A3

Jefferson Li, Arib Shaikh

09/03/2021

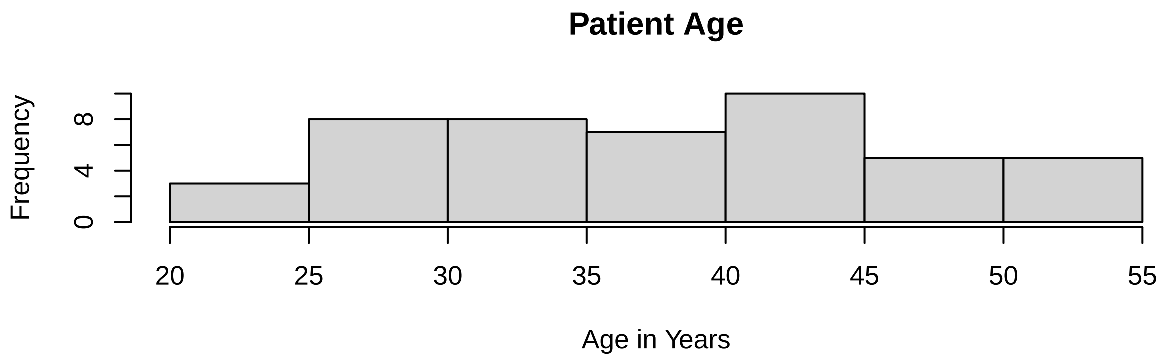
Q1

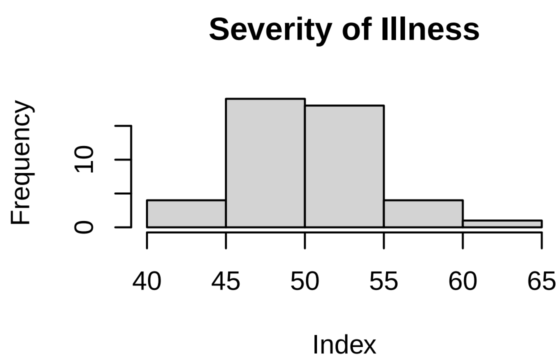
a)

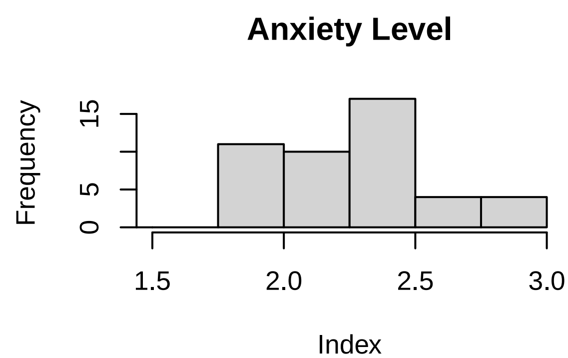
```
Data = read.table("PatientSatisfaction.txt",  col.names=c("Satisfaction", "Age", "Severity", "Anxiety"))

layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(Data$Age,
  main="Patient Age",
  xlab="Age in Years",
  xlim=c(20,55),
)
hist(Data$Severity,
  main="Severity of Illness",
  xlab="Index",
  xlim=c(40,65),
)
hist(Data$Anxiety,
  main="Anxiety Level",
  xlab="Index",
  xlim=c(1.5,3),
  breaks=c(1,1.25,1.5,1.75,2,2.25,2.5,2.75,3)

)
```



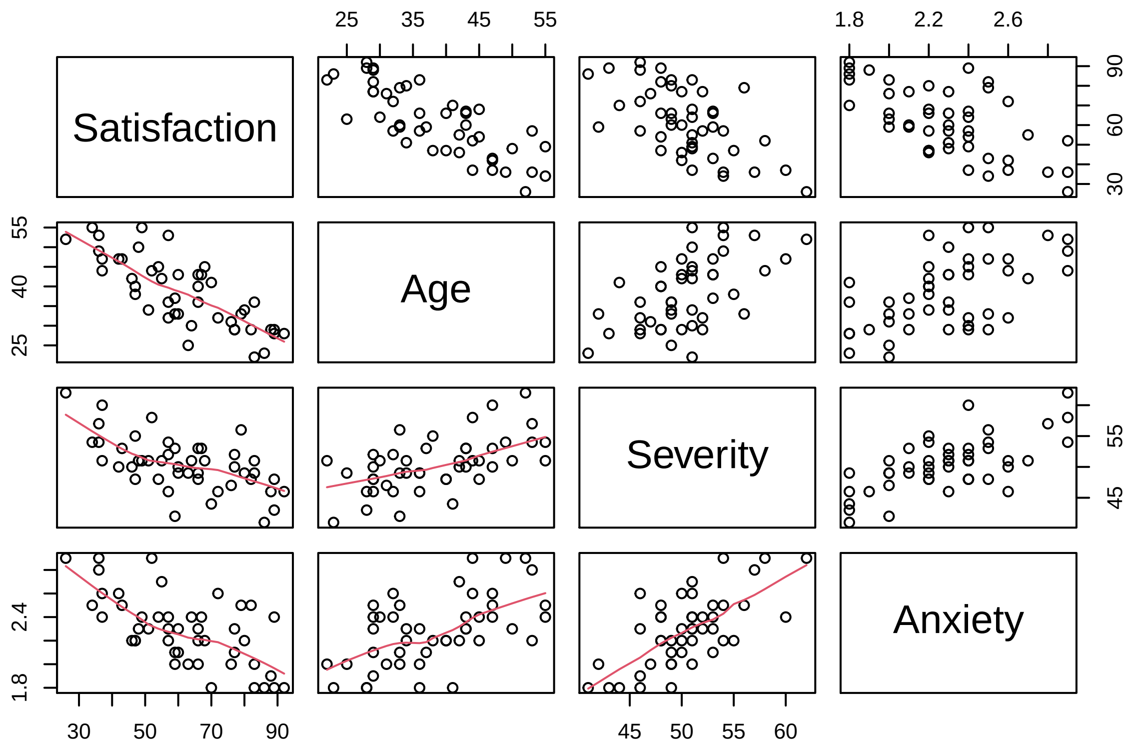




It is noteworthy that it seems all 3 plots are normally distributed.

b)

```
# scatter plot matrix
pairs(~Satisfaction +Age + Severity+ Anxiety, data = Data, lower.panel = panel.smooth)
```




```
cor(cbind(Data$Age, Data$Severity, Data$Anxiety))
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.5679505 0.5696775
## [2,] 0.5679505 1.0000000 0.6705287
## [3,] 0.5696775 0.6705287 1.0000000
```

Our scatter plot matrix shows that all 3 predictor value are positively correlated with each other, however, all 3 are negatively correlated with patient satisfaction.

