CSCI E-106:Fall 2020 Midterm Solutions

Instructions

1. Open book and open notes exam (textbooks (print or pdf), lecture slides, notes, practice exam, homework solutions, and TA slides, including all Rmd's*).
2. You are allowed to use RStudio Cloud (https://rstudio.cloud.), Microsoft Word, Power Point and PDF reader, and canvas on your laptop.
3. You need to have a camera on your laptop. Proctorio is required to start this exam. If you are prompted for an access code, you must Configure Proctorio on your machine.
4. Please read the list of recording and restrictions provided by Proctorio carefully before taking the exam
5. Please pay attention any timing and technical warnings that popped up your screen
6. The exam will be available from Friday October 23rd at 12 pm EST through Monday October 26th at 7:20 pm EST.
7. Once you start the exam, you have to complete the exam in 3 hours or by Monday October 26th at 7:20 pm EST, whichever comes first.
8. In order to receive full credit, please provide full explanations and calculations for each question
9. Make sure that you are familiar with the procedures for troubleshooting exam issues preview the document. Follow the protocol if there are any issues!

Problem 1

Refer to the Midterm Q1 Data set. (50 Points)

a-) Create development sample and hold out sample. Development sample is a random sample of 70% of the data and hold out sample is the remainder 30% of the data. Use set.seed(1023) to select the samples. (5 pts)

There are 496 observations, there will be 347 observations in the development sample and 149 observations on the holdout sample. Y and X are moderatly correlated.

```
Q1.DS <- read.csv("/cloud/project/Midterm Q1 Data Set.csv")
str(Q1.DS)
```

```
## 'data.frame':    496 obs. of  2 variables:
##  $ y: int  2300 2500 4000 500 4400 1900 3600 2200 2300 2200 ...
##  $ x: int  35 60 30 20 50 18 25 21 13 30 ...
```

```
cor(Q1.DS)
```

```
##           y         x
## y 1.0000000 0.4177872
## x 0.4177872 1.0000000
```

```
set.seed(1023)
IND<-sample(c(1:496),round(496*0.7))
dev.samp<-Q1.DS[IND,]
hold.out<-Q1.DS[-c(IND),]
```

b-) Build a regression model to predict Y as a function of X on the development sample. Write down the regression model, Is the regression model significant? (5 points)

the regression model is significant. Y = 1160.37 + 49.07 X. R^2 is 20% and the model is significant.

```
f1<-lm(y~x,data=dev.samp)
summary(f1)
```

```
##
## Call:
## lm(formula = y ~ x, data = dev.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2514.0  -888.1  -241.8   577.1  6731.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1160.37     141.96   8.174 5.74e-15 ***
## x              49.07       5.35   9.172  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1319 on 345 degrees of freedom
## Multiple R-squared:  0.1961, Adjusted R-squared:  0.1937
## F-statistic: 84.13 on 1 and 345 DF,  p-value: < 2.2e-16
```

c-) Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? Conduct the Breusch-Pagan Test to determine whether or not the error variances are constant . (10 points)

QQ plot indicates the data is skewed right. There could be outliers in the data set. Further examination is needed. Error vs. Fitted values indicate that error variances are not equal.

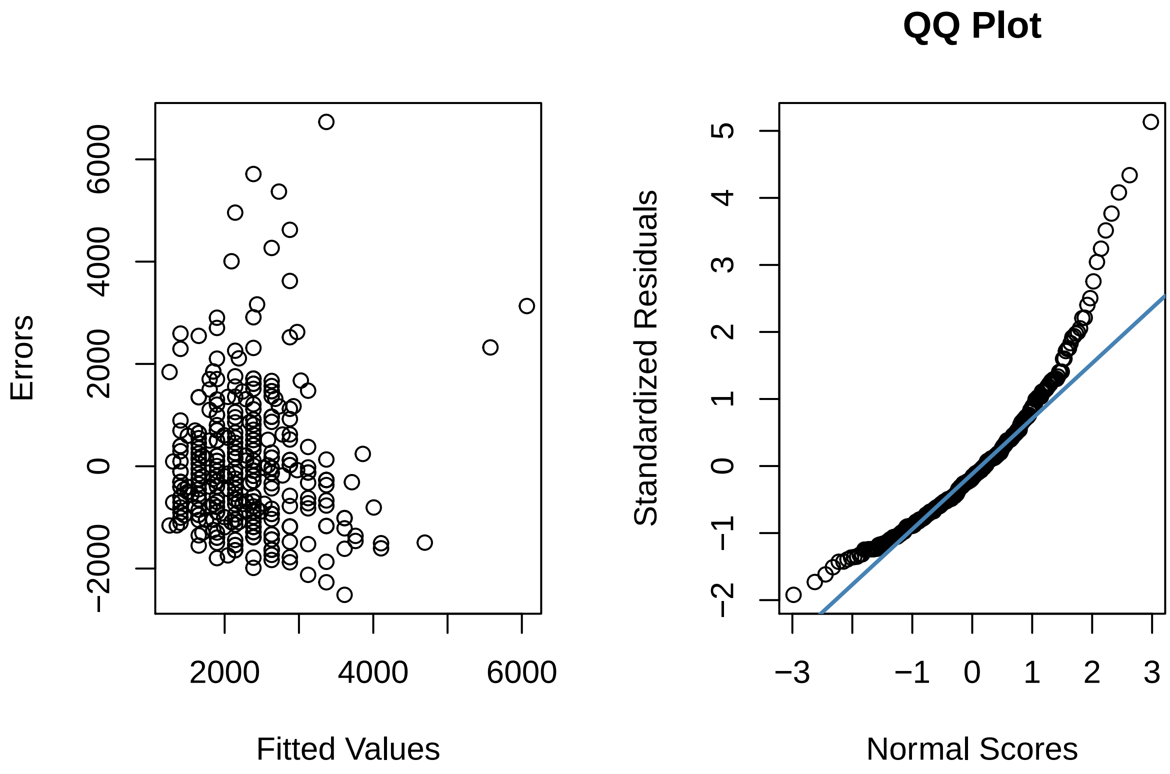
The Breusch-Pagan Test:

