Jonothan Meyer
Linear Regression
02/08/21
Project 2

Exercise 1

Data Set: divusa
Summary: This data set looks at the divorce rate in the United States from 1920-1996, looking at other factors such as female labor force, birth rate, and marriage rate. The goal was to distill the sample set down to plots and numbers that gave a good idea what the data looks like in bite sized format.

head(divusa):

```
> head(divusa)
  year divorce unemployed femlab marriage birth military
1 1920     8.0        5.2  22.70     92.0 117.9   3.2247
2 1921     7.2       11.7  22.79     83.0 119.8   3.5614
3 1922     6.6        6.7  22.88     79.7 111.2   2.4553
4 1923     7.1        2.4  22.97     85.2 110.5   2.2065
5 1924     7.2        5.0  23.06     80.3 110.9   2.2889
6 1925     7.2        3.2  23.15     79.2 106.6   2.1735
```

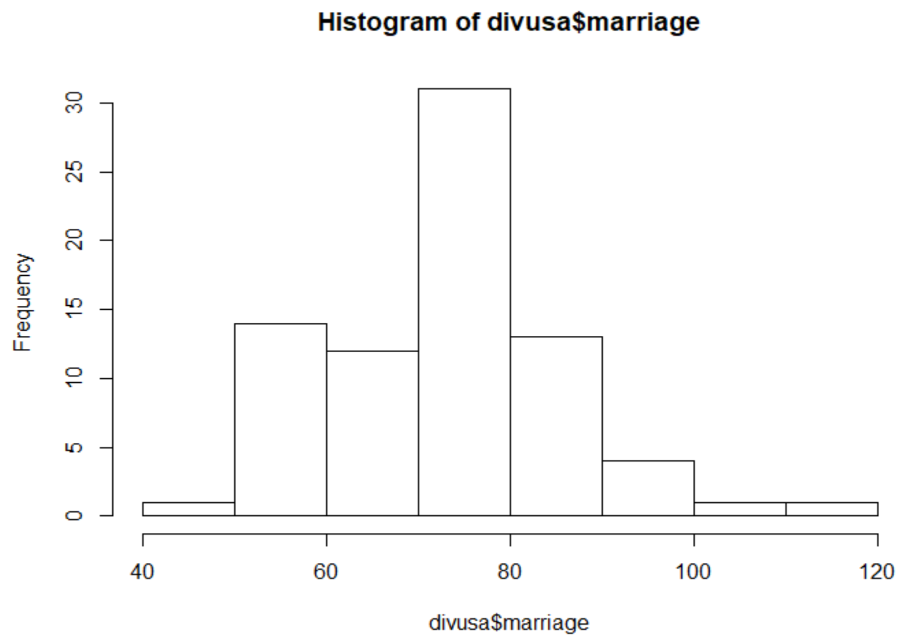
*This gives a basic numerical overview of the first 6 indices of data, giving a general idea of the factors looked at, as well as a few general numerical values for the data points to get an idea of the data and it's distribution.

summary(divusa):

```
> summary(divusa)
      year         divorce        unemployed         femlab         marriage          birth           military     
 Min.   :1920   Min.   : 6.10   Min.   : 1.200   Min.   :22.70   Min.   : 49.70   Min.   : 65.30   Min.   : 1.940  
 1st Qu.:1939   1st Qu.: 8.70   1st Qu.: 4.200   1st Qu.:27.47   1st Qu.: 61.90   1st Qu.: 68.90   1st Qu.: 3.469  
 Median :1958   Median :10.60   Median : 5.600   Median :37.10   Median : 74.10   Median : 85.90   Median : 9.102  
 Mean   :1958   Mean   :13.27   Mean   : 7.173   Mean   :38.58   Mean   : 72.97   Mean   : 88.89   Mean   :12.365  
 3rd Qu.:1977   3rd Qu.:20.30   3rd Qu.: 7.500   3rd Qu.:47.80   3rd Qu.: 80.00   3rd Qu.:107.30   3rd Qu.:14.266  
 Max.   :1996   Max.   :22.80   Max.   :24.900   Max.   :59.30   Max.   :118.10   Max.   :122.90   Max.   :86.641  
```