Since none of the correlations between the predictor variables exceed 0.7, the correlations are not extreme enough to raise any concerns of multicollinearity.

c)

```
fit = lm(Satisfaction ~ Age + Severity + Anxiety, data = Data)
summary(fit)
```

```
##
## Call:
## lm(formula = Satisfaction ~ Age + Severity + Anxiety, data = Data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## Age          -1.1416     0.2148  -5.315 3.81e-06 ***
## Severity     -0.4420     0.4920  -0.898   0.3741
## Anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

The estimated regression function is

$$Y = 158.4912517 + -1.1416118 \cdot X_1 + -0.4420043 \cdot X_2 + -13.4701632 \cdot X_3$$

$\hat{\beta}_2$ is interpreted as, controlling for X_1 and X_3, a 1-unit increase in X_2 corresponds to a predicted increase of -1.1416118 in patient satisfaction.

d)

$$H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0, \quad H_a : \hat{\beta}_1 \neq 0 \vee \hat{\beta}_2 \neq 0 \vee \hat{\beta}_3 \neq 0$$

Decision rule : reject if

$$\text{P Value} = pf(F, 3, n - p') < \alpha$$

then we reject H_0

```
alpha = 0.10

anova = anova(fit)
anova
```

```
## Analysis of Variance Table
##
## Response: Satisfaction
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Age        1 8275.4  8275.4 81.8026 2.059e-11 ***
## Severity   1  480.9   480.9  4.7539   0.03489 *
## Anxiety    1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSR =  sum(anova[2][1:3,1])
MSE = anova[3][4,1]

Fstat = SSR/MSE/3
Fstat
```

```
## [1] 30.05208
```

$$Fstatistic = 30.0520779$$

```
Pval = pf(Fstat,4-1, anova[1][4,1],lower.tail=F)
Pval
```

```
## [1] 1.541973e-10
```

Since the P value of $1.5419726 \times 10^{-10}$ is less than the level of significance of $\alpha = 0.1$we can reject H_0

My test implies that we are 90% confident that $\beta_1, \beta_2, \beta_3$ are not all 0.

e)

```
SSE = anova[2][4,1]
SST = (SSR + SSE)
SSE
```

```
## [1] 4248.841
```

```
R = SSR/SST
R
```

```
## [1] 0.6821943
```

```
Radj = 1 - ((anova[1][4,1]+4-1)/anova[1][4,1])*(SSE/SST)
Radj
```

```
## [1] 0.6594939
```

The R^2 value of 0.6821943 means that $100 \cdot (0.6821943)\%$ of the data fit the regression model. Furthermore, since the adjusted R^2 value is close to R^2, we can say most prediction variables are having an effect on predictions.

f)

```
#fit
pred = predict(fit, data.frame(Age = 35, Severity = 45, Anxiety = 2.2 ) , interval="prediction", level =
pred
```

```
##        fit      lwr      upr
## 1 69.01029 51.50965 86.51092
```

I am 90% confident that a future patient with age 35, severity of illness 45, and anxiety level 2.2 will have a patient satisfaction within the range

$$[51.5096525, 86.51092]$$

Q4

a)

```r
Data = read.table("egyptcttn.txt",  col.names=c("Variety", "Luminance", "lnGrade"))
D1 = as.numeric(Data$Variety=="Giza67")
D2 = as.numeric(Data$Variety=="Giza68")
D3 = as.numeric(Data$Variety=="Giza69")
D4 = as.numeric(Data$Variety=="Giza70")

# base = Menoufi

fullFit = lm(Luminance ~ lnGrade*D1 + lnGrade*D2 + lnGrade*D3 + lnGrade*D4, data = Data)
summary(fullFit)
```

```
##
## Call:
## lm(formula = Luminance ~ lnGrade * D1 + lnGrade * D2 + lnGrade *
##     D3 + lnGrade * D4, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66004 -0.05597 -0.00598  0.10859  0.32705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.8034     1.3976  56.386 7.47e-14 ***
## lnGrade       3.3137     0.4243   7.810 1.45e-05 ***
## D1            2.0524     1.9765   1.038  0.32352
## D2            5.0801     1.9765   2.570  0.02788 *
## D3            7.2233     1.9765   3.655  0.00443 **
## D4            5.1151     1.9765   2.588  0.02704 *
## lnGrade:D1   -1.1507     0.6000  -1.918  0.08411 .
## lnGrade:D2   -1.0797     0.6000  -1.800  0.10212
## lnGrade:D3   -2.2741     0.6000  -3.790  0.00354 **
## lnGrade:D4   -2.0709     0.6000  -3.452  0.00621 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2907 on 10 degrees of freedom
## Multiple R-squared:  0.9794, Adjusted R-squared:  0.9609
## F-statistic: 52.82 on 9 and 10 DF,  p-value: 2.986e-07
```

Full Model :

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 \cdot lnGrade + \beta_2 D1 + \beta_3 D2 + \beta_4 D3 + \beta_5 D4 + \\ & \beta_6 \cdot lnGrade \cdot D1 + \beta_7 \cdot lnGrade \cdot D2 + \beta_8 \cdot lnGrade \cdot D3 + \beta_9 \cdot lnGrade \cdot D4 \end{aligned}$$

b)

$$H_0 : \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

$$H_a : \text{One of } \beta_6, \beta_7, \beta_8, \beta_9 \text{ is not } 0$$

```
reducedFitB = lm(Luminance ~ lnGrade + D1 + D2 + D3 + D4, data = Data)
anova = anova(reducedFitB, fullFit)
anova
```

```
## Analysis of Variance Table
##
## Model 1: Luminance ~ lnGrade + D1 + D2 + D3 + D4
## Model 2: Luminance ~ lnGrade * D1 + lnGrade * D2 + lnGrade * D3 + lnGrade *
##     D4
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     14 2.39566
## 2     10 0.84501  4    1.5507 4.5876 0.02313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Pval = anova$`Pr(>F)`[2]
Pval
```

```
## [1] 0.02313061
```

since

$$\text{P value } = 0.0231306 < \alpha = 0.05$$

we reject H_0.

c)

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

$$H_a : \text{One of } \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9 \text{ is not } 0$$

```
reducedFitC = lm(Luminance ~ lnGrade, data = Data)
anova = anova(reducedFitC, reducedFitB)
anova
```

```
## Analysis of Variance Table
##
## Model 1: Luminance ~ lnGrade
## Model 2: Luminance ~ lnGrade + D1 + D2 + D3 + D4
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     18 31.639
## 2     14  2.396  4    29.243 42.723 1.066e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Pval = anova$`Pr(>F)`[2]
Pval
```

```
## [1] 1.066057e-07
```

since

$$\text{P value } = 1.0660575 \times 10^{-7} < \alpha = 0.05$$

we reject H_0.

d)

The model I would choose for this data is the full model, with all dummy variables and interations. This is because we rejected that the interactions have no effect on slope, so we know the interactions have some significant effects on the model. So we want to use the full model to capture that.

e)

