STAT 8030 HW 3

Question: 2

1. We're fitting the model using "Model <- lm(y~x1+x2+x3+x4,data = fitness)". Where we read the "Fitness_walking.cvs" file from the disk first as shown below.

```
> ##Question 2
> fitness <- read.csv(file = "C:/Users/sshri/Desktop/Fall 2021/Stat 8030/fitness_walking.csv",header=TRUE)
> head(fitness)
  Subject   y    x1   x2   x3    x4
1       1 1.5 139.8 19.1 18.1 133.6
2       2 2.1 143.3 21.1 15.3 144.6
3       3 1.8 154.2 21.2 15.3 164.6
4       4 2.2 176.6 23.2 17.7 139.4
5       5 2.2 154.3 22.4 17.1 127.3
6       6 2.0 185.4 22.1 16.4 137.3
>
> #2.1
> Model <- lm(y~x1+x2+x3+x4,data = fitness)
>
```

   Now, we're getting the R square value as shown below, in our case the R square value is 0.5815464

```
>
> Value <- summary(Model)$r.squared
> Value
[1] 0.5815464
>
```

   From the R square value, we can interpret how well the regression model fits the observed data. In our case we can say that 58.15% of the data fits the regression model perfectly.

2. R square found for the quadratic model that is fitted to the required data is 0.5925011 as shown below:

```
> #2.2
> x1_S <- I(fitness$x1^2)
> x2_S <- I(fitness$x2^2)
> x3_S <- I(fitness$x3^2)
> x4_S <- I(fitness$x4^2)
>
> Quad_Model  <- lm(y~x1+x2+x3+x4+x1_S+x2_S+x3_S+x4_S,data=fitness)
> Quad_Model <- summary(Quad_Model)
>
> Quad_Model$r.squared
[1] 0.5925011
>
```

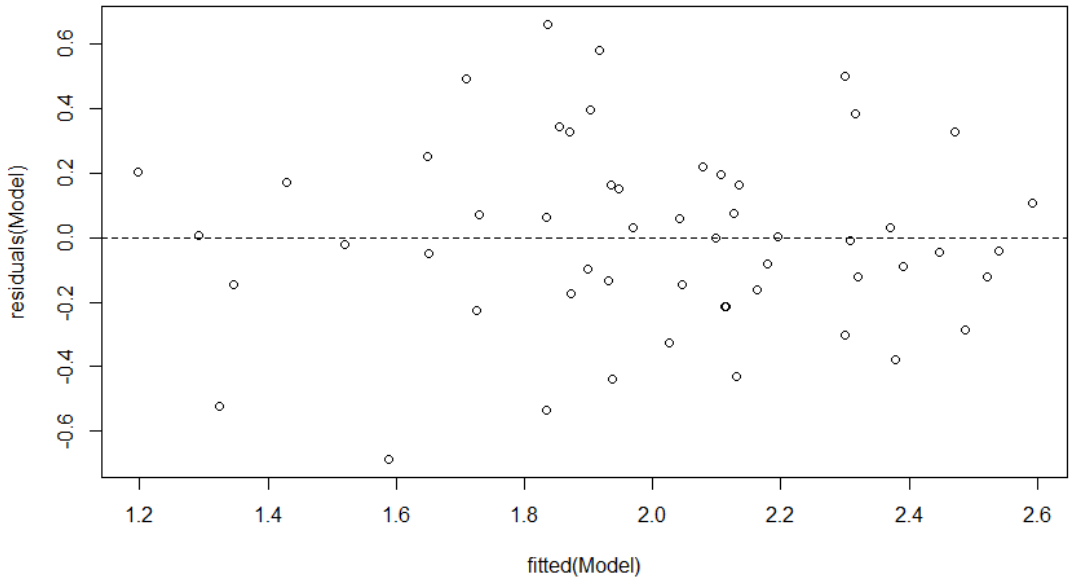
   The difference between the two R squares observed is 0.0109547. i.e. R square obtained here has an increase of 1.09%

3. For Hypothesis of Quadratic terms, we first see the Null Hypothesis. i.e.
   Null Hypothesis => $\beta 5 = \beta 6 = \beta 7 = \beta 8 = 0$
   Thus, the alternate hypothesis is that at least one of the β is non- zero. Now, we find the test statistic and p-value as shown below:

```
> #2.3
> Quad_Model  <- lm(y~x1+x2+x3+x4+x1_S+x2_S+x3_S+x4_S,data=fitness)
> anova(Model, Quad_Model)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4
Model 2: y ~ x1 + x2 + x3 + x4 + x1_S + x2_S + x3_S + x4_S
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 4.3938
2     45 4.2787  4   0.11502 0.3024 0.8748
> |
```

We can see that F-statistic = 0.3024 and P-value = 0.8748
We also observe that the P-value we got is greater than the $\alpha = 0.05$, which suggests that we fail to reject the NULL hypothesis. Thus, we do not have any evidence to support that adding quadratic terms significantly improves the model.

