STAC67A3

Yulun Wu

20/03/2022

Q3

(a)

```r
DataQ3 = read.table("SENIC.txt", header=F,
                    col.names = c("I","Y","X1","X2","X3","X4","X5","Z1","Z2","X6","X7","X8"))
quantdata = data.frame(DataQ3$Y,DataQ3$X1,DataQ3$X2,DataQ3$X3,
                       DataQ3$X4,DataQ3$X5,DataQ3$X6,DataQ3$X7,DataQ3$X8)
colnames(quantdata) = c("Y","X1","X2","X3","X4","X5","X6","X7","X8")
fitQ31 = lm(quantdata$Y~quantdata$X1+quantdata$X2+quantdata$X3+
              quantdata$X4+quantdata$X5+quantdata$X6+quantdata$X7+
              quantdata$X8)

summary(fitQ31)
```

```
##
## Call:
## lm(formula = quantdata$Y ~ quantdata$X1 + quantdata$X2 + quantdata$X3 +
##     quantdata$X4 + quantdata$X5 + quantdata$X6 + quantdata$X7 +
##     quantdata$X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5292 -0.9263 -0.1235  0.7756  6.6474
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.490678   1.665435   0.295  0.76887
## quantdata$X1  0.096009   0.029249   3.283  0.00140 **
## quantdata$X2  0.335551   0.130969   2.562  0.01184 *
## quantdata$X3  0.027130   0.015885   1.708  0.09063 .
## quantdata$X4  0.018671   0.007507   2.487  0.01447 *
## quantdata$X5 -0.009578   0.003618  -2.648  0.00937 **
## quantdata$X6  0.021627   0.004286   5.046 1.92e-06 ***
## quantdata$X7 -0.006577   0.002341  -2.810  0.00593 **
## quantdata$X8  0.001136   0.014202   0.080  0.93640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.321 on 104 degrees of freedom
## Multiple R-squared:  0.5568, Adjusted R-squared:  0.5227
## F-statistic: 16.33 on 8 and 104 DF,  p-value: 2.001e-15
```

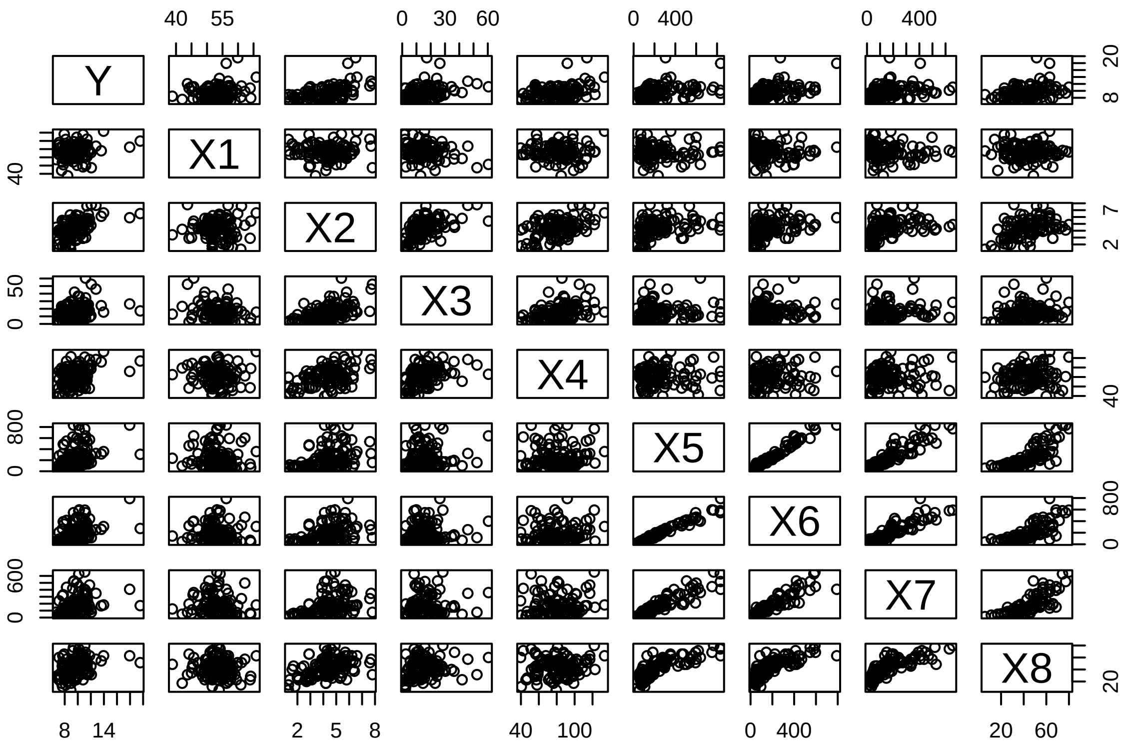
```
#X5 = factor(DataQ3[,8])
#Z= factor(DataQ3[,9])
```

(b)

Since p-value of β_0, β_3, β_8 are greater than 0.05, they are not significantly different from 0.

(c)

```
pairs(quantdata)
```




```
cor(quantdata)
```

```
##            Y           X1          X2         X3          X4          X5
## Y  1.0000000  0.188913972 0.533443831  0.3266838  0.38248193  0.40926525
## X1 0.1889140  1.000000000 0.001093166 -0.2258468 -0.01885490 -0.05882316
## X2 0.5334438  0.001093166 1.000000000  0.5591589  0.45339156  0.35977000
## X3 0.3266838 -0.225846789 0.559158869  1.0000000  0.42496204  0.13972495
## X4 0.3824819 -0.018854897 0.453391557  0.4249620  1.00000000  0.04581997
## X5 0.4092652 -0.058823160 0.359770000  0.1397249  0.04581997  1.00000000
## X6 0.4738855 -0.054774667 0.381411081  0.1429482  0.06291352  0.98099774
## X7 0.3403671 -0.082944616 0.393981340  0.1988998  0.07738133  0.91550415
## X8 0.3555379 -0.040451379 0.412600675  0.1851311  0.11192761  0.79452438
##             X6          X7          X8
## Y   0.47388550  0.34036706  0.35553792
## X1 -0.05477467 -0.08294462 -0.04045138
## X2  0.38141108  0.39398134  0.41260068
## X3  0.14294821  0.19889983  0.18513114
## X4  0.06291352  0.07738133  0.11192761
## X5  0.98099774  0.91550415  0.79452438
```

```
## X6  1.00000000  0.90789698  0.77806330
## X7  0.90789698  1.00000000  0.78350550
## X8  0.77806330  0.78350550  1.00000000
```

There is concern about multicollinearity, the correlation between some pair of X variables are too high. For example, cor(X5,X6)=0.98099774, cor(X5,X7)=0.91550415, cor(X6,X7)=0.90789698.