H_{0}:$\gamma_1 = 0$ H_{a}:$\gamma_1 \neq 0$

p value is <0.05. H_{0} is rejected, the error variances are not equal.

```
ei<-f1$residuals
yhat<-f1$fitted.values
par(mfrow=c(1,2))
```

```r
plot(yhat,ei,ylab="Errors",xlab="Fitted Values")
stdei<- rstandard(f1)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```



```r
anova(f1)
```

```
## Analysis of Variance Table
##
## Response: y
##              Df    Sum Sq   Mean Sq F value    Pr(>F)
## x             1 146277183 146277183  84.135 < 2.2e-16 ***
## Residuals 345 599821203   1738612
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ei2<-ei^2
g<-lm(ei2~dev.samp$x)
anova(g)
```

```
## Analysis of Variance Table
##
## Response: ei2
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## dev.samp$x   1 1.9435e+14 1.9435e+14  11.404 0.0008166 ***
## Residuals  345 5.8797e+15 1.7042e+13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSR=  194348752489180
SSE=  599821203
Chi.Square=(SSR / 2) / ((SSE/347)^2)
1-pchisq(Chi.Square,1)
```

```
## [1] 1.17897e-08
```

d-) Calculate the simultaneous 99% confidence interval for β_0,and β_1 and calculate the simultaneous 95% confidence intervals for the predicted new X values for 85 and 90. (10 pts)

The simultaneous 99% confidence interval for β_0,and β_1 are $759.31 \leq \beta_0 \leq 1561.42$ $33.96 \leq \beta_1 \leq 64.19$

The simultaneous prediction interval for the new X values for 85 and 90 are below in the r code.

```
confint(f1,level=1-0.01/2)
```

```
##                 0.25 %    99.75 %
## (Intercept) 759.30824 1561.42581
## x            33.95818   64.18807
```

```
pred<-predict.lm(f1,data.frame(x<-c(85,90)),se.fit = TRUE)
fit<-pred$fit
fit
```

```
##        1        2
## 5331.583 5576.949
```

```
s.pred<-sqrt(pred$se.fit^2+pred$residual.scale^2)
S=sqrt(2*qf(0.95,2,345))
B=qt(1-0.05/(2*2),345)
cbind(B,S)
```