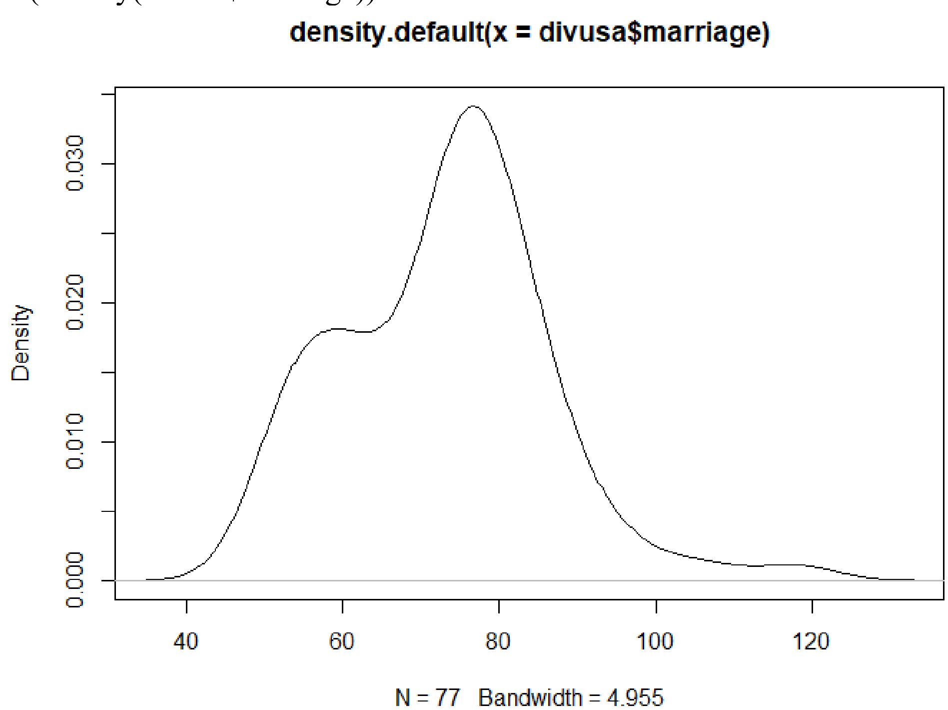
*The summary() function operates differently depending on the type of data set. For a table of information it gives information regarding the min/max, quartiles, mean, and median. Differs from the head function in that it doesn't give info from specific data points, but info of the data overall.

hist(divusa$marriage):



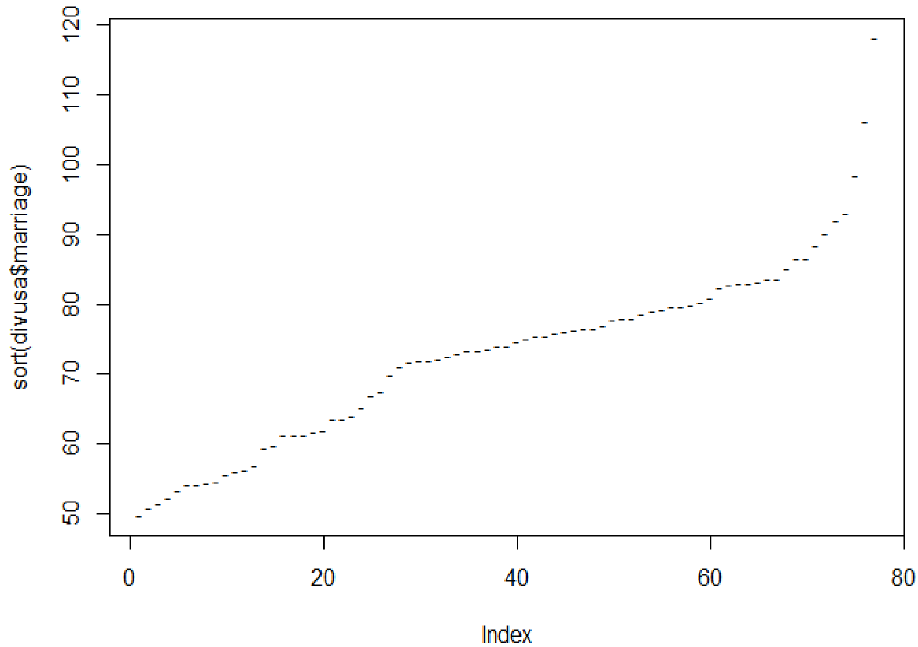
*I wanted to look at the distribution of marriage in the data, and first did so by plotting a histogram. This clearly shows that marriage largely normally distributed, with the mean between 70-80 marriages per 1000 women

plot(density(divusa$marriage)):