```
if(!require("ggplot2")) install.packages("ggplot2")
```

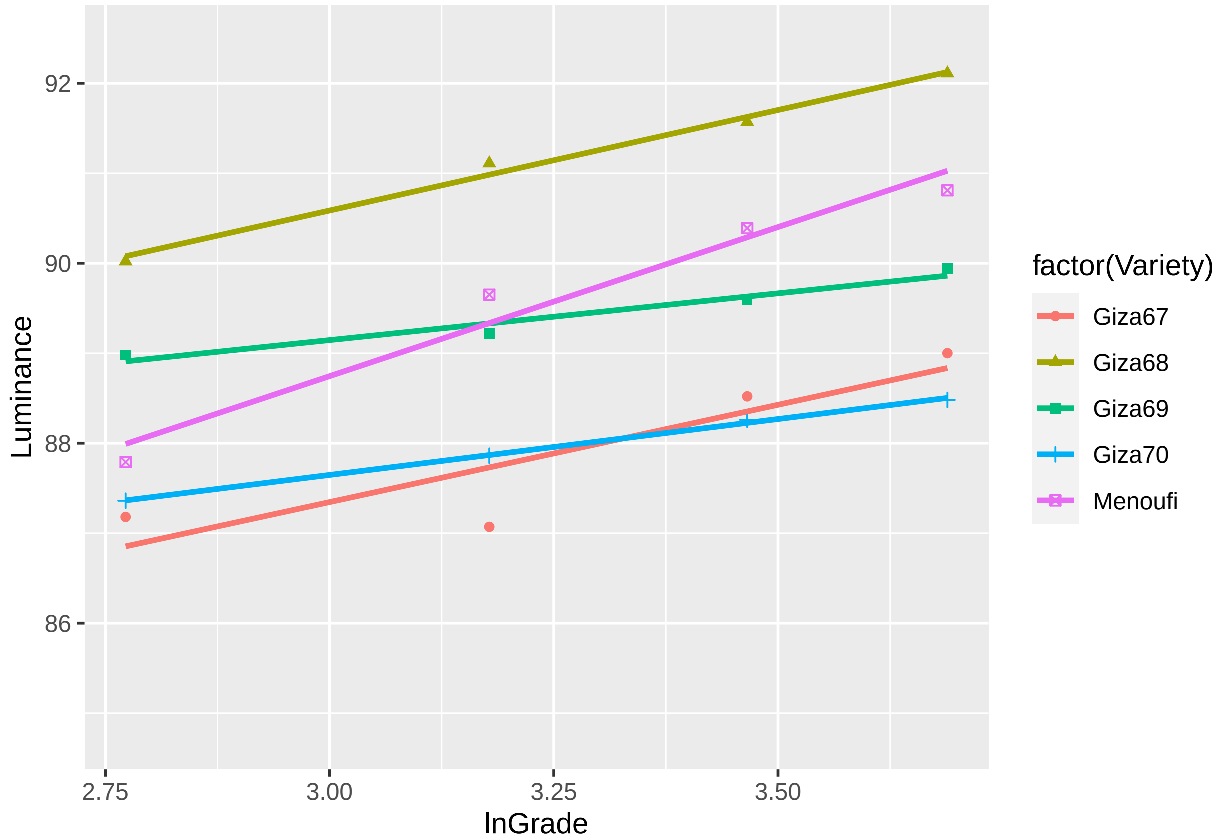
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
library(ggplot2)

ggplot(data=Data, aes(x=lnGrade, y=Luminance, color= factor(Variety), shape=factor(Variety))) + geom_poi
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Normality :

```r
p = shapiro.test(fullFit$residuals)$p.value
```

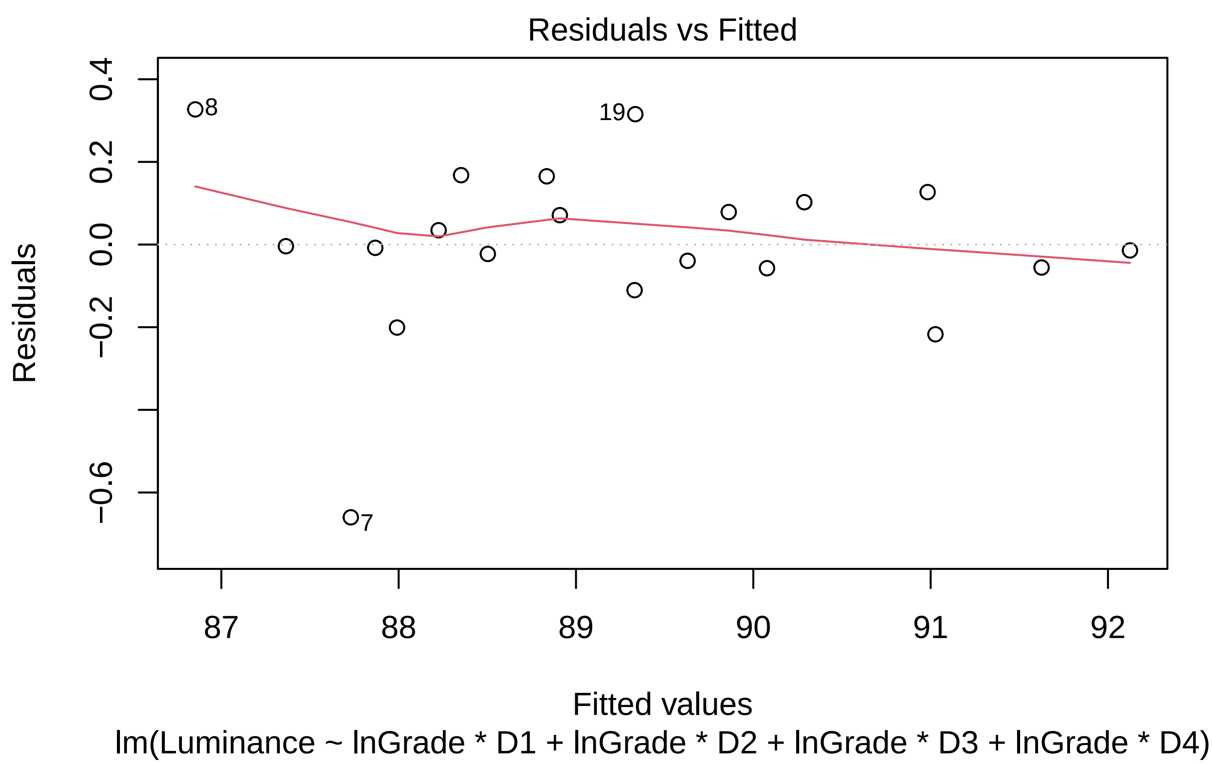
since

$$\text{P value} = 0.0190846 < \alpha = 0.05$$

we reject that this fit satisfies the normality assumption.

Equal Variance :

```r
plot(fullFit, which=1)
```



Visually, aside from 1 outlier at(87.7, -0.66), the residuals are uniformly distributed, and Equal Variance holds.

Linearity :