(d)

```
fitQ32 = lm(quantdata$Y~quantdata$X1+quantdata$X2+
              quantdata$X3+quantdata$X4+quantdata$X6+
              quantdata$X7+quantdata$X8)

summary(fitQ32)
```

```
##
## Call:
## lm(formula = quantdata$Y ~ quantdata$X1 + quantdata$X2 + quantdata$X3 +
##     quantdata$X4 + quantdata$X6 + quantdata$X7 + quantdata$X8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6431 -0.8342 -0.0612  0.7101  6.8879
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.366894   1.711763   0.214 0.830700
## quantdata$X1  0.095676   0.030074   3.181 0.001928 **
## quantdata$X2  0.377485   0.133677   2.824 0.005679 **
## quantdata$X3  0.025825   0.016325   1.582 0.116680
## quantdata$X4  0.019855   0.007705   2.577 0.011361 *
## quantdata$X6  0.011623   0.002080   5.588 1.82e-07 ***
## quantdata$X7 -0.008166   0.002326  -3.510 0.000661 ***
## quantdata$X8 -0.006802   0.014274  -0.477 0.634665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.358 on 105 degrees of freedom
## Multiple R-squared:  0.5269, Adjusted R-squared:  0.4953
## F-statistic:  16.7 on 7 and 105 DF,  p-value: 1.183e-14
```

R^2 before I dropped X5 is 0.5568, after I dropped X5 is 0.5269. R^2 didn't change much after dropping X5, means that adding X5 to our model does not improve the prediction on Y much.

Q4

(a)

```
DataQ4 = read.csv("kidiq-1.csv", header=T)
colnames(DataQ4) = c("Y","Z1","X1","Z2","X2")
fitQ4a = lm(Y~Z1,data=DataQ4)
summary(fitQ4a)
```

```
##
```

```
## Call:
## lm(formula = Y ~ Z1, data = DataQ4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57.55 -13.32   2.68  14.68  58.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   77.548      2.059  37.670  < 2e-16 ***
## Z1            11.771      2.322   5.069 5.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.85 on 432 degrees of freedom
## Multiple R-squared:  0.05613,    Adjusted R-squared:  0.05394
## F-statistic: 25.69 on 1 and 432 DF,  p-value: 5.957e-07
```

β_0: When kids' mother not graduated from high school, we expect kids' iq score to be 77.548. When kids' mother graduated from high school, we expect kids' iq score to be 89.319.
β_1: The kid's iq is 11.771 bigger for Kids with mother graduated from high school compared to Kids with mother not graduated from high school.

(b)

```r
fitQ4b = lm(Y~Z1+X1,data=DataQ4)
summary(fitQ4b)
```

```
##
## Call:
## lm(formula = Y ~ Z1 + X1, data = DataQ4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.873 -12.663   2.404  11.356  49.545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
## Z1           5.95012    2.21181   2.690  0.00742 **
## X1           0.56391    0.06057   9.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 431 degrees of freedom
## Multiple R-squared:  0.2141, Adjusted R-squared:  0.2105
## F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

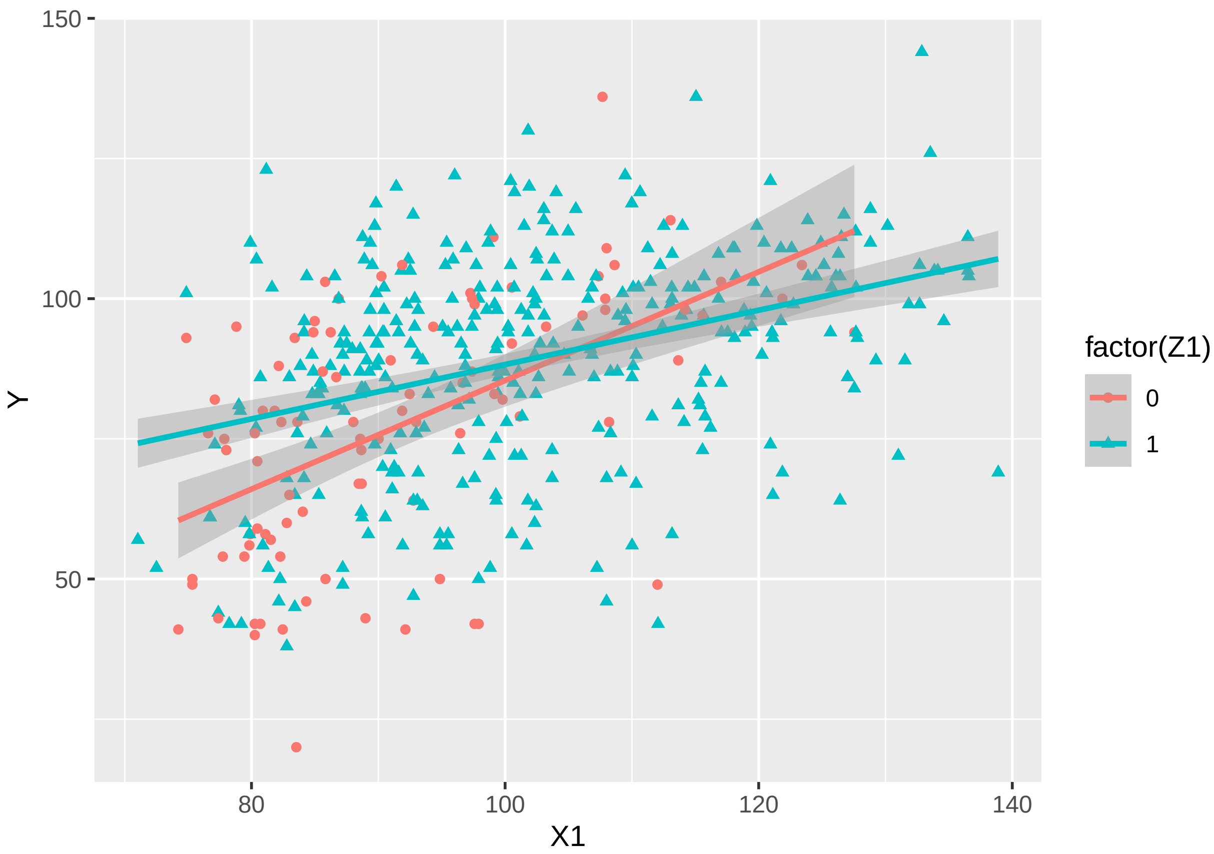
β_0: When mother of kid is not graduated from high school and has 0 iq score, we expect the kids' iq score to be 25.73154.
β_1: The iq score of kids is expected to be 5.95012 bigger for kids with mother graduated from high school compared to kids with mother not graduated from high school, for any given level of mothers' iq score.
β_2: The kid's iq score will increase by 0.56391 if mother's iq score is increased by 1 for kids with mother that has any of two graduation status.

(c)