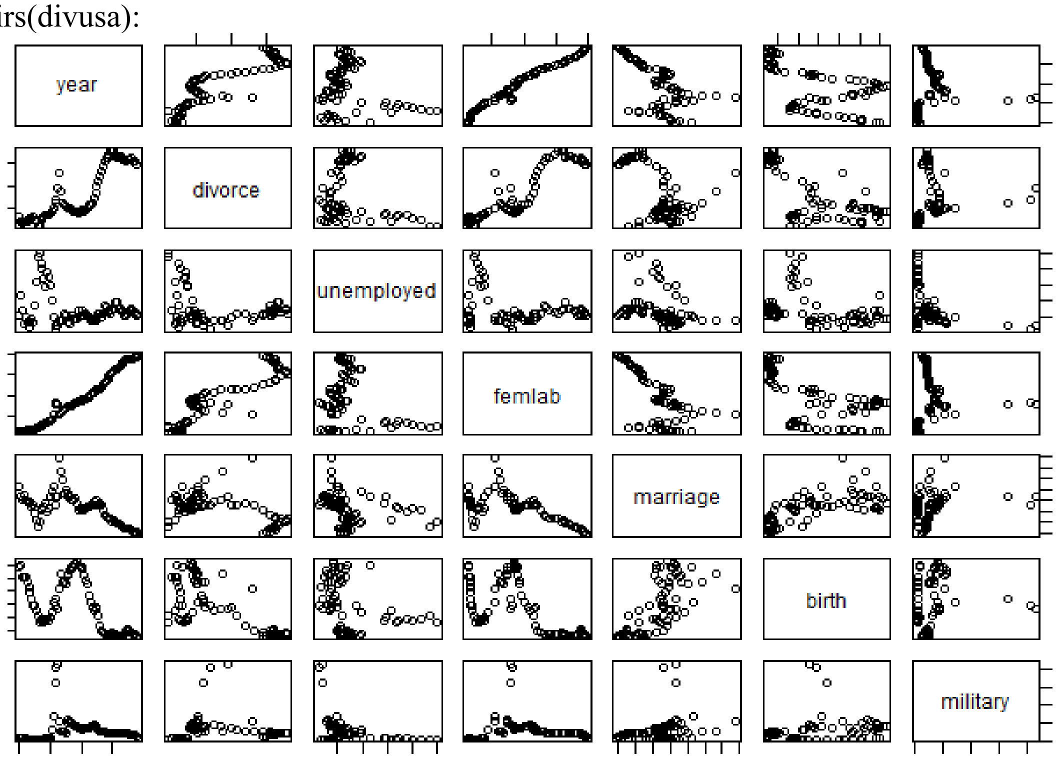
*Kernal Density Estimate of the same data seen in the previous histogram. Since this graph is continuous it smooths over the blockiness of the histogram. At a glance this plot is able to give a slightly better view of the marriage data distribution.

plot(sort(divusa$marriage),pch="-"):



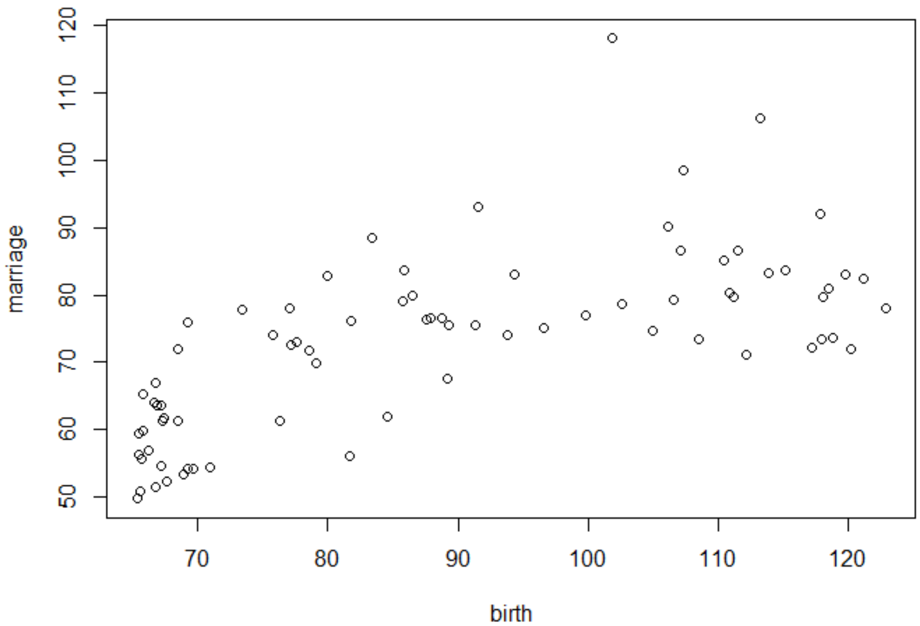
*Index Plot has the advantage of showing all the cases individually and is useful for finding outliers.