```r
summary(fullFit)
```

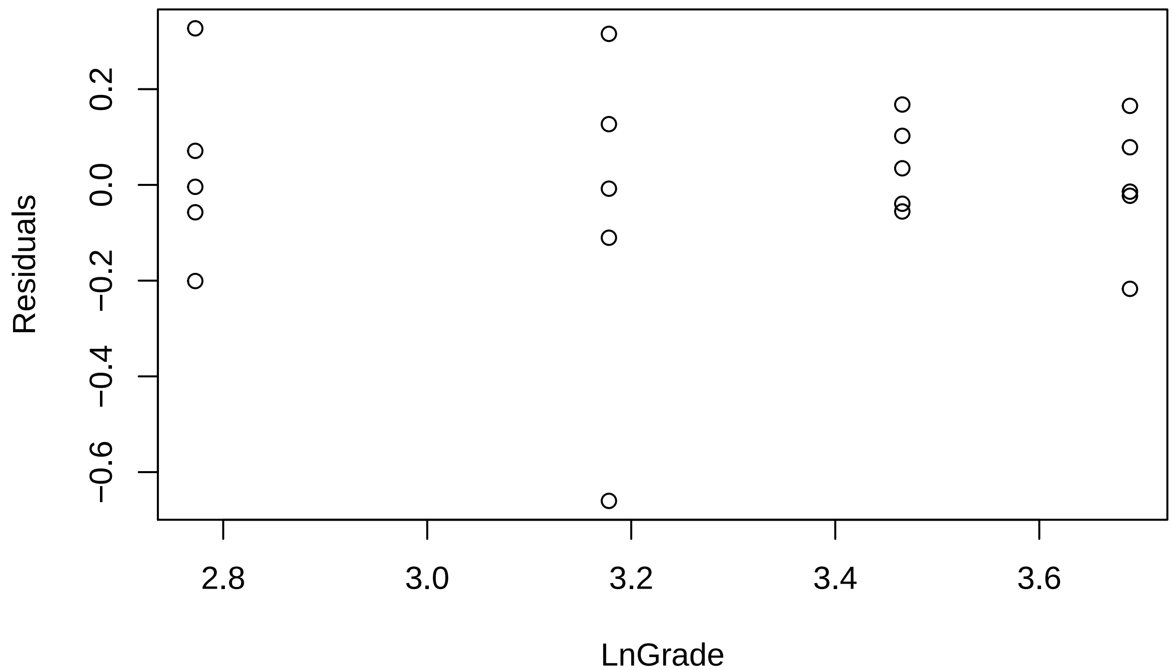
```
##
## Call:
## lm(formula = Luminance ~ lnGrade * D1 + lnGrade * D2 + lnGrade *
##     D3 + lnGrade * D4, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66004 -0.05597 -0.00598  0.10859  0.32705
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.8034     1.3976  56.386 7.47e-14 ***
## lnGrade       3.3137     0.4243   7.810 1.45e-05 ***
## D1            2.0524     1.9765   1.038  0.32352
## D2            5.0801     1.9765   2.570  0.02788 *
## D3            7.2233     1.9765   3.655  0.00443 **
## D4            5.1151     1.9765   2.588  0.02704 *
## lnGrade:D1   -1.1507     0.6000  -1.918  0.08411 .
## lnGrade:D2   -1.0797     0.6000  -1.800  0.10212
## lnGrade:D3   -2.2741     0.6000  -3.790  0.00354 **
## lnGrade:D4   -2.0709     0.6000  -3.452  0.00621 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2907 on 10 degrees of freedom
## Multiple R-squared:  0.9794, Adjusted R-squared:  0.9609
## F-statistic: 52.82 on 9 and 10 DF,  p-value: 2.986e-07
```

As we can see from the P values, we reject that any of the 5 slopes are 0. Also, from the graph we can confidently see that there is a linear relationship between luminance and lnGrade for all 5 categories.
from these 2 observations, we can say our model satisfies linearity.

Independent/uncorrelated error terms :

```
plot(Data$lnGrade, resid(fullFit),
     ylab="Residuals", xlab="LnGrade",)
```



From this plot, I visually access that there is no major deviate patterns, therefor Independent/uncorrelated error terms is satisfied.

Question 5.

```
WoolStrengthData <- read.table("StrengthWool.txt", header = TRUE)
```

Part a.

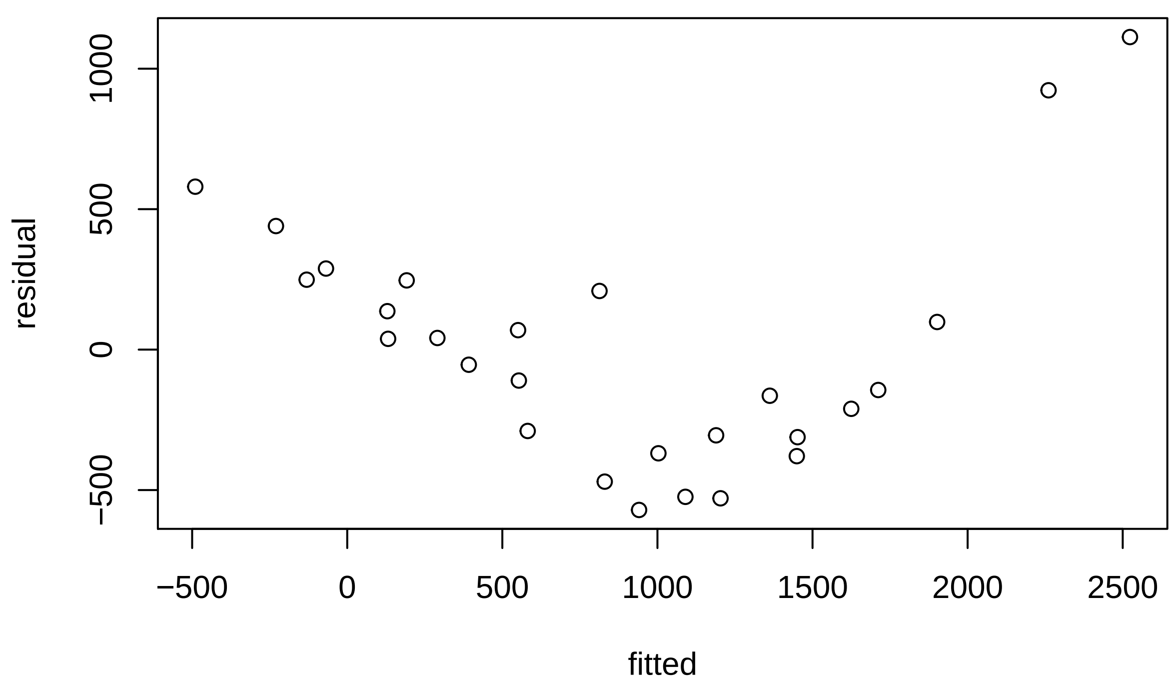
Fit a linear model with just main effects for the three class variables and show that this is not a good fit to the data