```
##             B        S
## [1,] 2.251228 2.458413
```

```
cbind(pred$fit-B*s.pred,pred$fit+B*s.pred)
```

```
##       [,1]     [,2]
## 1 2266.564 8396.602
## 2 2496.704 8657.194
```

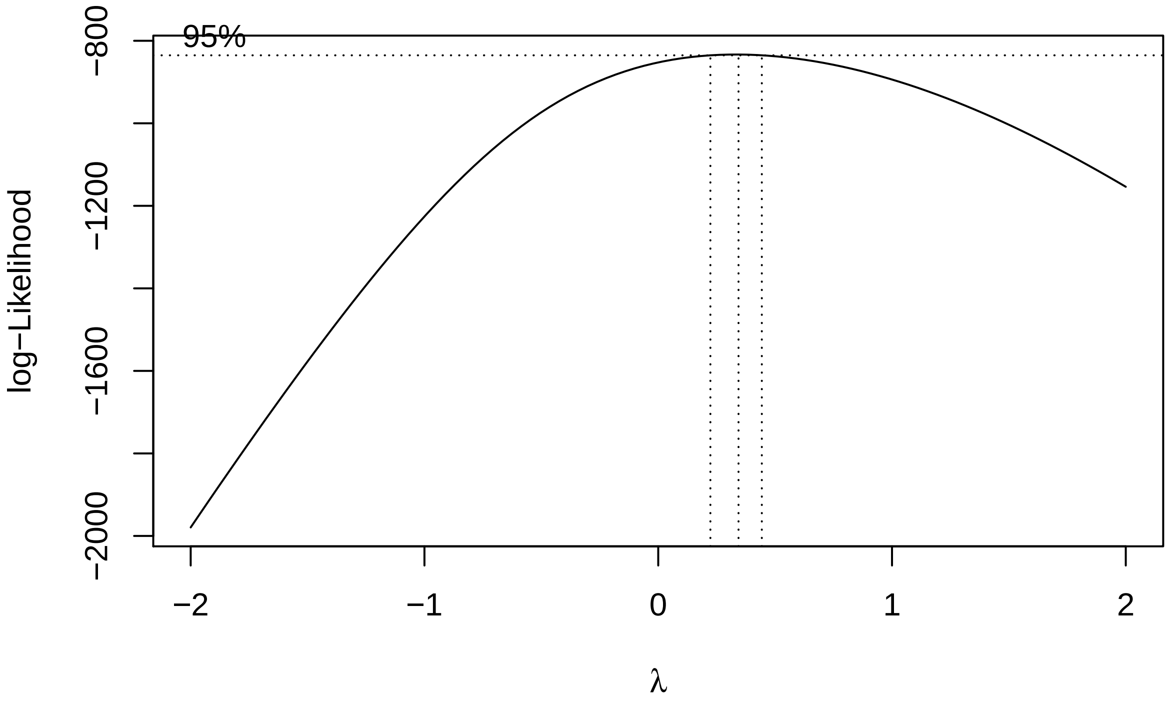
e-) Use the Box-Cox procedure to find an appropriate power transformation and perform the transformation. Obtain, the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show? (10 pts)

It looks like λ is 0.34. However, you could also use either λ 0.5 (square root transformation) or λ is 0 (log transformation) to make the transformation more easier. I used λ is 0.34.

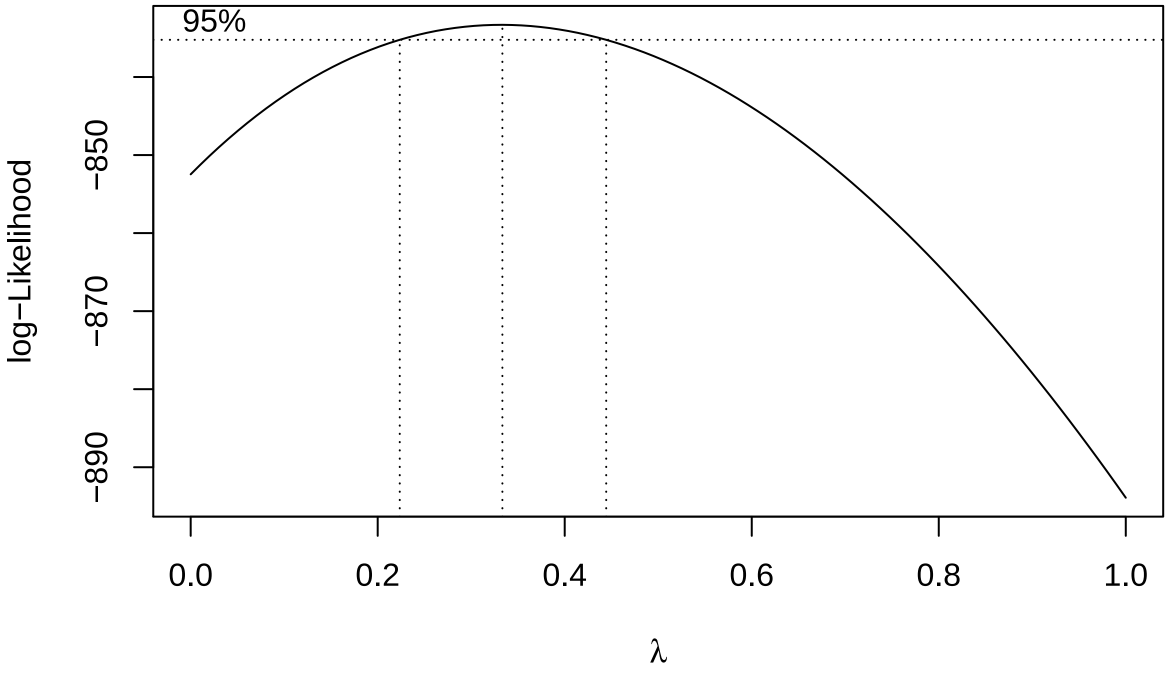
the regression model is significant. $Y^{0.34} = 11.03 + 0.098$ X. R^2 is 20% and the model is significant.

QQ plots look normal. However, error vs predicted graph still shows a V shape. Unequal variances still persists.

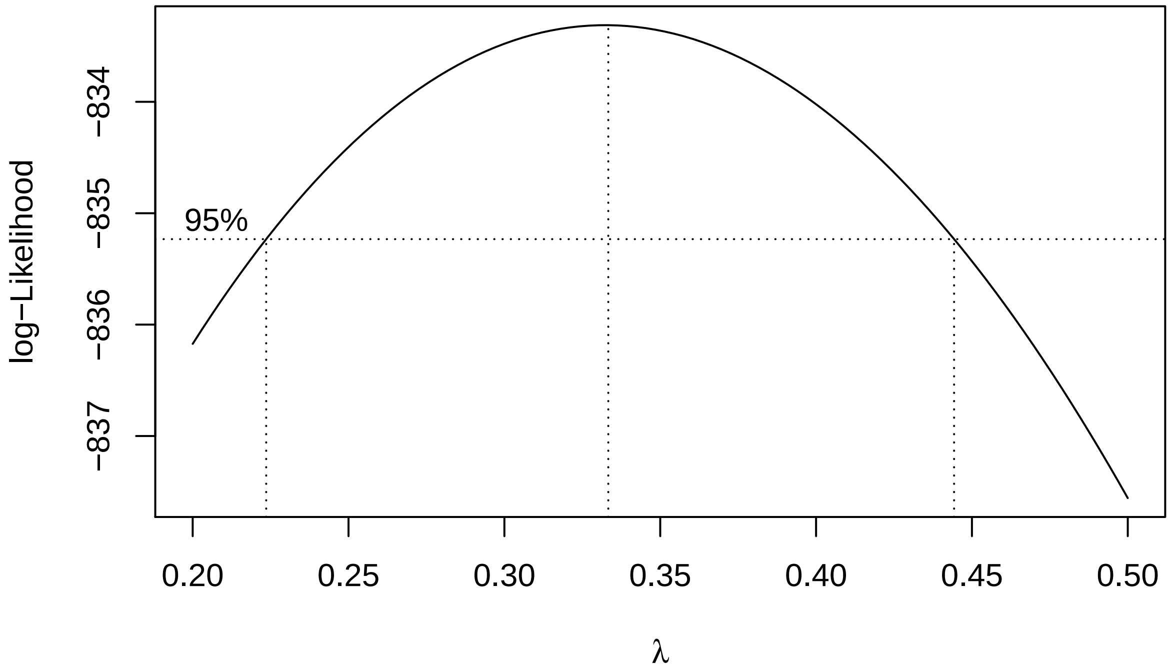
```
library(MASS)
par(mfrow=c(1,1))
boxcox(f1,lambda=seq(-2,2,by=0.1))
```