pairs(divusa):



*This has to be my favorite and one of the most useful quick plots of any data set. It shows all of the different variables plotted against each other. This could be very useful for quickly finding linear interactions. In terms of a predictor and response that are linearly regressive birth~marriage appears to be, along with femlab~year.

plot(marriage~birth,divusa):



*After looking at the pairs(divusa) function I decided to make a linear regression model using birth (per 1000 women) as my predictor variable, and marriage (per 1000 women) as my response. Does appear to have a vaguely linear relationship.

summary(marriage.birth.lm):

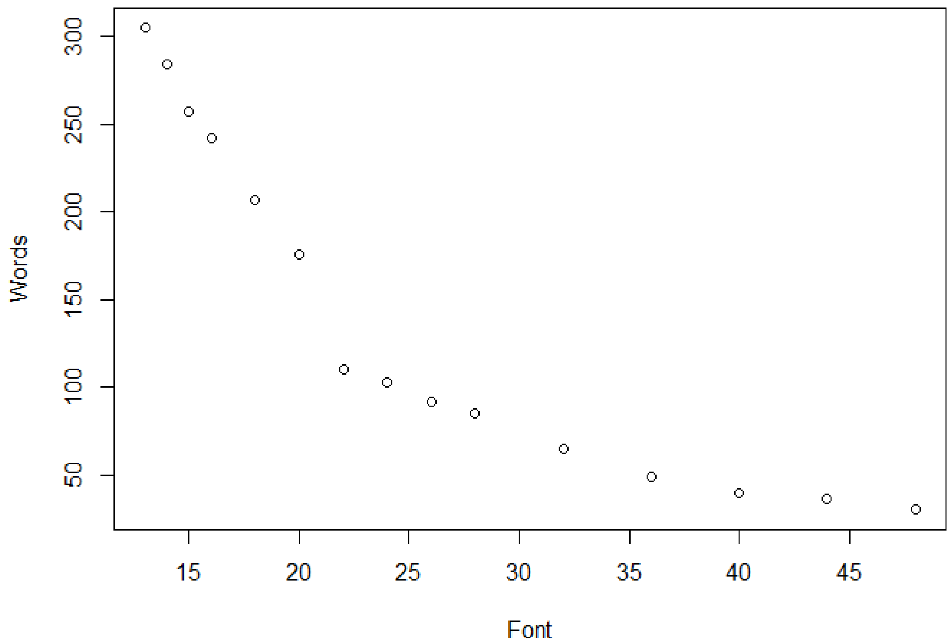
```
Residuals:
    Min      1Q  Median      3Q     Max 
-15.157  -6.406  -1.103   5.141  39.233 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 32.70558    5.21995   6.266 2.13e-08 ***
birth        0.45301    0.05738   7.896 1.89e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.763 on 75 degrees of freedom
Multiple R-squared:  0.4539,    Adjusted R-squared:  0.4466 
F-statistic: 62.34 on 1 and 75 DF,  p-value: 1.885e-11

> marriage.birth.lm

Call:
lm(formula = marriage ~ birth, data = divusa)

Coefficients:
(Intercept)        birth  
     32.706        0.453  
```

*Info that gives insight into the relationship between the predictor and response variables in question. r-squared=.45, b1=32.706, and b0=.453 with 75 degrees of freedom.