```
Cycles = WoolStrengthData$Cycles
lenCut <- cut(WoolStrengthData$Len, breaks=c(0,250, 300, 350), labels = c("250", "300", "350"))
ampCut <- cut(WoolStrengthData$Amp, breaks=c(0,8,9,10), labels = c("8", "9", "10"))
loadCut <- cut(WoolStrengthData$Load, breaks=c(0,40,45,50), labels=c("40", "45", "50"))
fit <- lm(Cycles ~ lenCut + ampCut + loadCut, data = WoolStrengthData)
summary(fit)
```

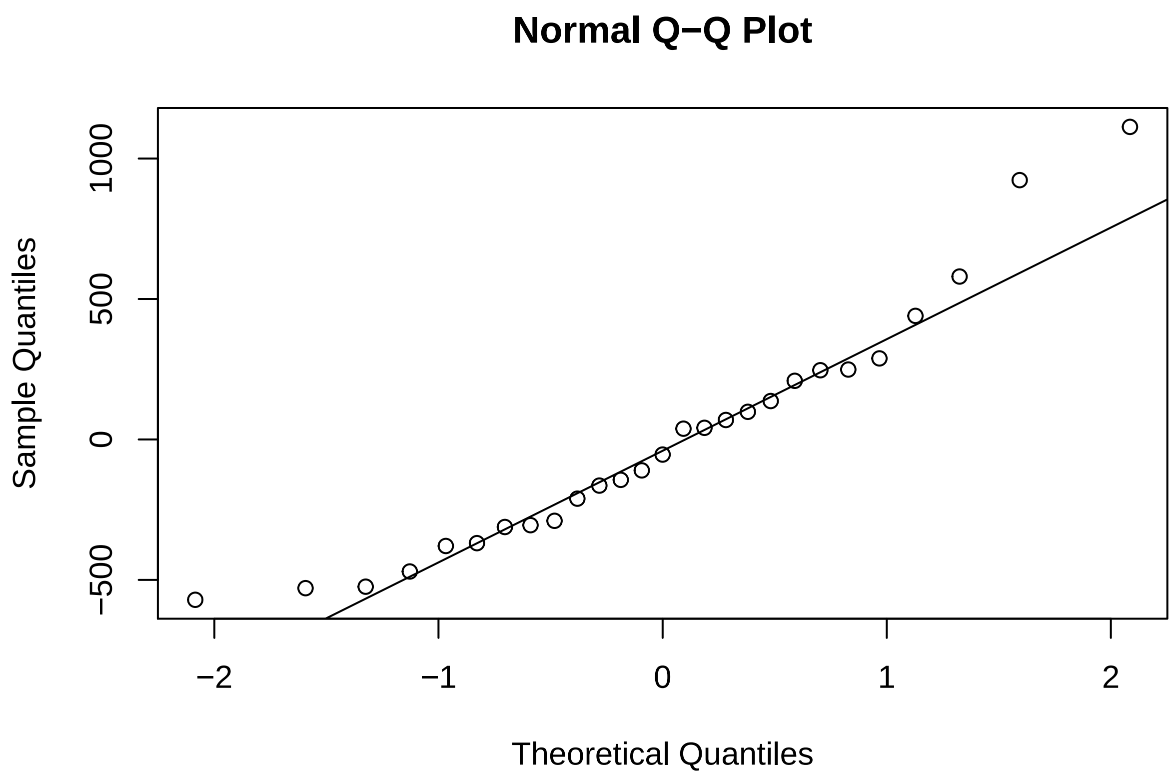
```
##
## Call:
## lm(formula = Cycles ~ lenCut + ampCut + loadCut, data = WoolStrengthData)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -570.81 -308.43  -53.81  227.57 1112.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1203.4      246.0   4.891 8.83e-05 ***
## lenCut300      421.4      227.8   1.850 0.079096 .
## lenCut350     1320.0      227.8   5.795 1.14e-05 ***
## ampCut9       -811.6      227.8  -3.563 0.001948 **
## ampCut10     -1071.7      227.8  -4.705 0.000136 ***
## loadCut45     -262.6      227.8  -1.153 0.262611
## loadCut50     -621.7      227.8  -2.729 0.012918 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 483.2 on 20 degrees of freedom
## Multiple R-squared:  0.7692, Adjusted R-squared:  0.6999
## F-statistic: 11.11 on 6 and 20 DF,  p-value: 1.769e-05
```

```r
residual <- fit$residuals
fitted <- fit$fitted.values
plot(x=fitted, y=residual)
```



```
qqnorm(residual)
qqline(residual)
```



If we look at the plot of the Residual Plot as well as the QQ Plot of the model, we can see that this model is not a good fit to the data, where the QQ Plot shows many points not following the linear line, as well as the curved shape in the residual plot not showing signs of a good fit.

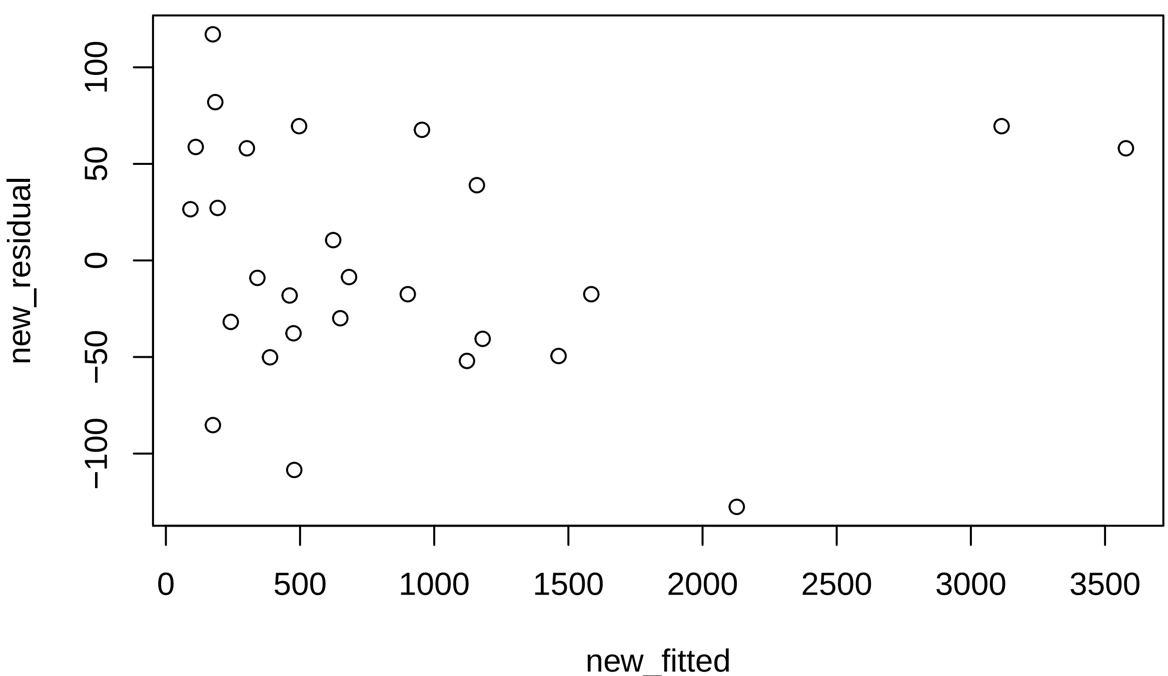
#5.b)