```
boxcox(f1,lambda=seq(0.0,1,by=0.1))
```



```
boxcox(f1,lambda=seq(0.2,0.5,by=0.1))
```



```
f1.2<-lm(y^0.34~x,data=dev.samp)
summary(f1.2)
```

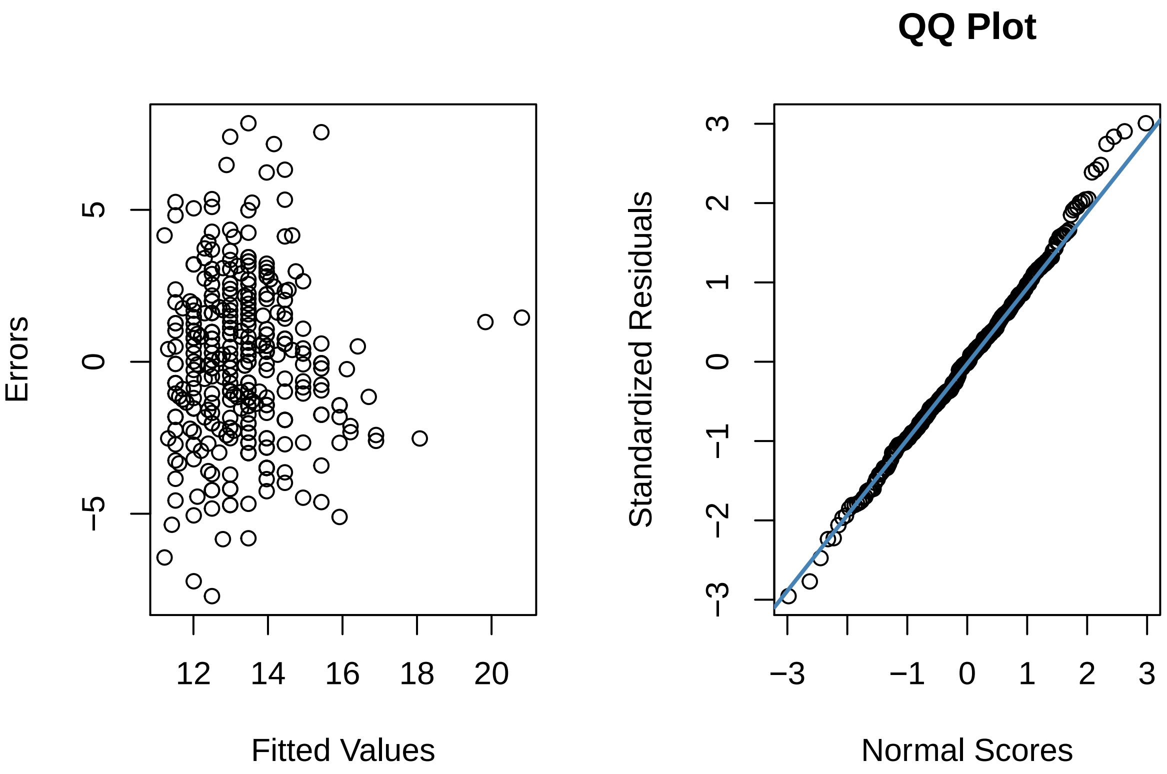
```
##
## Call:
## lm(formula = y^0.34 ~ x, data = dev.samp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7107 -1.7419 -0.0488  1.6105  7.8492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.02928    0.28149  39.182   <2e-16 ***
## x            0.09785    0.01061   9.224   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.615 on 345 degrees of freedom
## Multiple R-squared:  0.1978, Adjusted R-squared:  0.1955
## F-statistic: 85.08 on 1 and 345 DF,  p-value: < 2.2e-16
```

```
ei<-f1.2$residuals
yhat<-f1.2$fitted.values
par(mfrow=c(1,2))
plot(yhat,ei,ylab="Errors",xlab="Fitted Values")
ei2<-ei^2
g<-lm(ei2~dev.samp$x)
summary(g)
```