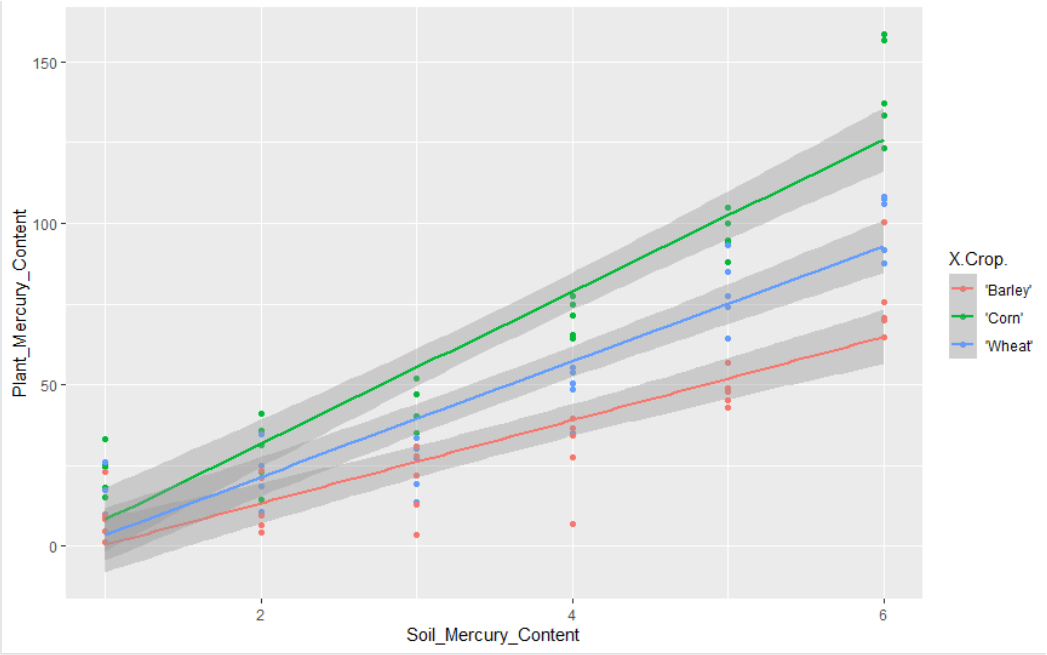
4. Since we already figured that the second model i.e., quadratic model doesn't prove to be improving the initial model. I would be choosing the initial model to make the Residual Plot. The Residual Plot is shown below:



I feel from the residual plot observed, the assumptions of equal variances, normal errors, and linear mean function appear to be reasonable because there is no pattern in the plot, nor the residuals are bounded close to the 0.0 line, suggesting that there is constant variance.

Question: 3

1. Plot showing SoilMerCon and PlantMerCon is shown below:



2. Though a few points have a different pattern, the relationship between soil mercury content and plant mercury content appears to be linear. Since most of the points fall in the line.

   And the relationship between soil mercury content and plant mercury content appears to be inface the same for all three crops.

3. The fitted model ignoring the crop type is shown below:

```
> #3.3
> Model <- lm(X.PlantMerCon.~X.SoilMerCon.,data = mercury)
> summary(Model)

Call:
lm(formula = X.PlantMerCon. ~ X.SoilMerCon., data = mercury)

Residuals:
    Min      1Q  Median      3Q     Max
-51.514 -12.984   0.596  13.010  64.118

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -14.022      4.986  -2.812  0.00607 **
X.SoilMerCon.  18.084      1.280  14.124  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.74 on 88 degrees of freedom
Multiple R-squared:  0.6939,    Adjusted R-squared:  0.6904
F-statistic: 199.5 on 1 and 88 DF,  p-value: < 2.2e-16

> |
```