```
new_fit <- lm(Cycles ~ ampCut + lenCut + loadCut + ampCut*lenCut + ampCut*loadCut + lenCut*loadCut, data
summary(new_fit)
```

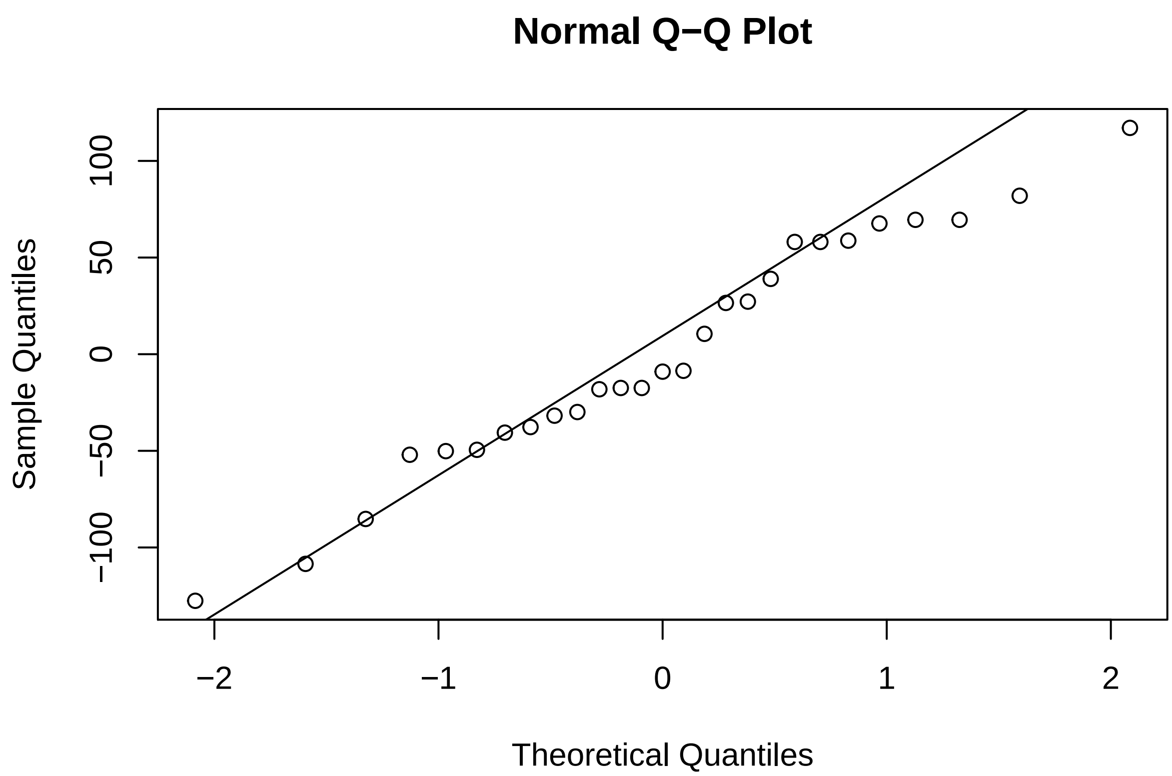
```
##
## Call:
## lm(formula = Cycles ~ ampCut + lenCut + loadCut + ampCut * lenCut +
##     ampCut * loadCut + lenCut * loadCut, data = WoolStrengthData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.593  -39.148   -9.037   58.074  117.074
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.826e+02  9.237e+01   7.390 7.69e-05 ***
## ampCut9             -2.944e+02  1.161e+02  -2.537 0.034879 *
## ampCut10            -5.713e+02  1.161e+02  -4.923 0.001160 **
```

```
## lenCut300           7.809e+02  1.161e+02   6.728 0.000148 ***
## lenCut350           2.895e+03  1.161e+02  24.946 7.13e-09 ***
## loadCut45          -2.041e+02  1.161e+02  -1.759 0.116697
## loadCut50          -5.077e+02  1.161e+02  -4.374 0.002368 **
## ampCut9:lenCut300   -2.147e+02  1.271e+02  -1.688 0.129813
## ampCut10:lenCut300  -4.310e+02  1.271e+02  -3.390 0.009502 **
## ampCut9:lenCut350   -1.698e+03  1.271e+02 -13.355 9.45e-07 ***
## ampCut10:lenCut350  -1.826e+03  1.271e+02 -14.362 5.40e-07 ***
## ampCut9:loadCut45    1.255e-12  1.271e+02   0.000 1.000000
## ampCut10:loadCut45   1.843e+02  1.271e+02   1.450 0.185155
## ampCut9:loadCut50    3.613e+02  1.271e+02   2.842 0.021747 *
## ampCut10:loadCut50   5.717e+02  1.271e+02   4.496 0.002012 **
## lenCut300:loadCut45 -1.003e+02  1.271e+02  -0.789 0.452782
## lenCut350:loadCut45 -2.593e+02  1.271e+02  -2.040 0.075709 .
## lenCut300:loadCut50 -3.323e+02  1.271e+02  -2.614 0.030944 *
## lenCut350:loadCut50 -9.427e+02  1.271e+02  -7.414 7.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 110.1 on 8 degrees of freedom
## Multiple R-squared:  0.9952, Adjusted R-squared:  0.9844
## F-statistic: 92.25 on 18 and 8 DF,  p-value: 2.537e-07
```

```r
new_residual <- new_fit$residuals
new_fitted <- new_fit$fitted.values
plot(x=new_fitted, y=new_residual)
```



```r
qqnorm(new_residual)
qqline(new_residual)
```



If we look at the new residual plot of the new model, we can see that this model is a better fit to the data, as we can see that the variance is linear around 0 but does not have constant variance. Hence, the model which considers all the interactions between pairs of the class variables is a better model for fitting the data.

Part c.

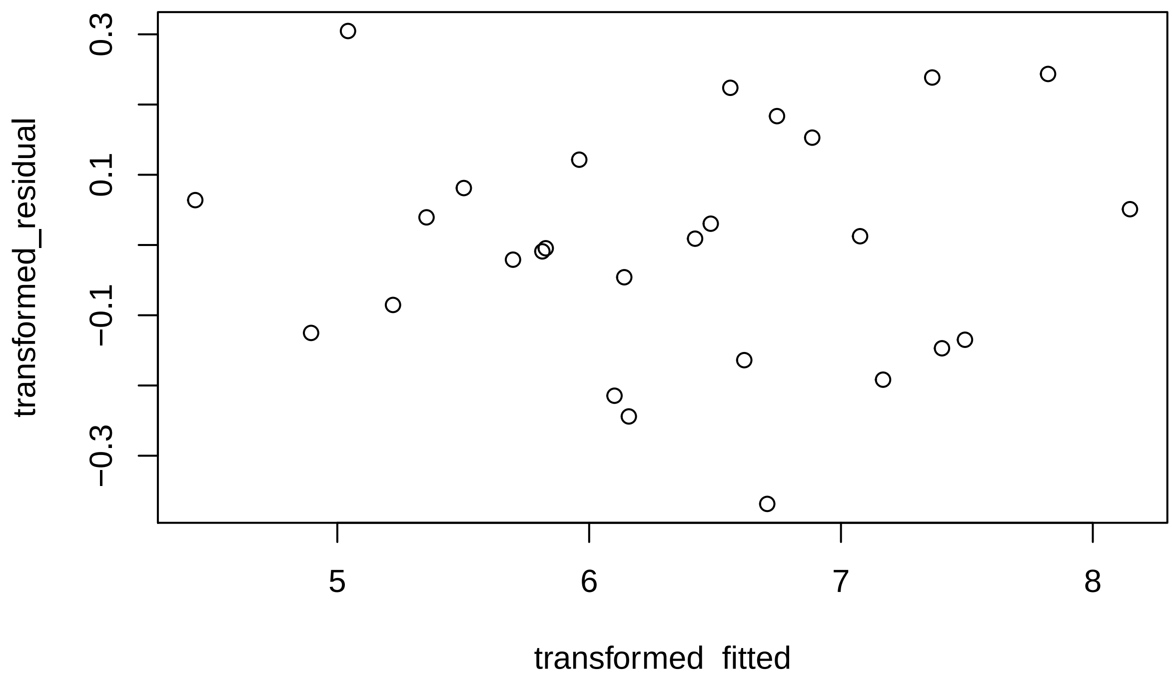
```r
transformed_fit <- lm(log(Cycles)~(lenCut + ampCut + loadCut))