```r
fitQ4c = lm(Y~Z1+X1+Z1:X1,data=DataQ4)
summary(fitQ4c)
```

```
##
## Call:
## lm(formula = Y ~ Z1 + X1 + Z1:X1, data = DataQ4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.092 -11.332   2.066  11.663  43.880
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.4820    13.7580  -0.835 0.404422
## Z1           51.2682    15.3376   3.343 0.000902 ***
## X1            0.9689     0.1483   6.531 1.84e-10 ***
## Z1:X1        -0.4843     0.1622  -2.985 0.002994 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.97 on 430 degrees of freedom
## Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
## F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16
```

```r
library(ggplot2)
ggplot(data=DataQ4, aes(x=X1,y=Y,color=factor(Z1),shape=factor(Z1))) + geom_point() +
  geom_smooth(method='lm',formula=y~x)
```



β_0:

When mother of kid is not graduated from high school and has 0 iq, we expect the kids' iq to be -11.4820. β_1: The iq of kids is expected to be 51.2682 bigger for kids with mother graduated from high school compared to Kids with mother not graduated from high school, for any given level of mothers' iq.
β_2: The kid's iq score will increase by 0.9689 if kids' mother's iq score is increased by 1 for kids with mother not graduated from high school. β_3: The kids' iq score will have additional decrease of 0.4843 if kids' mother's iq score is increased by 1 for kids with mother graduated from high school compared to kids with mother not graduated from high school that have the same level of mother's iq score.
Children of mothers who are graduated from high school: Y=39.7862+0.4846*X1*
*Children of mothers who are not graduated from high school: Y=-11.4820+0.9689*X1
(d) p-value for coefficient of interaction term is 0.002994<0.05, so we reject the null hypothesis that coefficient of interaction term is 0. We can conclude that slopes relating mother's IQ to child test scores depends on maternal high school indicator.
(e)

```r
fitQ4e = lm(Y~Z1+X1+Z1:X1+factor(Z2)+X2,data=DataQ4)
fitQ4e_r1 = lm(Y~Z1+X1+Z1:X1+X2,data=DataQ4)
summary(fitQ4e)
```

```
##
## Call:
## lm(formula = Y ~ Z1 + X1 + Z1:X1 + factor(Z2) + X2, data = DataQ4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.053 -11.439   1.884  11.465  44.417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.5745    16.2245  -1.206 0.228303
## Z1           51.6398    15.5817   3.314 0.000998 ***
## X1            0.9641     0.1500   6.426 3.51e-10 ***
## factor(Z2)2   1.9464     2.8083   0.693 0.488633
## factor(Z2)3   4.9426     3.2275   1.531 0.126411
## factor(Z2)4   0.7867     2.5019   0.314 0.753353
## X2            0.3344     0.3328   1.005 0.315554
## Z1:X1        -0.4940     0.1647  -2.998 0.002874 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.97 on 426 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2248
## F-statistic: 18.94 on 7 and 426 DF,  p-value: < 2.2e-16
```

```r
anova(fitQ4e_r1, fitQ4e)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ Z1 + X1 + Z1:X1 + X2
## Model 2: Y ~ Z1 + X1 + Z1:X1 + factor(Z2) + X2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    429 138511
## 2    426 137568  3    942.97 0.9733 0.4052
```

