 0.98231, p-value = 0.883
```

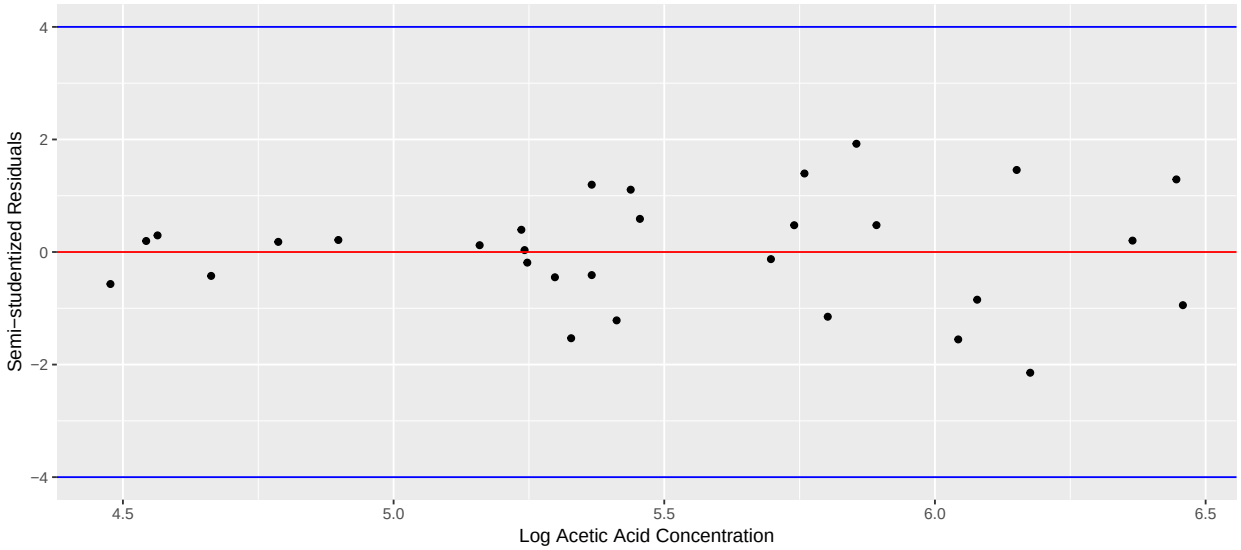
The Standardized Residuals are near the theoretical line but stray somewhat towards the tails. A quick Shapiro-Wilk test indicates that there is no evidence to suggest that data are non-normal (p-value = 0.833).

Outliers

```
MSE <- anova(cheese.model1)$`Mean Sq`[2]
e.star <- cheese.model1$residuals / sqrt(MSE)