```
##
```

```
## Call:
## lm(formula = ei2 ~ dev.samp$x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.287 -6.046 -3.977  1.422 54.869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.42981    1.09495   6.786 5.05e-11 ***
## dev.samp$x  -0.02753    0.04127  -0.667    0.505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.17 on 345 degrees of freedom
## Multiple R-squared:  0.001288,   Adjusted R-squared:  -0.001607
## F-statistic: 0.4449 on 1 and 345 DF,  p-value: 0.5052
```

```r
stdei<- rstandard(f1.2)
qqnorm(stdei,ylab="Standardized Residuals",xlab="Normal Scores", main="QQ Plot")
qqline(stdei,col = "steelblue", lwd = 2)
```



f-) Calculate R Square on the hold out sample (hint: calculate SSE, SSR and SST on the hold out sample). Is the model performance robust? (10 pts)

R^2 on the hold out sample is 9.2%. Whereas, the R^2 on the development sample is 20%. The model performance was significantly worsened on the hold out sample. The model is not robust.

```
SST<-var(hold.out$y)*(length(hold.out$y)-1)
pred<-predict(f1.2,hold.out)
ei<-hold.out$y-(pred^(1/0.34))
SSE<-sum(ei^2)
R.SQ=1-(SSE/SST)
R.SQ
```

```
## [1] 0.09229792
```

Problem 2

Refer to the Midterm Q2 Data set (30 Points)

Perform one factor analysis by finding the best variable to explain Y. Fit one variable regression model with Y as a dependent variable against remaining variables, as an independent variable one at a time. Choose the variable with highest R^2 that explains Y and comment on the QQ plot and error vs. fitted values graph for the model assumptions.

X_4, X_5, and X_4, X_6 are NOT significant. X_1, X_2, and X_3, X_6 are significant. X_1 is the best variable. The R^2s are below.

model with x1: R^2 is 94% model with x2: R^2 is 38% model with x3: R^2 is 54% model with x4: R^2 is 0% model with x5: R^2 is 0.02% model with x6: R^2 is 0%

QQ plot indicates heavy tails, Residual vs. Fitted graph shows an evidence of unequal variances.

```
Q2.DS <- read.csv("/cloud/project/Midterm Q2 Data Set.csv")
f1<-lm(Y~X1,data=Q2.DS)
f2<-lm(Y~X2,data=Q2.DS)
f3<-lm(Y~X3,data=Q2.DS)
f4<-lm(Y~X4,data=Q2.DS)
f5<-lm(Y~X5,data=Q2.DS)
f6<-lm(Y~X6,data=Q2.DS)
summary(f1)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = Q2.DS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.791 -15.894   3.507  12.088  78.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.0480     7.1138  -2.959  0.00622 **
## X1            4.0721     0.1899  21.443  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.95 on 28 degrees of freedom
## Multiple R-squared:  0.9426, Adjusted R-squared:  0.9406
## F-statistic: 459.8 on 1 and 28 DF,  p-value: < 2.2e-16
```

```r
summary(f2)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = Q2.DS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.495 -53.431 -29.045   3.423 306.137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 63.78286   17.52442   3.640 0.001094 **
## X2           0.08196    0.01971   4.158 0.000275 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.73 on 28 degrees of freedom
## Multiple R-squared:  0.3817, Adjusted R-squared:  0.3596
## F-statistic: 17.29 on 1 and 28 DF,  p-value: 0.0002748
```

```r
summary(f3)
```

```
##
## Call:
## lm(formula = Y ~ X3, data = Q2.DS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.319  -30.721  -14.690    4.634  259.180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.33511   19.20529   0.590     0.56
## X3           0.20079    0.03465   5.795 3.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.66 on 28 degrees of freedom
## Multiple R-squared:  0.5454, Adjusted R-squared:  0.5291
## F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

```r
summary(f4)
```

```
##
## Call:
## lm(formula = Y ~ X4, data = Q2.DS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84.16 -71.78 -41.66   9.84 357.70
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.3720    26.2035   3.296  0.00267 **
## X4           -0.1132     1.5175  -0.075  0.94107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 116.7 on 28 degrees of freedom
## Multiple R-squared:  0.0001986,  Adjusted R-squared:  -0.03551
## F-statistic: 0.005563 on 1 and 28 DF,  p-value: 0.9411
```