summary(transformed_fit)
```

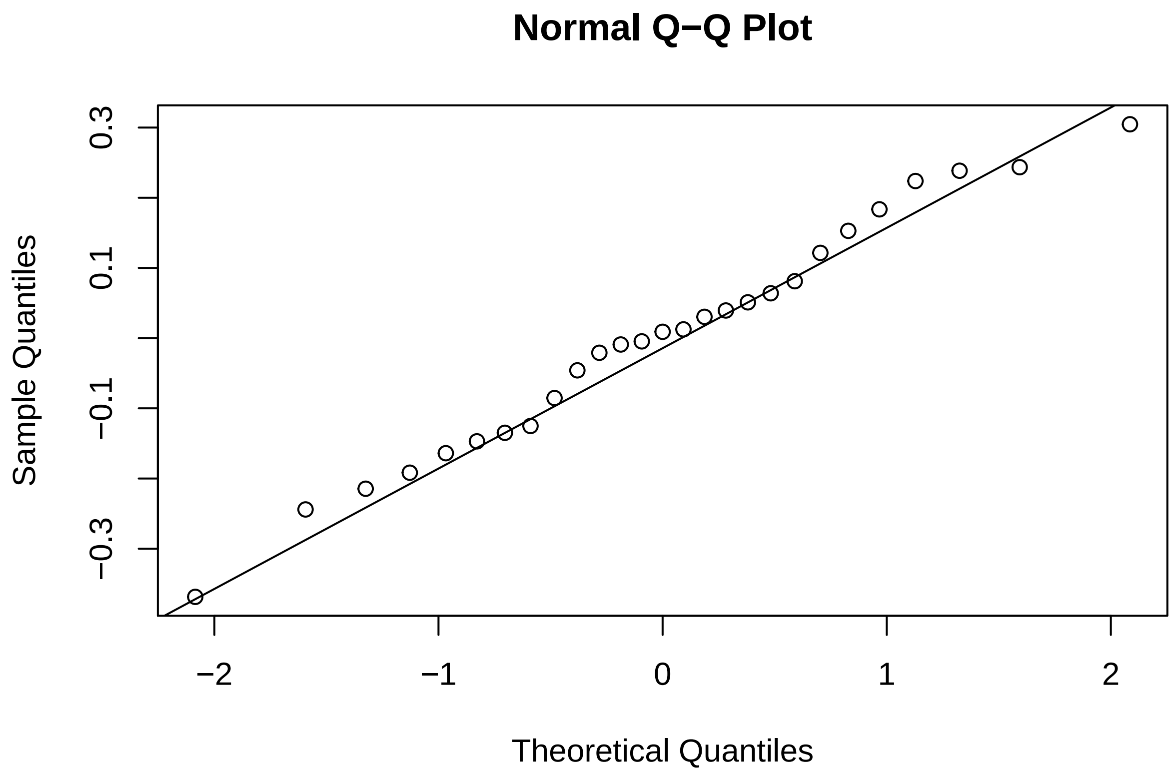
```
##
## Call:
## lm(formula = log(Cycles) ~ (lenCut + ampCut + loadCut))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36860 -0.13002  0.00902  0.10129  0.30469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.48287    0.09644  67.225  < 2e-16 ***
```

```
## lenCut300    0.91833    0.08928  10.286 1.97e-09 ***
## lenCut350    1.66477    0.08928  18.646 4.10e-14 ***
## ampCut9     -0.65521    0.08928  -7.339 4.31e-07 ***
## ampCut10    -1.26173    0.08928 -14.132 7.19e-12 ***
## loadCut45   -0.32529    0.08928  -3.643  0.00162 **
## loadCut50   -0.78524    0.08928  -8.795 2.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1894 on 20 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.9598
## F-statistic: 104.5 on 6 and 20 DF,  p-value: 4.979e-14
```

```r
transformed_residual <- transformed_fit$residuals
transformed_fitted <- transformed_fit$fitted.values
plot(x=transformed_fitted, y=transformed_residual)
```




```r
qqnorm(transformed_residual)
qqline(transformed_residual)
```



After performing the log transformation on the model, we can see that the variance is constant and linear around 0. The high R/Radj values also show that almost all of the variance can be explained by the model. This is a better fit for the model.

Part d.

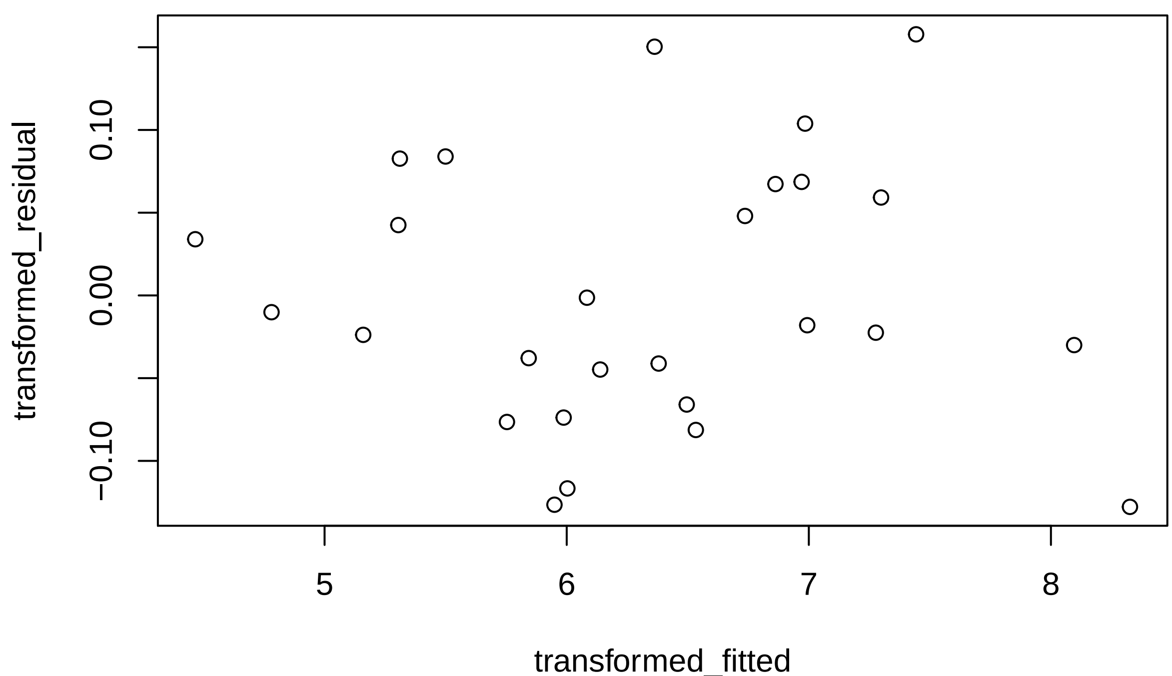
```
transformed_fit2 <- lm(log(Cycles)~(ampCut + lenCut + loadCut + ampCut*lenCut + ampCut*loadCut + lenCut*