Exercise 2

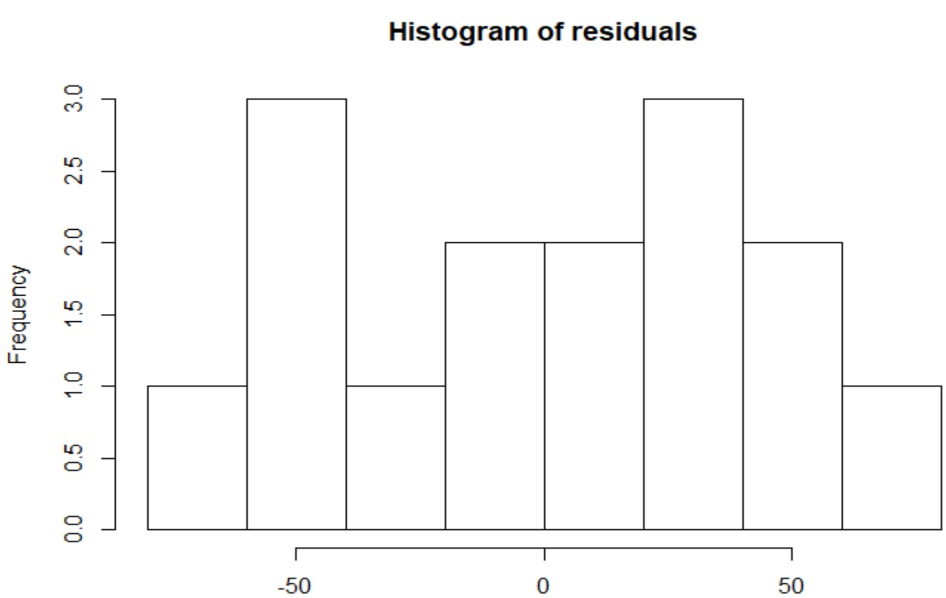
(1) This is a controlled experiment because I, as the experimenter, can "control" the values of the predictor variable by setting the font. After which I record the corresponding response value by counting the amount of words on the page of the document.

(2) Scatterplot



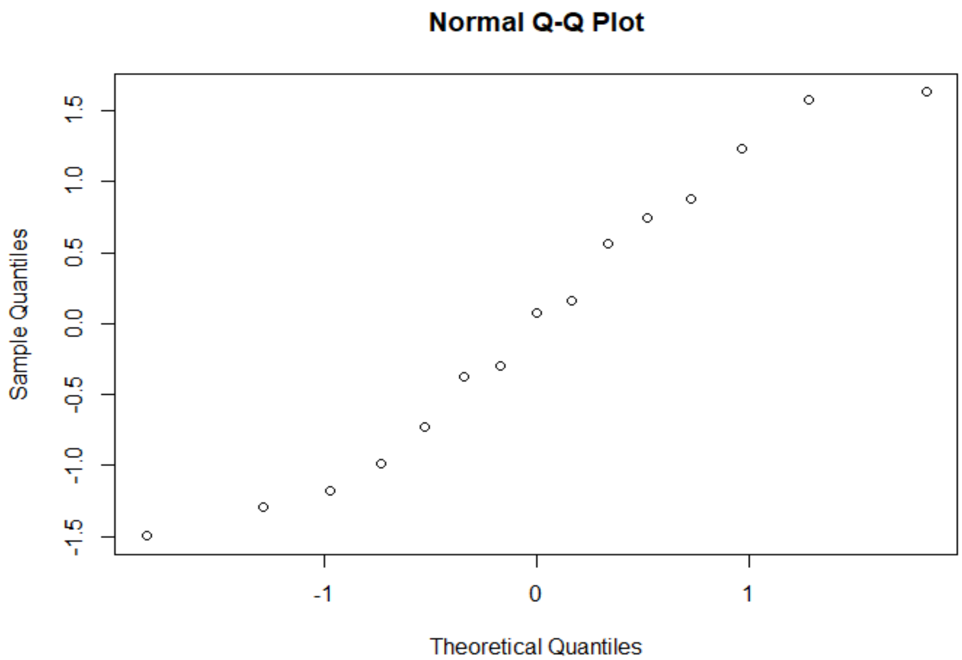
*Using the "eye test" the data does not appear to be exactly linear. While you can clearly see a relationship between the predictor and response it is not a straight line, and appears more logarithmic.