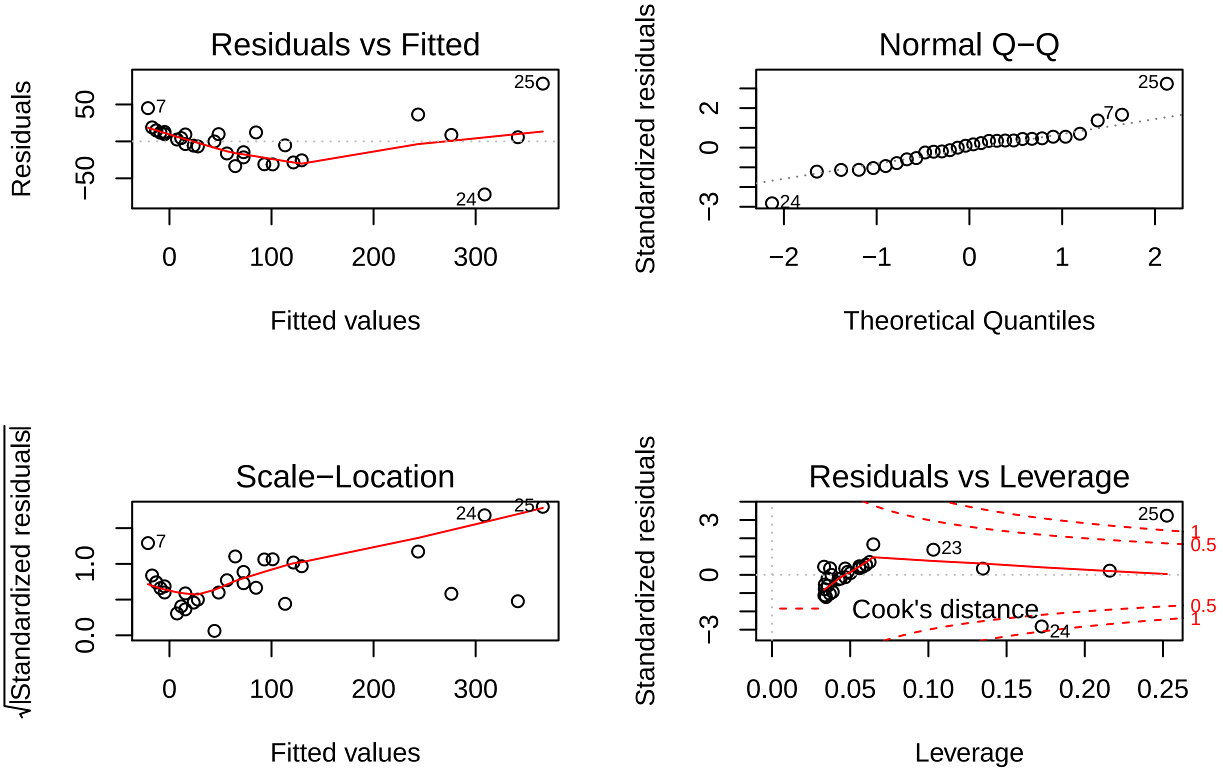
```r
summary(f5)
```

```
##
## Call:
## lm(formula = Y ~ X5, data = Q2.DS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -99.12 -77.60 -34.99  16.47 342.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.6638    27.5665   3.688 0.000964 ***
## X5           -0.2884     0.3137  -0.919 0.365862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.9 on 28 degrees of freedom
## Multiple R-squared:  0.02929,    Adjusted R-squared:  -0.005378
## F-statistic: 0.8449 on 1 and 28 DF,  p-value: 0.3659
```

```r
summary(f6)
```

```
##
## Call:
## lm(formula = Y ~ X6, data = Q2.DS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -85.46 -71.41 -42.60   7.62 359.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 84.32743   22.27411   3.786 0.000744 ***
## X6           0.00347    0.02505   0.139 0.890831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 116.6 on 28 degrees of freedom
## Multiple R-squared:  0.0006847,  Adjusted R-squared:  -0.03501
## F-statistic: 0.01918 on 1 and 28 DF,  p-value: 0.8908
```

```r
par(mfrow=c(2,2))
plot(f1)
```



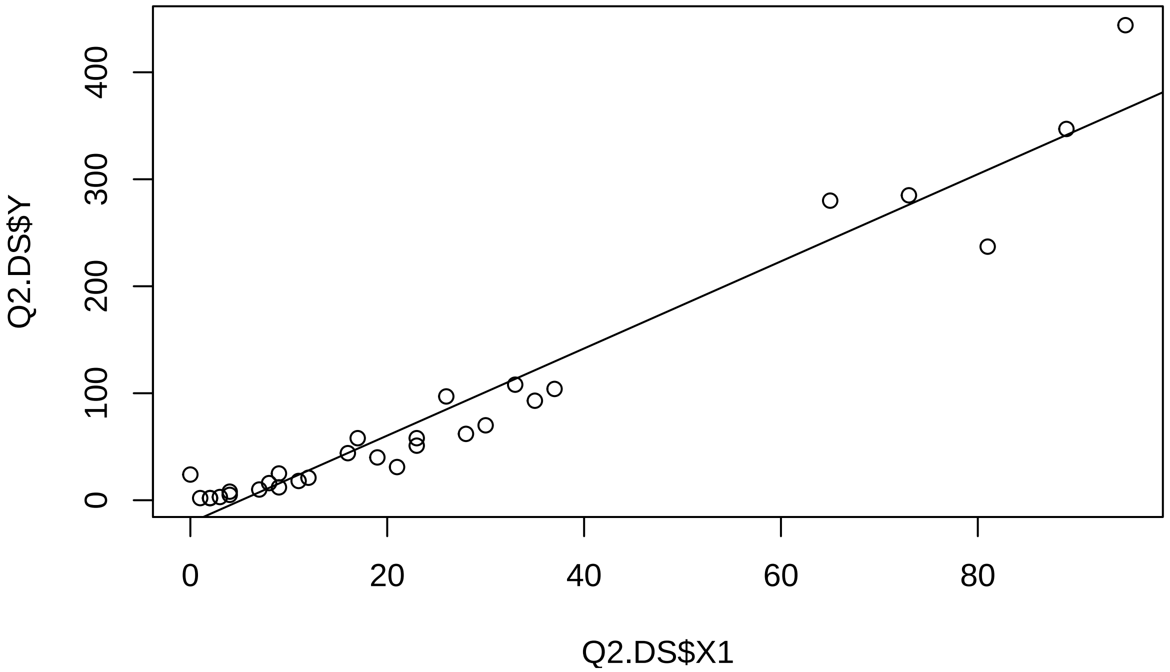
Problem 3

Based on your final model selected on problem 2.