```r
anova(fitQ4c, fitQ4e)
```

```
## Analysis of Variance Table
```

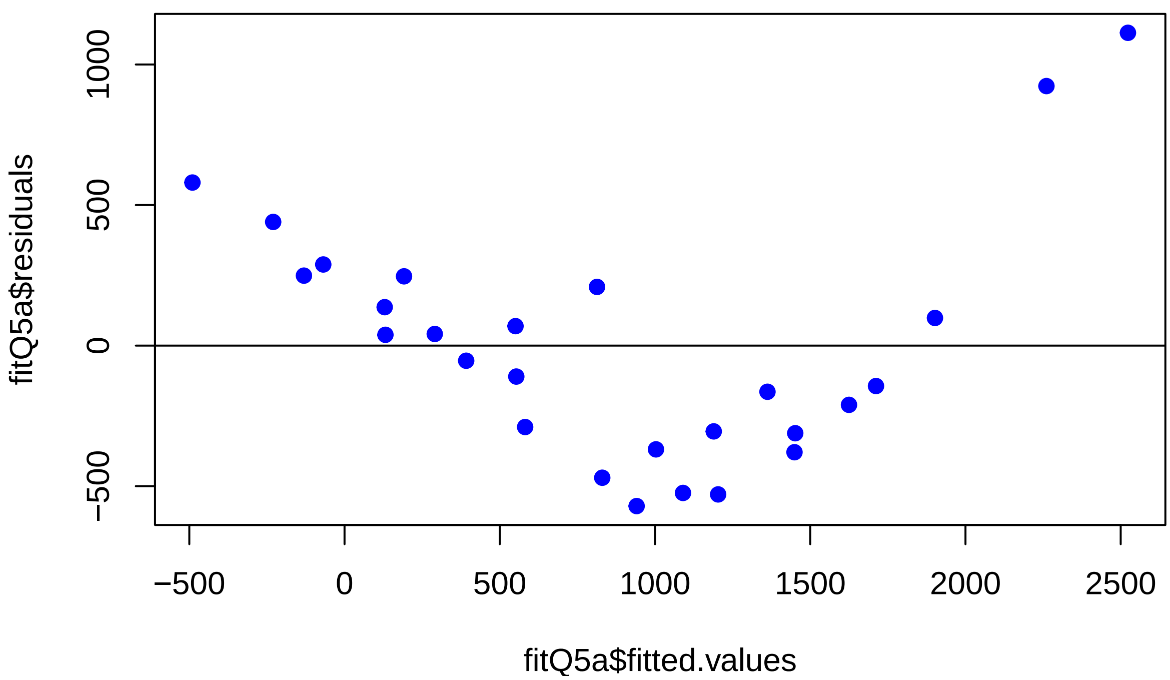
```
##
## Model 1: Y ~ Z1 + X1 + Z1:X1
## Model 2: Y ~ Z1 + X1 + Z1:X1 + factor(Z2) + X2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    430 138879
## 2    426 137568  4    1310.9 1.0149 0.3993
```

Test whether mom's work status is a significant predictor with $\alpha = 0.05$: p-value of F-test of full model(Y ~ Z1 + X1 + Z1:X1 + factor(Z2) + X2) against reduced model(Y ~ Z1 + X1 + Z1:X1 + X2) is 0.4052>0,05, so we fail to reject the null hypothesis that mom's work status is not a significant predictor. Therefore, we conclude that mom's work status is not a significant predictor.
Test whether both mom's work status and mom's age are significant predictors with $\alpha = 0.05$: p-value of F-test of full model(Y ~ Z1 + X1 + Z1:X1 + factor(Z2) + X2) against reduced model(Y ~ Z1 + X1 + Z1:X1) is 0.3993>0,05, so we fail to reject the null hypothesis that both mom's work status and mom's age are not significant predictor. Therefore, we conclude that both mom's work status and mom's age are not significant predictor.

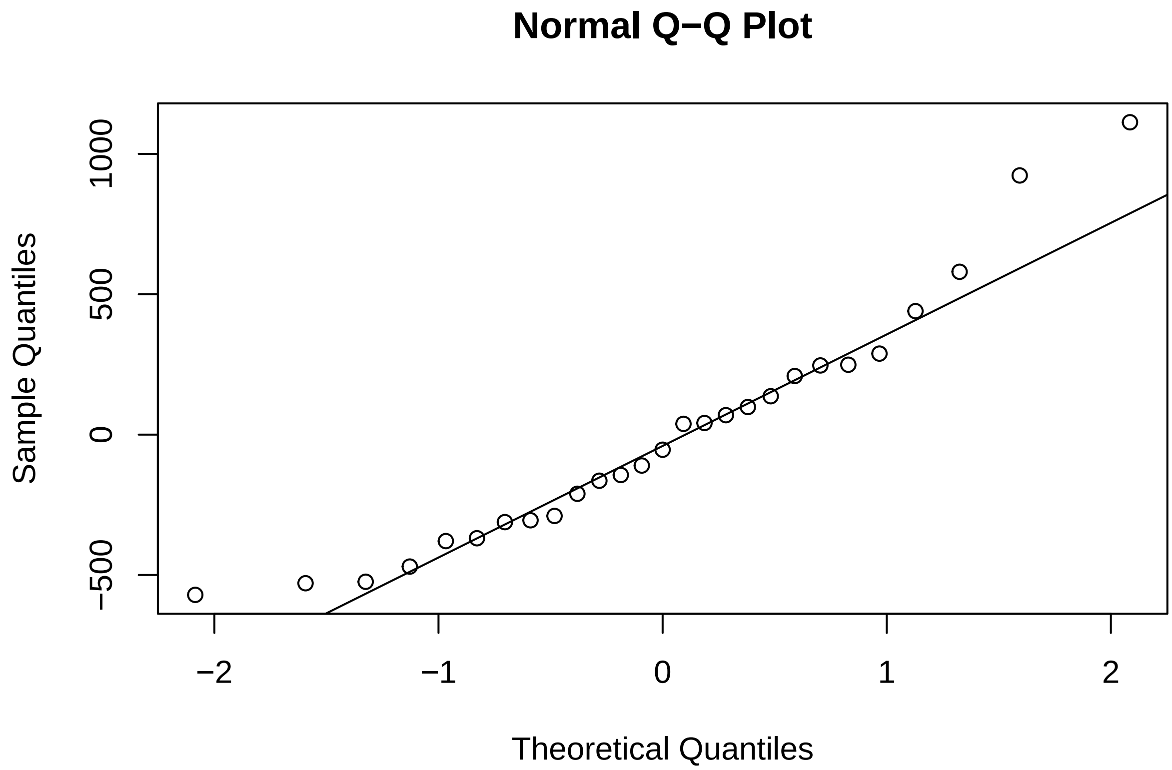
Q5

(a)

```r
DataQ5 = read.table("StrengthWool.txt", header=T)
fitQ5a = lm(Cycles~factor(Len)+factor(Amp)+factor(Load),data=DataQ5)
plot(fitQ5a$fitted.values, fitQ5a$residuals, pch=20, cex=1.5, col="blue")
abline(c(0,0))
```