4. The single fitted model as required is shown below:

```
> #3.4
> Interaction <-lm(X.PlantMerCon.~X.SoilMerCon.* X.Crop., data = mercury)
> summary(Interaction)

Call:
lm(formula = X.PlantMerCon. ~ X.SoilMerCon. * X.Crop., data = mercury)

Residuals:
    Min      1Q  Median      3Q     Max
-32.178  -8.305  -1.203   7.772  35.569

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -12.528      5.550  -2.257   0.0266 *
X.SoilMerCon.                   12.877      1.425   9.036 5.02e-14 ***
X.Crop.'Corn'                   -2.832      7.848  -0.361   0.7191
X.Crop.'Wheat'                  -1.649      7.848  -0.210   0.8341
X.SoilMerCon.:X.Crop.'Corn'     10.678      2.015   5.298 9.24e-07 ***
X.SoilMerCon.:X.Crop.'Wheat'     4.945      2.015   2.454   0.0162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.33 on 84 degrees of freedom
Multiple R-squared:  0.8794,    Adjusted R-squared:  0.8722
F-statistic: 122.5 on 5 and 84 DF,  p-value: < 2.2e-16

> |
```

We can see from the above summary that, for the whole model the intercept = -12.528 and the Model slope = 12.877

From this we draw,
The intercept of Barley = -12.528 and slop of Barley = 12.877
The intercept of corn = -2.832 + (-12.528) = - 15.36 and slop of corn = 10.678 + 12.877 = 23.555
The intercept of wheat = -1.649 + (-12.528) = -14.177 and slop of wheat = 4.945 + 12.877 = 17.822

5. No, from the above observation, it does not appear that the intercepts are significantly different across crop types.

Inferential analysis for the same is as follows:

Null Hypothesis => β3 = 0; Where, β3 is X.crops.'Wheat'
Alternate Hypothesis => β3 is not equal to 0

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -12.528      5.550  -2.257   0.0266 *
X.SoilMerCon.                   12.877      1.425   9.036 5.02e-14 ***
X.Crop.'Corn'                   -2.832      7.848  -0.361   0.7191
X.Crop.'Wheat'                  -1.649      7.848  -0.210   0.8341
```

From the summary, we can see, t-statistic = -0.210 and p-value = 0.8341

We also observe that the P-value we got is greater than the α = 0.05, which suggests that we fail to reject the NULL hypothesis. Thus, we do not have any evidence to support that the intercept of crop 'wheat' is different from others.

6. Yes, from the above observation, it appears that the slopes are significantly different across crop types.

   Inferential analysis for the same is as follows:

   Null Hypothesis => β3 = 0; Where, β3 is X.crops.'Wheat'
   Alternate Hypothesis => β3 is not equal to 0

```
X.SoilMerCon.:X.Crop.'Corn'    10.678      2.015   5.298 9.24e-07 ***
X.SoilMerCon.:X.Crop.'Wheat'    4.945      2.015   2.454   0.0162 *
---
```

   From the summary, we can see, t-statistic = 2.454 and p-value = 0.0162

   We also observe that the P-value we got is less than the α = 0.05, which suggests that we reject the NULL hypothesis. Thus, we have the evidence to support that the intercept of crop 'wheat' is different from all other crop slopes.

7. The single fitted model as required is shown below:

```
> #3.7
> Quad_model <- lm(X.PlantMerCon.~X.SoilMerCon.+I(X.SoilMerCon.^2)+X.SoilMerCon.*X.Crop., data = mercury)
> summary(Quad_model)

Call:
lm(formula = X.PlantMerCon. ~ X.SoilMerCon. + I(X.SoilMerCon.^2) +
    X.SoilMerCon. * X.Crop., data = mercury)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5799  -6.3137  -0.5595   5.7779  23.5706

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 21.0664     5.5129   3.821 0.000256 ***
X.SoilMerCon.              -12.3193     3.0236  -4.074 0.000105 ***
I(X.SoilMerCon.^2)           3.5994     0.4062   8.861 1.24e-13 ***
X.Crop.'Corn'               -2.8320     5.6600  -0.500 0.618149
X.Crop.'Wheat'              -1.6493     5.6600  -0.291 0.771470
X.SoilMerCon.:X.Crop.'Corn'  10.6777     1.4533   7.347 1.28e-10 ***
X.SoilMerCon.:X.Crop.'Wheat'  4.9446     1.4533   3.402 0.001031 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.613 on 83 degrees of freedom
Multiple R-squared:  0.938,     Adjusted R-squared:  0.9335
F-statistic: 209.3 on 6 and 83 DF,  p-value: < 2.2e-16

>
```

   From the summary,

   a. Yes, there appears to be a difference in slopes for the three crops since the slopes are different for all the crops.

b. No, there appears to be no difference in the intercepts for the three crops.

8. We can see the F test for the subsets below:

```
> #3.8
> anova(Interaction,Quad_model)
Analysis of Variance Table

Model 1: X.PlantMerCon. ~ X.SoilMerCon. * X.Crop.
Model 2: X.PlantMerCon. ~ X.SoilMerCon. + I(X.SoilMerCon.^2) + X.SoilMerCon. *
    X.Crop.
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     84 14925
2     83  7670  1    7255.2 78.511 1.241e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