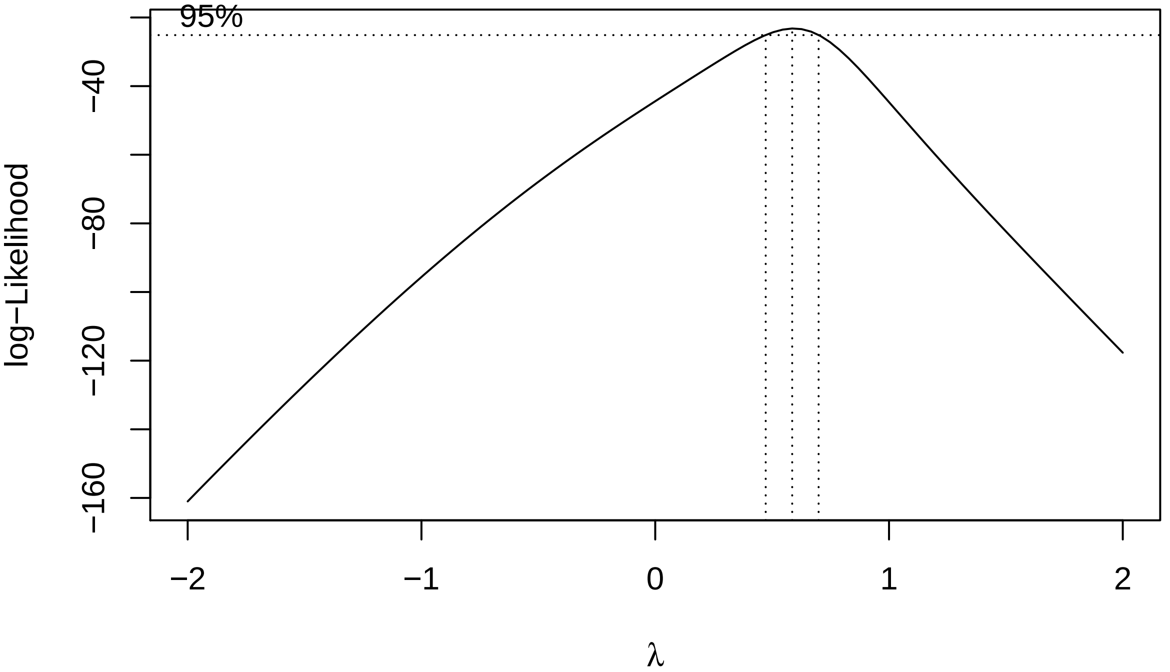
a-) is the linear fit appropriate? If not, transform the data and find an appropriate fit. Comment on the model and regression model assumptions. (10 points)

The linear fit is appropriate, however Box-Cox transformation suggest the square root of transformation of Y, which is a good solution for heavy tail distribution. After the transformation, QQ plots looks normal and there is no further evidence of unequal variances exist. The new model is $\sqrt{Y} = 2.5 + 0.19X_1$ and R^2 is 95% and the model is signficant, the graphs indicate that all regression model assumptionsn are met.

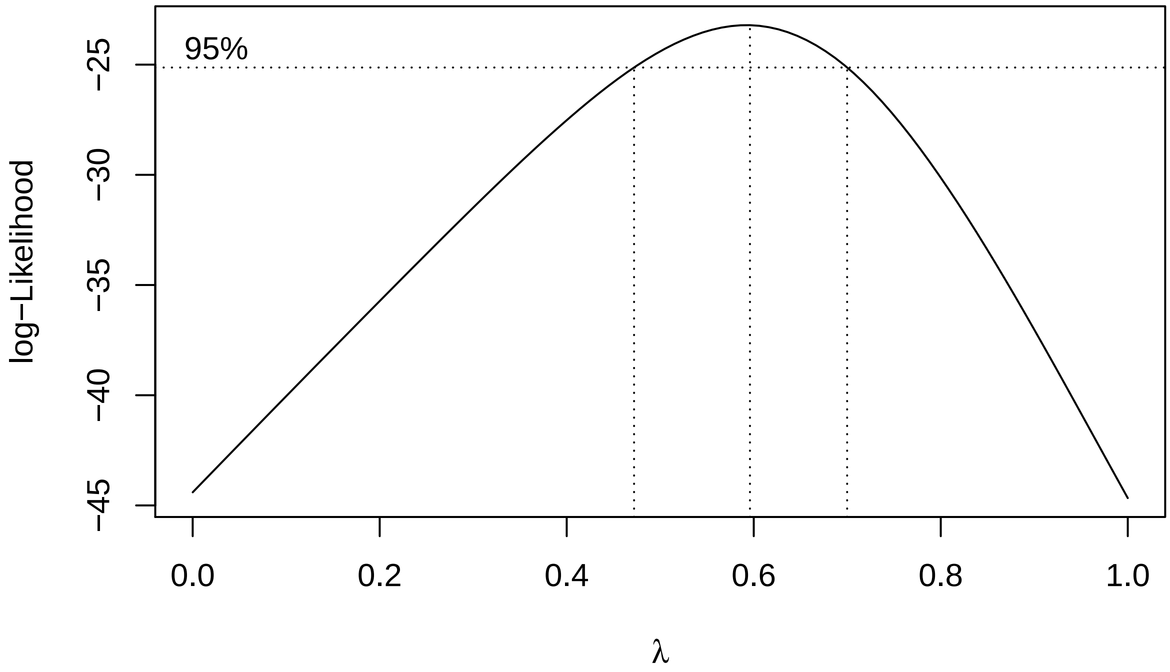
```r
par(mfrow=c(1,1))
plot(Q2.DS$X1,Q2.DS$Y)
abline(f1)
```



```r
boxcox(f1,lambda=seq(-2,2,by=0.1))
```