summary(transformed_fit2)
```

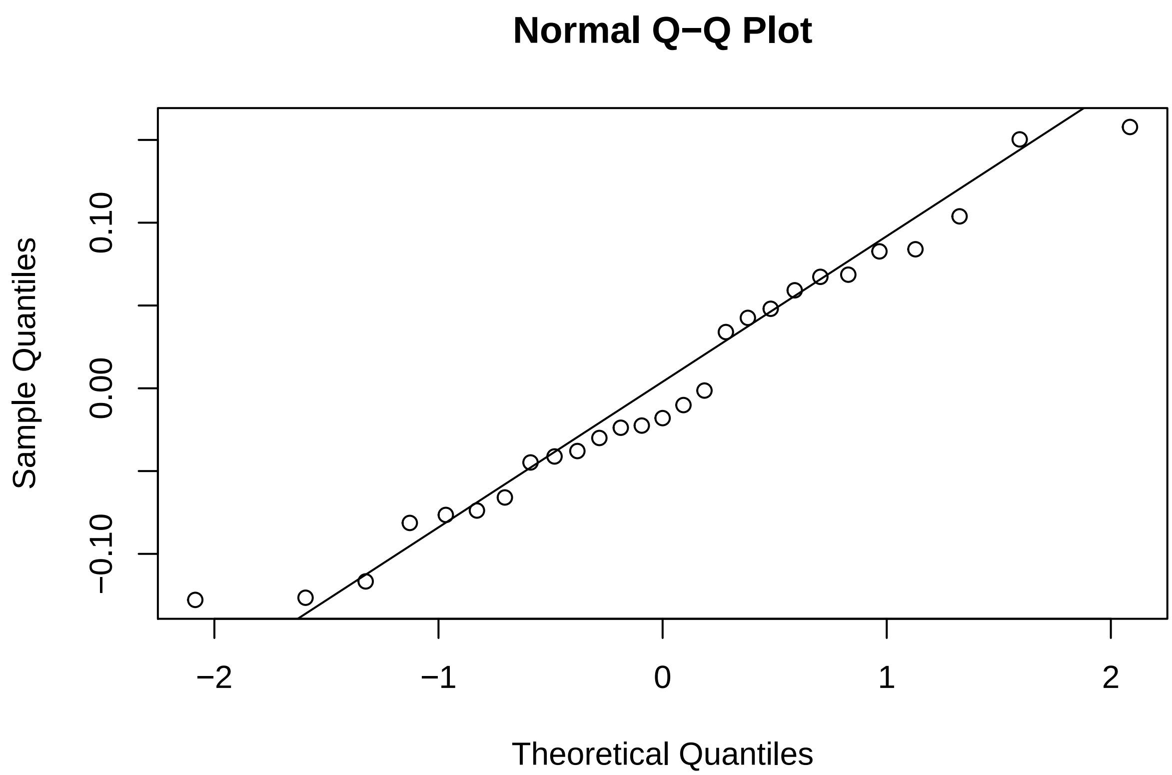
```
##
## Call:
## lm(formula = log(Cycles) ~ (ampCut + lenCut + loadCut + ampCut *
##     lenCut + ampCut * loadCut + lenCut * loadCut))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12779 -0.05537 -0.01802  0.06325  0.15780
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.362917   0.120807  52.670 1.87e-11 ***
## ampCut9               -0.413379   0.151801  -2.723 0.026121 *
## ampCut10              -1.203298   0.151801  -7.927 4.67e-05 ***
```

```
## lenCut300            0.913780   0.151801   6.020 0.000316 ***
## lenCut350            1.963516   0.151801  12.935 1.21e-06 ***
## loadCut45           -0.375588   0.151801  -2.474 0.038457 *
## loadCut50           -0.609676   0.151801  -4.016 0.003861 **
## ampCut9:lenCut300   -0.001114   0.166290  -0.007 0.994817
## ampCut10:lenCut300   0.064964   0.166290   0.391 0.706242
## ampCut9:lenCut350   -0.614678   0.166290  -3.696 0.006074 **
## ampCut10:lenCut350  -0.152966   0.166290  -0.920 0.384537
## ampCut9:loadCut45   -0.074416   0.166290  -0.448 0.666379
## ampCut10:loadCut45  -0.003211   0.166290  -0.019 0.985067
## ampCut9:loadCut50   -0.035285   0.166290  -0.212 0.837264
## ampCut10:loadCut50  -0.084089   0.166290  -0.506 0.626717
## lenCut300:loadCut45  0.083463   0.166290   0.502 0.629248
## lenCut350:loadCut45  0.145059   0.166290   0.872 0.408448
## lenCut300:loadCut50 -0.133655   0.166290  -0.804 0.444766
## lenCut350:loadCut50 -0.273658   0.166290  -1.646 0.138450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.144 on 8 degrees of freedom
## Multiple R-squared:  0.9928, Adjusted R-squared:  0.9768
## F-statistic: 61.71 on 18 and 8 DF,  p-value: 1.236e-06
```

```
transformed_residual <- transformed_fit2$residuals
transformed_fitted <- transformed_fit2$fitted.values
plot(x=transformed_fitted, y=transformed_residual)
```



```
qqnorm(transformed_residual)
qqline(transformed_residual)
```




```
anova(transformed_fit, transformed_fit2)
```

```
## Analysis of Variance Table
##
## Model 1: log(Cycles) ~ (lenCut + ampCut + loadCut)
## Model 2: log(Cycles) ~ (ampCut + lenCut + loadCut + ampCut * lenCut +
##     ampCut * loadCut + lenCut * loadCut)
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     20 0.71742
## 2      8 0.16591 12   0.55151 2.216 0.1325
```

We can see that the result of this is very similar to that of part c. However, it is not a significant change in terms of deciding a better fit, and we can use F-test. Since the p-value of F-test is $0.1325 > 0.05$, With 95% confidence we cannot reject H0 as there is evidence that the interaction terms have no effect on the model with transformations