```r
qqnorm(fitQ5a$residuals)
qqline(fitQ5a$residuals)
```




```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
shapiro.test(fitQ5a$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitQ5a$residuals
## W = 0.9331, p-value = 0.08234
```

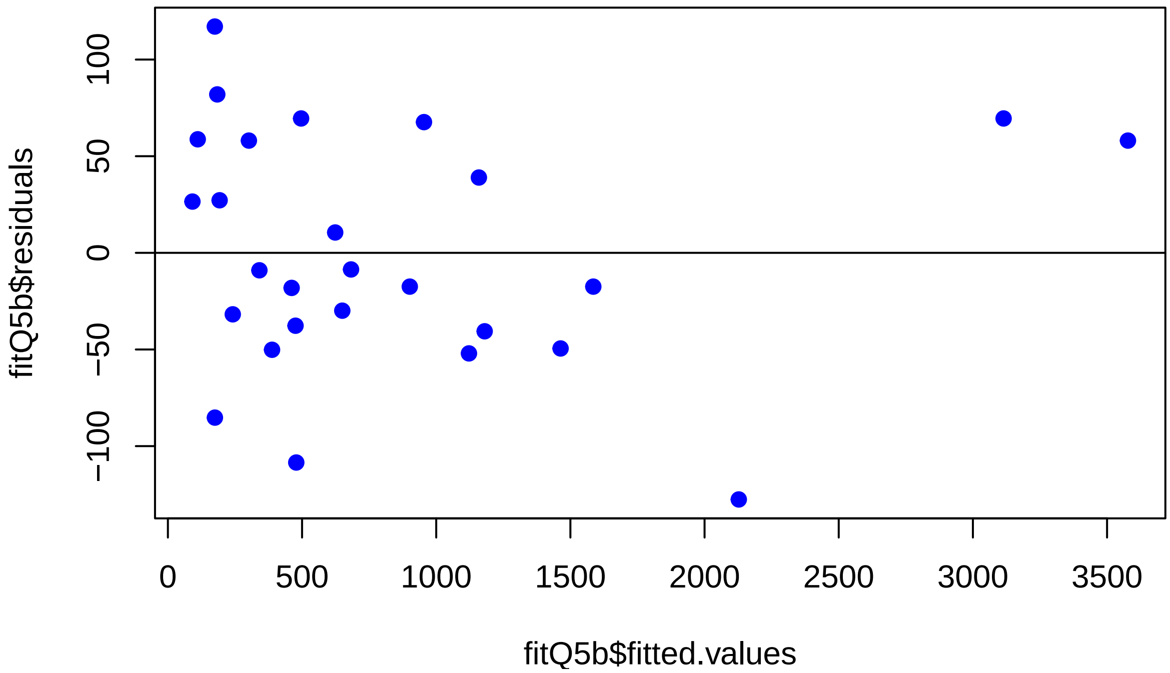
```r
bptest(fitQ5a)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitQ5a
## BP = 9.7675, df = 6, p-value = 0.1348
```

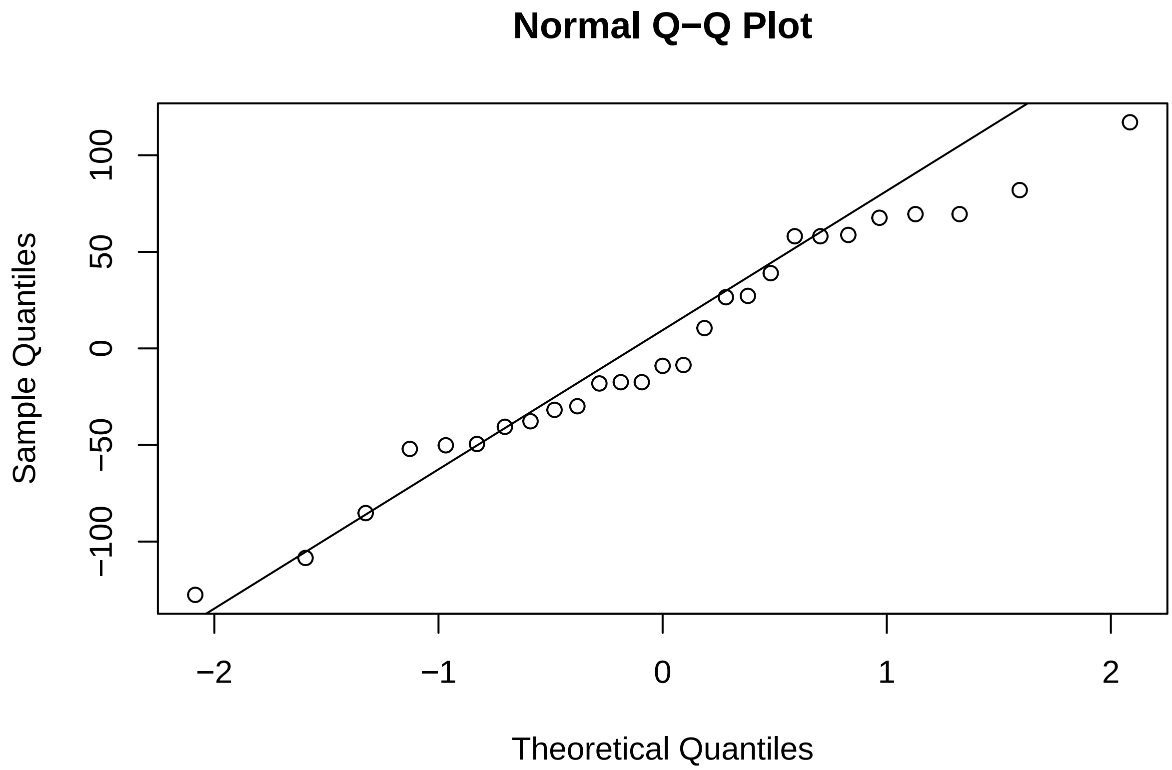
The residual vs fitted value graph looks like a quadratic graph, so linearity assumption is violated. The normal Q-Q plot line is not really close to y=x line, but p-value for SW test is $0.08234 > \alpha = 0.05$, so we fail to reject the null hypothesis that that sample is from normal population. Since p-value for BP test is 0.1348 $> \alpha = 0.05$, so we fail to reject the null hypothesis that the residuals are distributed with equal variance.

(b)

```
fitQ5b = lm(Cycles~factor(Len)+factor(Amp)+factor(Load)+
              factor(Len):factor(Amp) +factor(Len):factor(Load)+
              factor(Amp):factor(Load),data=DataQ5)
plot(fitQ5b$fitted.values, fitQ5b$residuals, pch=20, cex=1.5, col="blue")
abline(c(0,0))
```