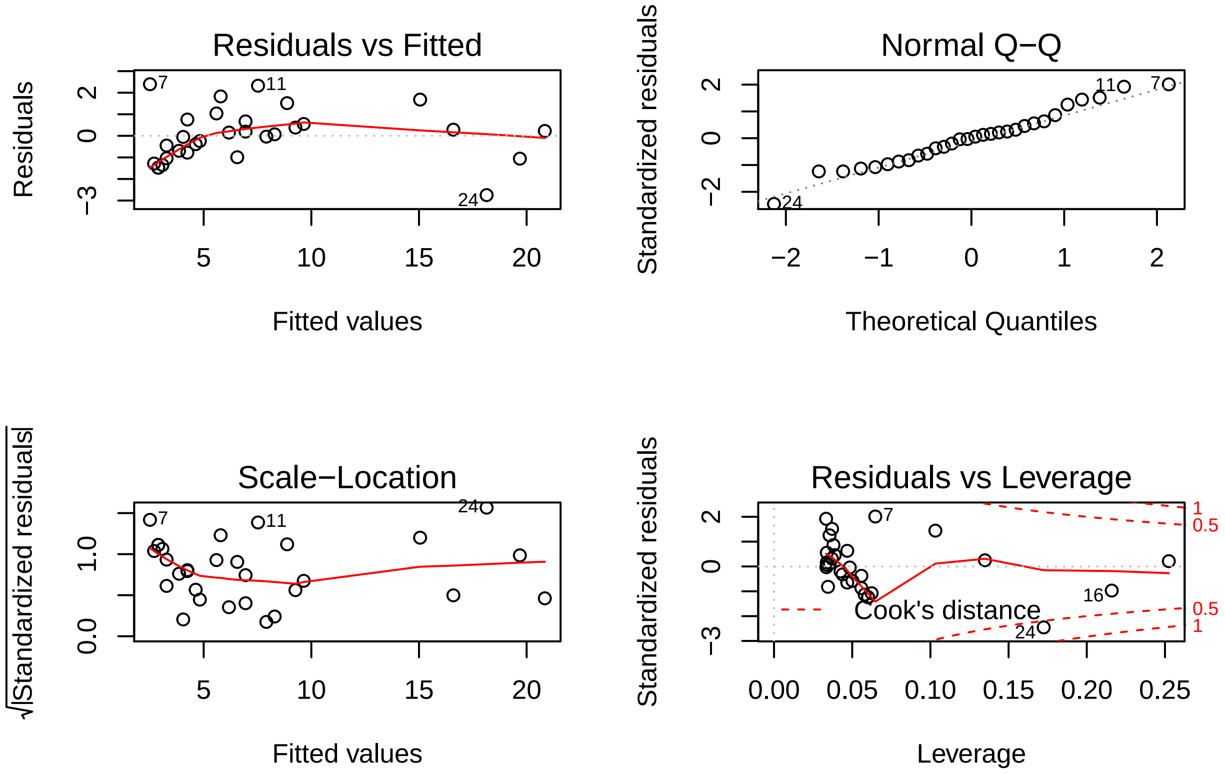
```r
boxcox(f1,lambda=seq(0.0,1,by=0.1))
```



```
f2.1<-lm(sqrt(Y)~X1,data=Q2.DS)
summary(f2.1)
```

```
##
## Call:
## lm(formula = sqrt(Y) ~ X1, data = Q2.DS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74682 -0.93860  0.01647  0.64009  2.39315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.505832   0.313294   7.998 1.04e-08 ***
## X1          0.193034   0.008363  23.081  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.231 on 28 degrees of freedom
## Multiple R-squared:  0.9501, Adjusted R-squared:  0.9483
## F-statistic: 532.7 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(f2.1)
```



b-) Predict Y when X_1=33,X_2=18,X_3=450,X_4=11,X_5=12,X_6=0.05 , and calculate the 99% confidence interval (10 points). Y is predicted to be 79 and 99% confidence interval is $68 \leq \hat{(}Y) \leq 91$. ‘

```
pred<-predict.lm(f2.1,data.frame(X1 = 33),interval =  "confidence",level=0.99)
pred^2
```

```
##        fit      lwr      upr
## 1 78.78284 67.81224 90.57558
```