We perform the hypothesis testing. We see from above that F-statistic = 78.511 and p-value = 1.241e-13

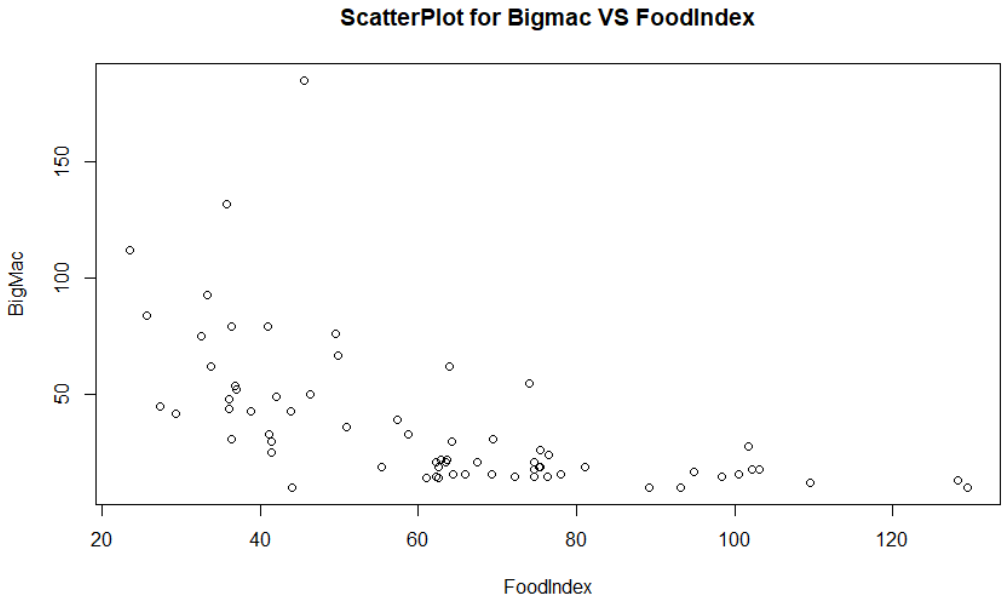
We also observe that the P-value we got is less than the α = 0.05, which suggests that we reject the NULL hypothesis. Thus, we have the evidence to support that adding quadratic terms significantly improved the model.

Question: 4

1. The model data is as follows:

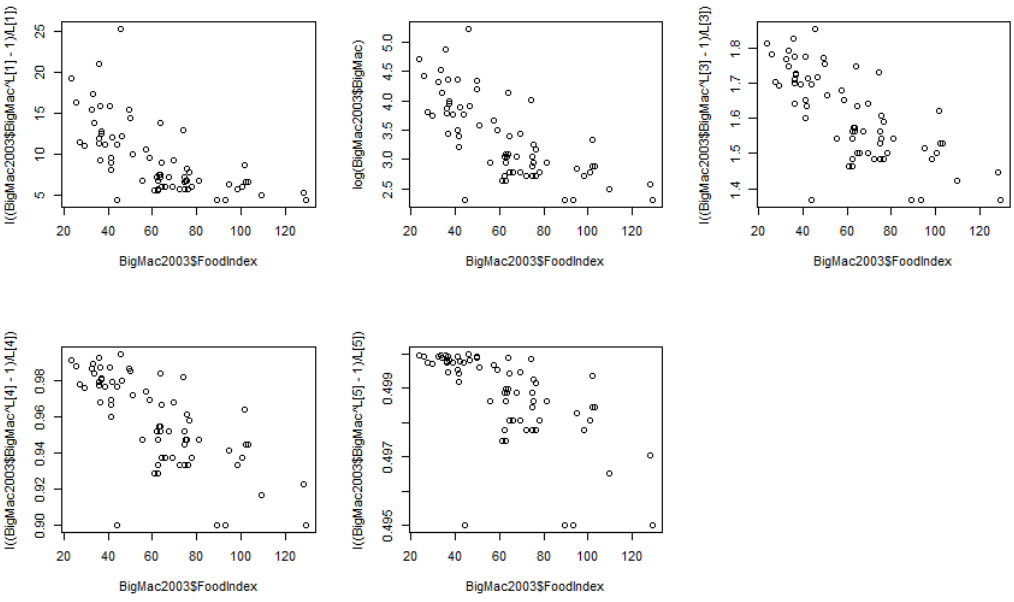
```
> ##Question 4
>
> #4.1
> library(alr4)
>
> BigMacData <- lm(BigMac~FoodIndex, data = BigMac2003)
> summary(BigMacData)

Call:
lm(formula = BigMac ~ FoodIndex, data = BigMac2003)

Residuals:
    Min      1Q  Median      3Q     Max
-40.442 -15.123  -6.898   9.729 135.733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.7520     8.5064   9.728 1.93e-14 ***
FoodIndex    -0.7343     0.1278  -5.746 2.44e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.91 on 67 degrees of freedom
Multiple R-squared:  0.3301,    Adjusted R-squared:  0.3201
F-statistic: 33.02 on 1 and 67 DF,  p-value: 2.435e-07

>
```