```
qqnorm(fitQ5b$residuals)
qqline(fitQ5b$residuals)
```




```
shapiro.test(fitQ5b$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitQ5b$residuals
## W = 0.97117, p-value = 0.6331
```

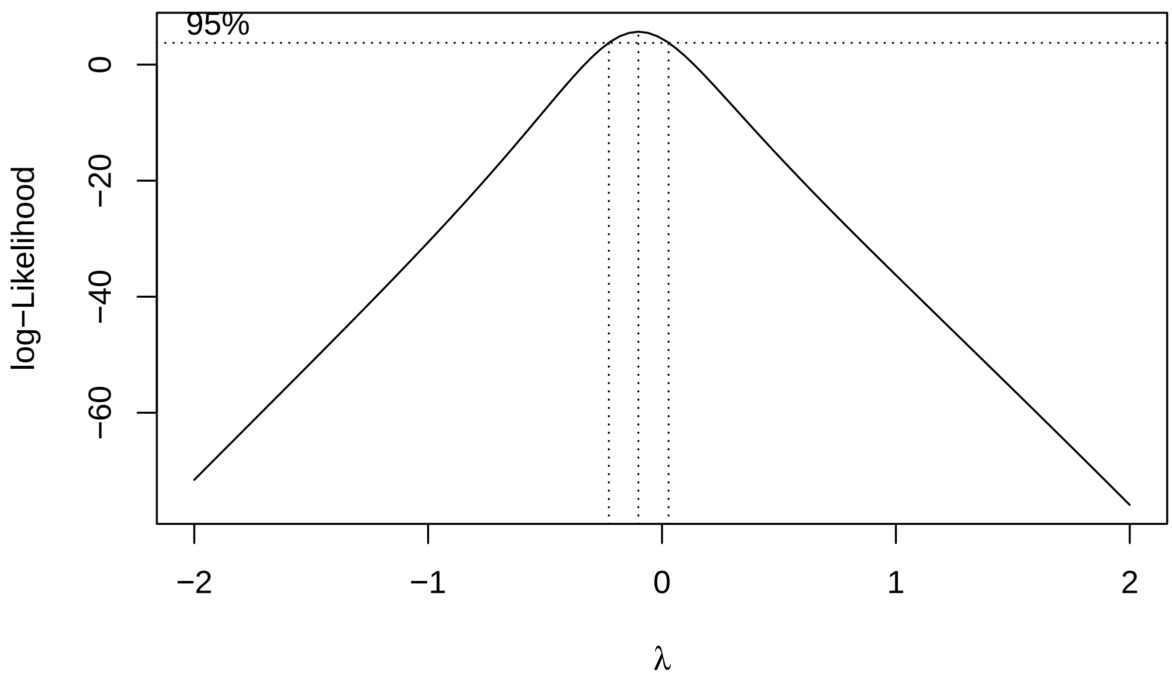
```
bptest(fitQ5b)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitQ5b
## BP = 22.335, df = 18, p-value = 0.2174
```

The residual vs fitted value graph looks like it has more variability when fitted value is small, so equal variance assumption may be violated. The normal Q-Q plot line is not really close to y=x line, but p-value for SW test is $0.6331 > \alpha = 0.05$, so we fail to reject the null hypothesis that that sample is from normal population. Since p-value for BP test is .2174, p-value $> \alpha = 0.05$, so we fail to reject the null hypothesis that the residuals are distributed with equal variance. This model doesn't really fit better than model in (a).

(c)

```
library(MASS)
result = boxcox(fitQ5a)
```



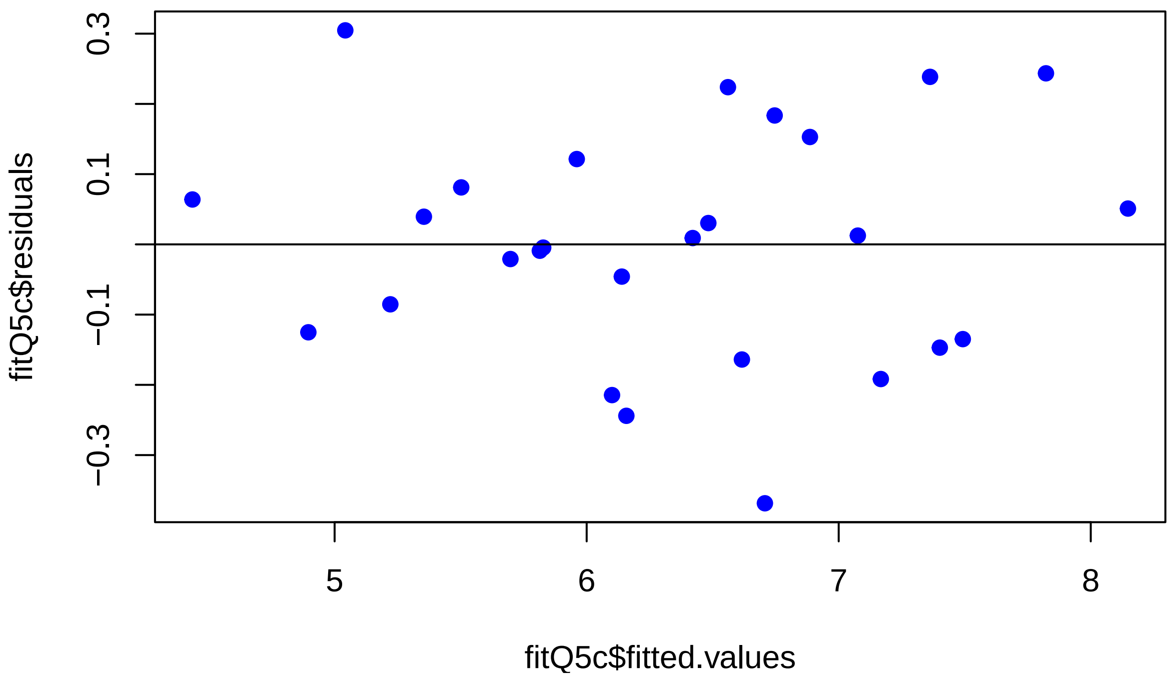