ggplot(cheese.model1, aes(x = Acetic, y = e.star)) +
```

```
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  geom_hline(yintercept = c(-4,4), color = "blue") +
  xlab("Log Acetic Acid Concentration") +
  ylab("Semi-studentized Residuals")
```

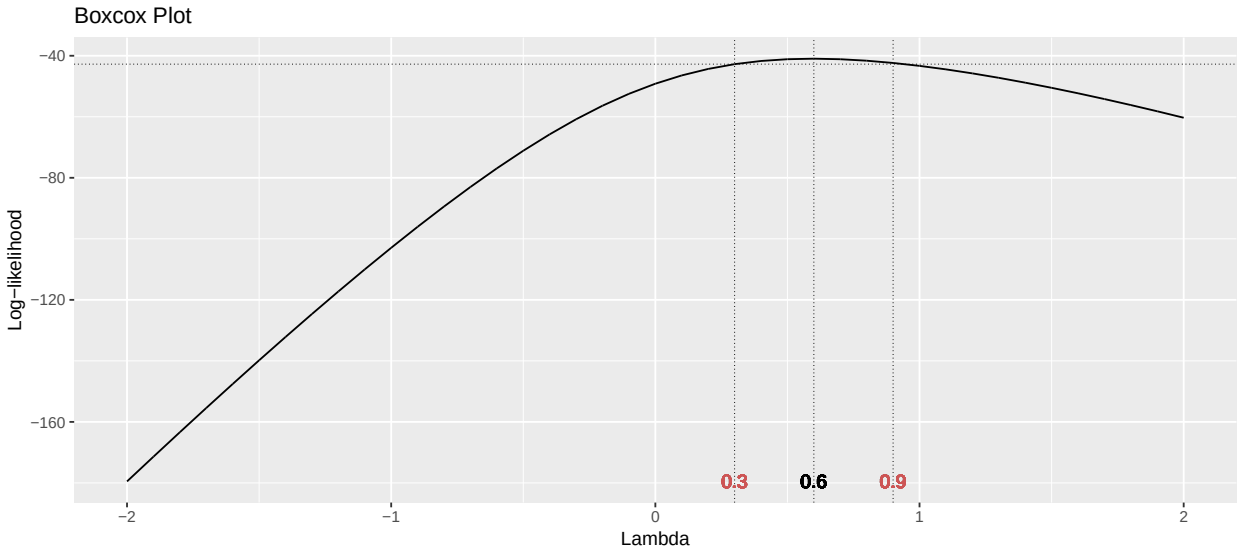


A semi-studentized residual plot indicates that there are residuals greater than abs(4) thus there are no outliers.

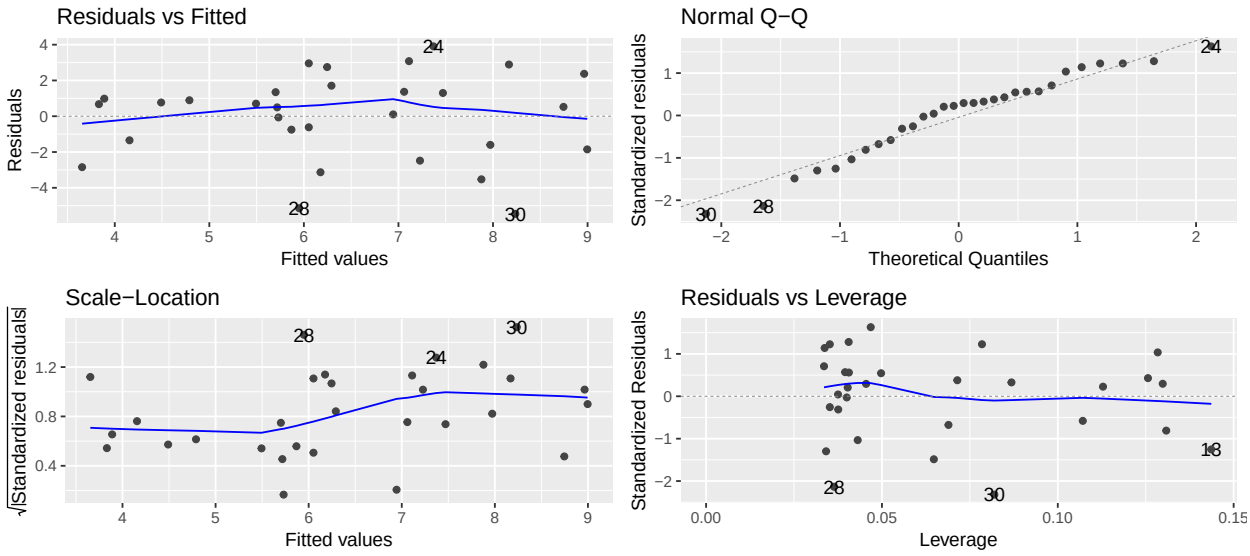
e

The residual plots and subsequent tests indicate that Constant Variance assumption has been violated. This can be attempted to be resolved with a transformation on Taste Score. In order to find the optimal transformation on Y, a Box-Cox plot can be employed to find the optimal transformation to obtain the smallest SSE.

```
lindia::gg_boxcox(cheese.model1)
```



```
cheese.model2 <- lm(taste^0.6 ~ Acetic, data = cheese)
autoplot(cheese.model2)
```



The Box-Cox plot indicates that $\lambda = 0.6$ is the optimal value for a transformation to Y. The transformed model per these results is $\hat{Y_i}^{0.6} = -8.4187 + 2.6967$ Acetic.

```
levene.test(cheese.model2$residuals, group)
```

```
##
##  Modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  cheese.model2$residuals
## Test Statistic = 1.1066, p-value = 0.3018
```

```
lmtest::bptest(cheese.model2, student = F)
```

```
##
##  Breusch-Pagan test
##
## data:  cheese.model2
## BP = 1.9033, df = 1, p-value = 0.1677
```

The **Residuals vs Fitted** plot indicates that the variance is more constant than before. The Brown-Forsyte test indicates that there is no evidence suggesting the variance is not constant(p-value = 0.3018). Additionally, the Bruesch-Pagan tests indicates the same result (p-value = 0.1677). Normality and Linearity assumptions still appear to be met with the new model.

f

Using the new model that was estimated for part (e), find the 95% confidence interval for mean taste when acetic acid has a value of 4.5. Compare the interval you found in Problem 1(g). Which interval is more likely to have 95% confidence? Explain.

```
predict(cheese.model2, newdata = data.frame(Acetic=c(4.5)), interval = "confidence")
```

```
##        fit     lwr      upr
```

```
## 1 3.716514 1.84454 5.588488
```

With 95% confidence, the average taste score when Log Acetic Acid Concentration is 4.5 is between -1.6468 and 9.08. Compared to the confidence interval for the non-transformed model which is [-1.6285, 19.4612]. The confidence interval is approximately halved in the transformed model when compared to the original model; however, this makes sense as the transformed model scales the response. The transformed model has a more likely to have confidence since the transformed model meets all the required assumptions whereas the original model does not.

g

Are there any other problems remaining with the model after what was done in part (e)? Justify your answer. Note that you do not need to actually implement any changes to the model.

The R^2 value decreased to 0.2896 for the transformed model. This value is low so additional predictors should be tested for significance to achieve a higher R^2 value and a better fit overall.