(3) Normal Probability Plot

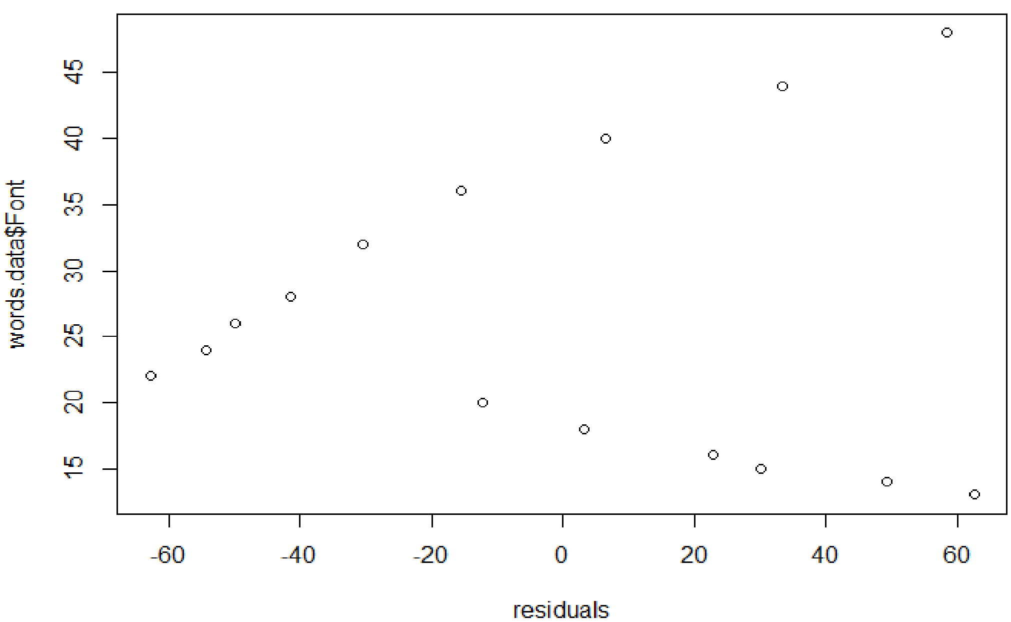


*The residuals do not appear to be approximately normally distributed for the error term in the model to be satisfied.

(4) QQ Plot



Residuals Plotted Against Predictor



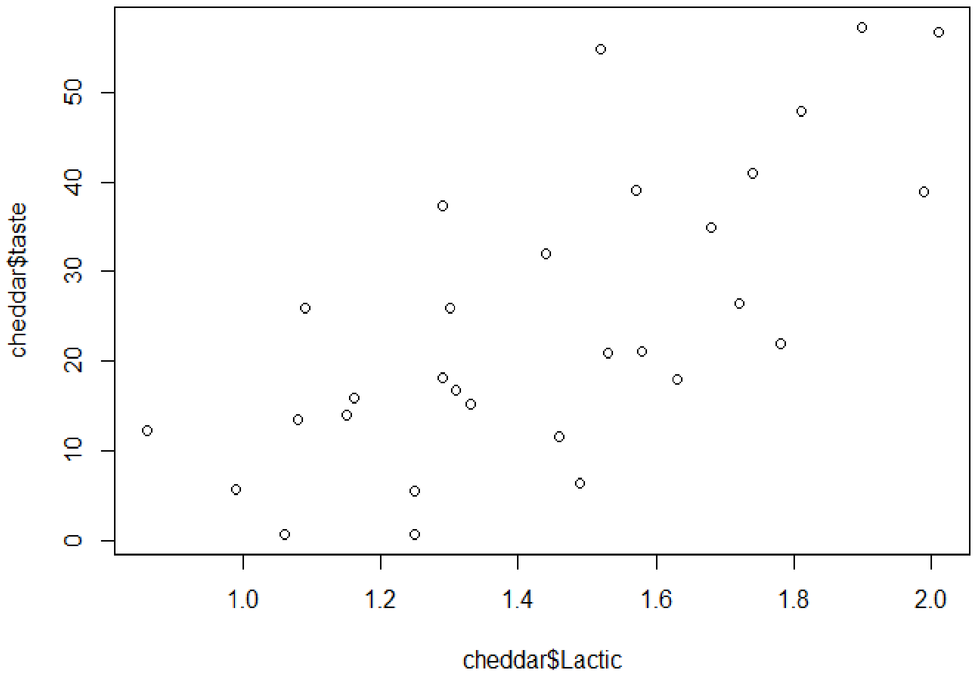
*It does not appear that the normality assumption for the error term is satisfied because the data points appear to be linear when they should appear random. The normal QQ Plot should appear as a straight line of data points if the current linear regression model is appropriate, and it appears to be a bit off from that.

(5)
b0 = 342.998
b1 = -7.737 y = -7.37x + 342.998
(6)
r^2 = .813, 81.3% of the data is represented by the linear regression model.

(7)
I do not think a Linear Regression model is appropriate for this data because it's slope appears logarithmic, not linear. If the y-axis was changed to be logarithmic then I think this model would work quite well.