```
mylambda = result$x[which.max(result$y)]
mylambda
```

```
## [1] -0.1010101
```

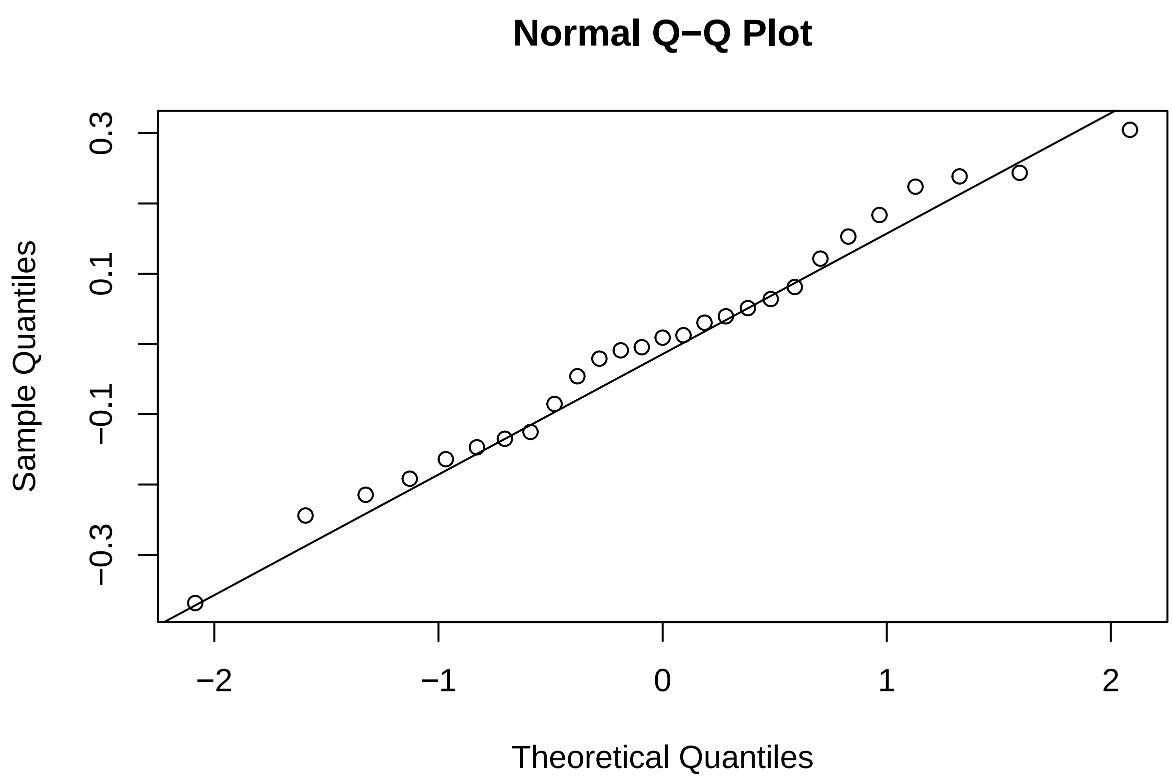
```
Y_star=log(DataQ5$Cycles)
fitQ5c = lm(Y_star~factor(Len)+factor(Amp)+factor(Load),data=DataQ5)
summary(fitQ5c)
```

```
##
## Call:
## lm(formula = Y_star ~ factor(Len) + factor(Amp) + factor(Load),
##     data = DataQ5)
##
## 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 ***
## factor(Len)300   0.91833    0.08928  10.286 1.97e-09 ***
## factor(Len)350   1.66477    0.08928  18.646 4.10e-14 ***
## factor(Amp)9    -0.65521    0.08928  -7.339 4.31e-07 ***
## factor(Amp)10   -1.26173    0.08928 -14.132 7.19e-12 ***
## factor(Load)45 -0.32529    0.08928  -3.643  0.00162 **
## factor(Load)50 -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
plot(fitQ5c$fitted.values, fitQ5c$residuals, pch=20, cex=1.5, col="blue")
abline(c(0,0))
```




```r
qqnorm(fitQ5c$residuals)
qqline(fitQ5c$residuals)
```



```
shapiro.test(fitQ5c$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitQ5c$residuals
## W = 0.98443, p-value = 0.9458
```

```
bptest(fitQ5c)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitQ5c
## BP = 11.995, df = 6, p-value = 0.06208
```

The residual vs fitted value graph now looks like a scatter plot. The normal Q-Q plot line is now closer to y=x line, and p-value for SW test is $0.9458 > \alpha = 0.05$, so we fail to reject the null hypothesis that that sample is from normal population. Since p-value for BP test is $0.06208 > \alpha = 0.05$, so we fail to reject the null hypothesis that the residuals are distributed with equal variance.

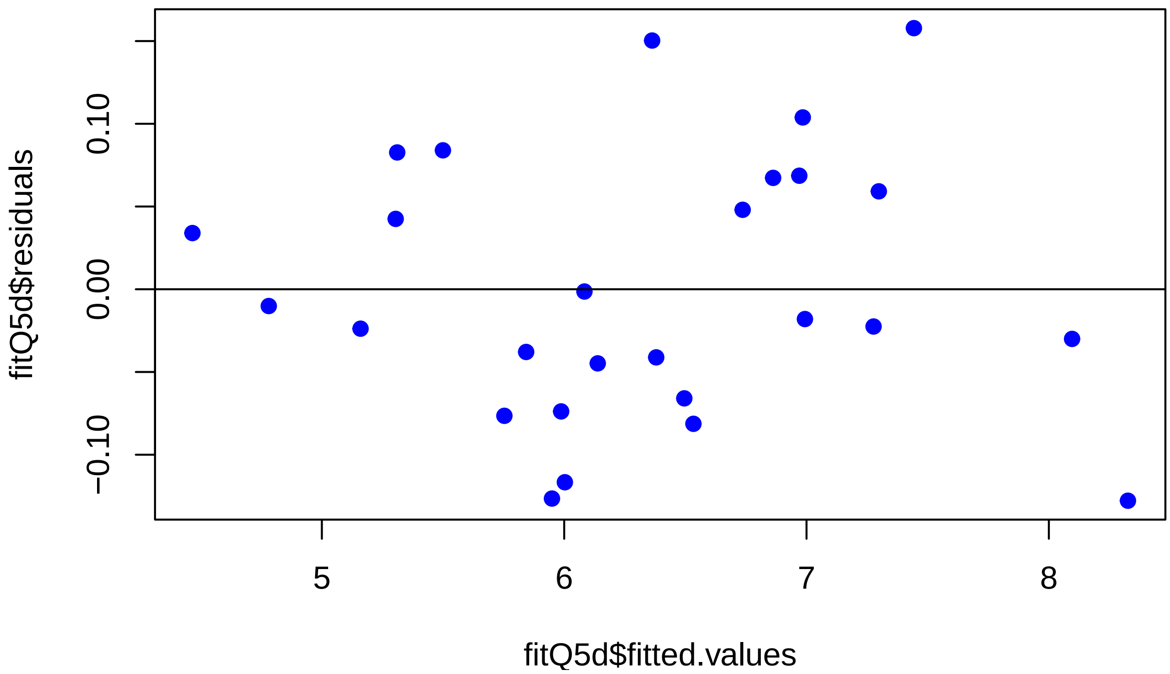
(d)