Scatter Plot for it is as follows:



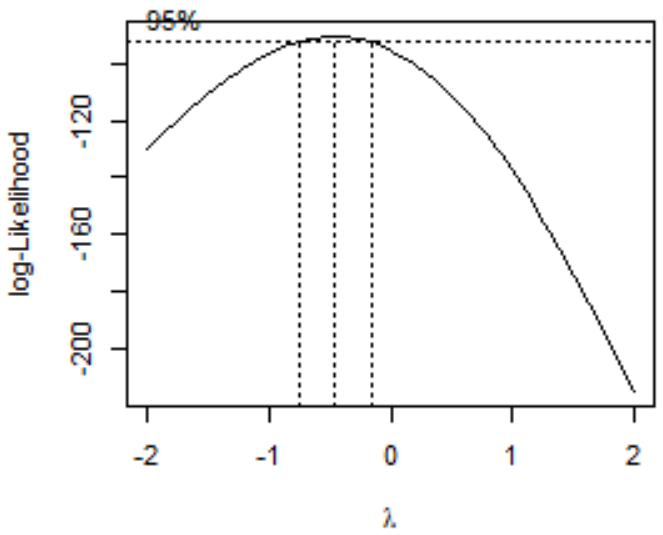
From above, we observe that there is no linear relationship between BigMac and FoodIndex since there is no pattern observed.

2. Transformation Plots is shown below:



We observe from the above plots that, the plot with Lambda = -1/2 appears to look most linear.

3. Plot drawn for log-likelihood vs Lambda is shown below:



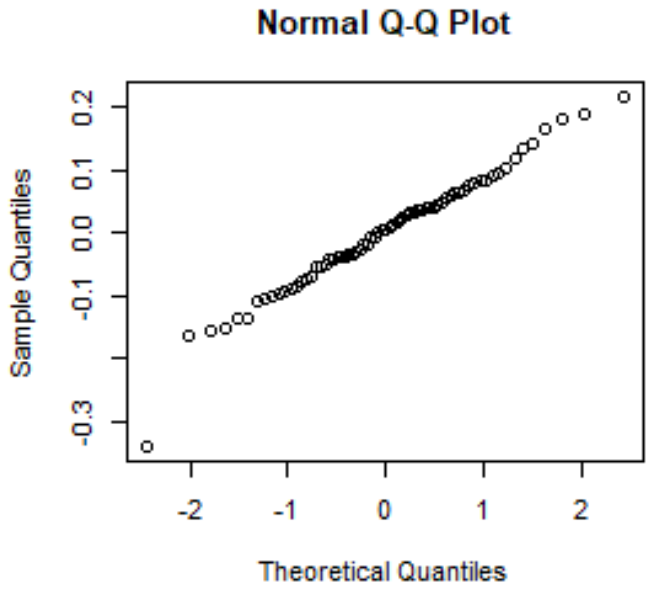
The optimal λ value as required can be found in the "lambda" given in the code. It is -0.464646 in our case as shown below:

Values	
L	num [1:5] 0.5 0 -0.5 -1 -2
Lambda	-0.464646464646465

4. Estimated regression coefficients under the transformed model can be seen below:

```
> #4.4
>
> LModel <- lm(I((BigMac2003$BigMac^Lambda - 1)/Lambda)~FoodIndex, data = BigMac2003)
> summary(LModel)

Call:
lm(formula = I((BigMac2003$BigMac^Lambda - 1)/Lambda) ~ FoodIndex,
    data = BigMac2003)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33972 -0.05379  0.00462  0.05903  0.21491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.9361874  0.0313815  61.698  < 2e-16 ***
FoodIndex   -0.0041498  0.0004714  -8.802 8.65e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09558 on 67 degrees of freedom
Multiple R-squared:  0.5363,    Adjusted R-squared:  0.5294
F-statistic: 77.48 on 1 and 67 DF,  p-value: 8.649e-13

> |
```

Normal quantile plot of the residuals is shown below:



From the above plot, we can clearly say that they look normal.