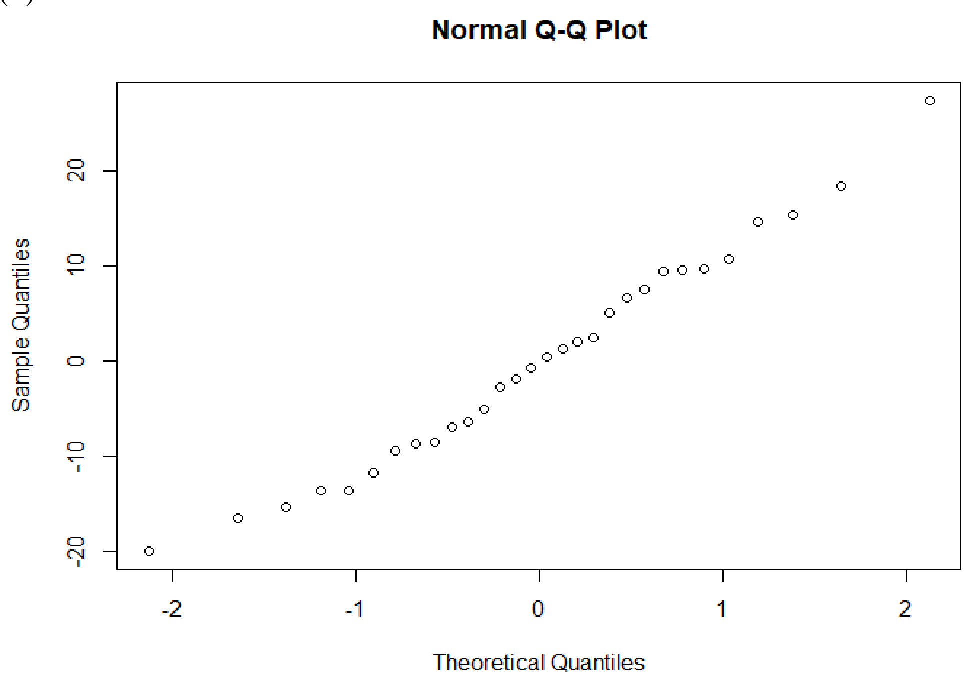
Exercise 3

(1) Lactic Acid Plotted Against Taste



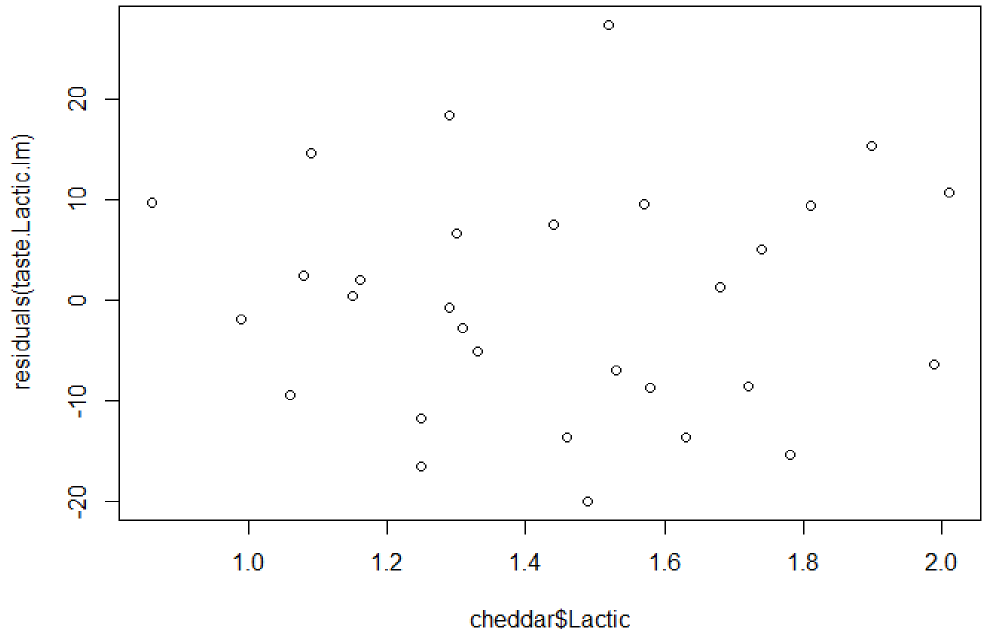
*It does appear that the data passes the "eye test" and satisfies the criteria for a linear regression.