```
fitQ5d = lm(Y_star~factor(Len)+factor(Amp)+factor(Load)+
              factor(Len):factor(Amp) +factor(Len):factor(Load)+
              factor(Amp):factor(Load),data=DataQ5)
summary(fitQ5d)
```

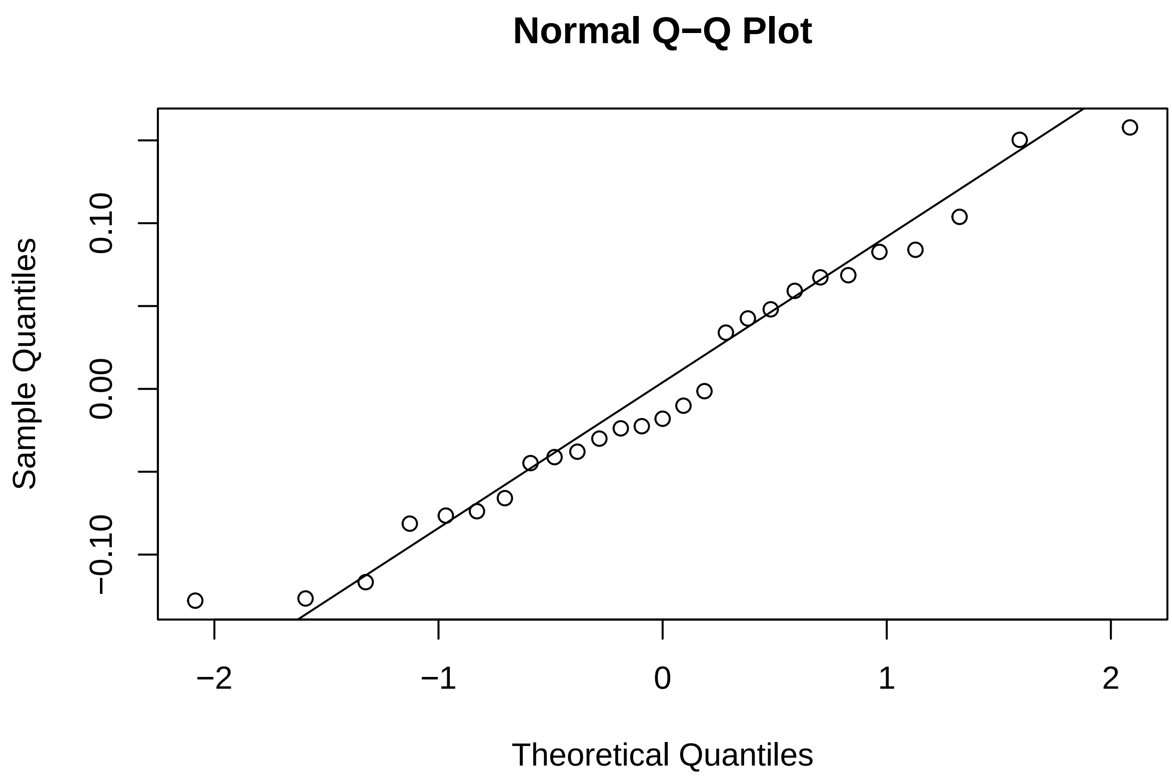
```
##
## Call:
## lm(formula = Y_star ~ factor(Len) + factor(Amp) + factor(Load) +
##     factor(Len):factor(Amp) + factor(Len):factor(Load) + factor(Amp):factor(Load),
##     data = DataQ5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12779 -0.05537 -0.01802  0.06325  0.15780
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.362917   0.120807  52.670 1.87e-11 ***
## factor(Len)300               0.913780   0.151801   6.020 0.000316 ***
## factor(Len)350               1.963516   0.151801  12.935 1.21e-06 ***
## factor(Amp)9                -0.413379   0.151801  -2.723 0.026121 *
## factor(Amp)10               -1.203298   0.151801  -7.927 4.67e-05 ***
## factor(Load)45              -0.375588   0.151801  -2.474 0.038457 *
## factor(Load)50              -0.609676   0.151801  -4.016 0.003861 **
## factor(Len)300:factor(Amp)9  -0.001114   0.166290  -0.007 0.994817
## factor(Len)350:factor(Amp)9  -0.614678   0.166290  -3.696 0.006074 **
## factor(Len)300:factor(Amp)10  0.064964   0.166290   0.391 0.706242
## factor(Len)350:factor(Amp)10 -0.152966   0.166290  -0.920 0.384537
## factor(Len)300:factor(Load)45  0.083463   0.166290   0.502 0.629248
## factor(Len)350:factor(Load)45  0.145059   0.166290   0.872 0.408448
## factor(Len)300:factor(Load)50 -0.133655   0.166290  -0.804 0.444766
## factor(Len)350:factor(Load)50 -0.273658   0.166290  -1.646 0.138450
```

```
## factor(Amp)9:factor(Load)45   -0.074416   0.166290  -0.448 0.666379
## factor(Amp)10:factor(Load)45  -0.003211   0.166290  -0.019 0.985067
## factor(Amp)9:factor(Load)50   -0.035285   0.166290  -0.212 0.837264
## factor(Amp)10:factor(Load)50  -0.084089   0.166290  -0.506 0.626717
## ---
## 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
```

```
plot(fitQ5d$fitted.values, fitQ5d$residuals, pch=20, cex=1.5, col="blue")
abline(c(0,0))
```




```
qqnorm(fitQ5d$residuals)
qqline(fitQ5d$residuals)
```




```r
shapiro.test(fitQ5d$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fitQ5d$residuals
## W = 0.96517, p-value = 0.4806
```

```r
bptest(fitQ5d)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitQ5d
## BP = 22.955, df = 18, p-value = 0.1923
```

The residual vs fitted value graph looks like a scatter plot. The normal Q-Q plot line is close to y=x line, and p-value for SW test is $0.4806 > \alpha = 0.05$, so we fail to reject the null hypothesis that that sample is from normal population. Since p-value for BP test is 0.1923, p-value $> \alpha = 0.05$, so we fail to reject the null hypothesis that the residuals are distributed with equal variance. We get the same conclusion from this model compare to model in (c), and compare $R^2 = 0.9691$ from model in (c) is 0.9691 to $R^2 = 0.9928$ from model here, there is no significant improvement on R^2, so this model is no better than the model with main effects only.