(2)



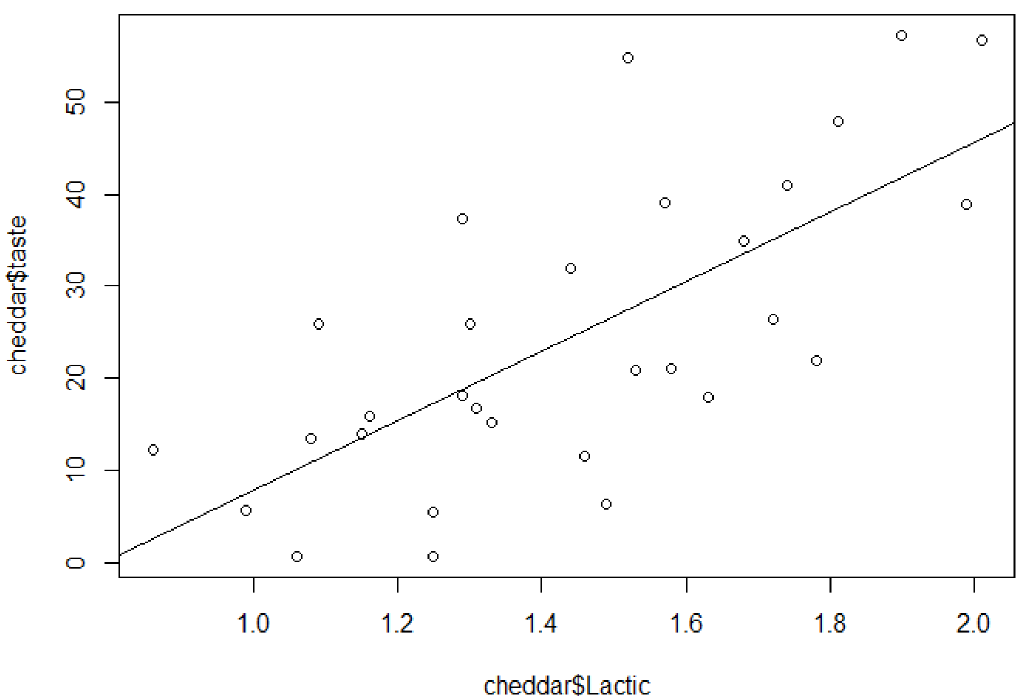
*The data appears as basically a straight line, so it does appear that the normaility assumption for the error term in the model is satisfied.

(3) Residual Values Associated with the Linear Regression Plotted as Function of Predictor Variable



*The plot appears to be completely random, so it does appear that constant variance assumption associated with the model is satisfied.

(4)
b0 = -29.86   y = 37.72x - 29.86
b1 = 37.72



*After calculating the intercept and plotting the regression line there seems to be a problem. The intercept is negative despite the intercept appearing positive when plotted. I was unable to find the problem with this after checking my work.

(5)
sxy = 100.753
sxx = 2.67
(sxy/sxx) = 37.71995 = b1

Residual Standard Error (found in R): 11.75, 11.75^2 = 138.06
Computation:
sse = 3862.489
n = 30, df = 28
3862.489/28 = 137.946,
sqr(137.946) = 11.75 = sigma hat squared found in R

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -29.859     10.582  -2.822  0.00869 ** 
cheddar$Lactic   37.720      7.186   5.249 1.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 28 degrees of freedom
Multiple R-squared:  0.4959,	Adjusted R-squared:  0.4779 
F-statistic: 27.55 on 1 and 28 DF,  p-value: 1.405e-05
```

(6)
r^2 = .4959

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -29.859     10.582  -2.822  0.00869 ** 
cheddar$Lactic   37.720      7.186   5.249 1.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.75 on 28 degrees of freedom
Multiple R-squared:  0.4959,	Adjusted R-squared:  0.4779 
F-statistic: 27.55 on 1 and 28 DF,  p-value: 1.405e-05
```

(7)
The estimated slope of the regression line in terms of the real world response is 37.72. For every one unit increase in Lactic Acid it is expected that taste will increase by 37.72 units.

(8)
x = 1.5
y = (1.5)(37.75) - 29.859
y = 26.72 expected units of taste if Lactic Acid is at 1.5

(9)
I do not think a simple linear regression model is appropriate for the underlying bivariate population in this problem. One issue is the negative intercept found when we know there is an asymptote at y = 0.

Also, the r^2 seems to be fairly low at .50, so a simple linear regression is not the best way to model this data for prediction.