Lab4_Transformations

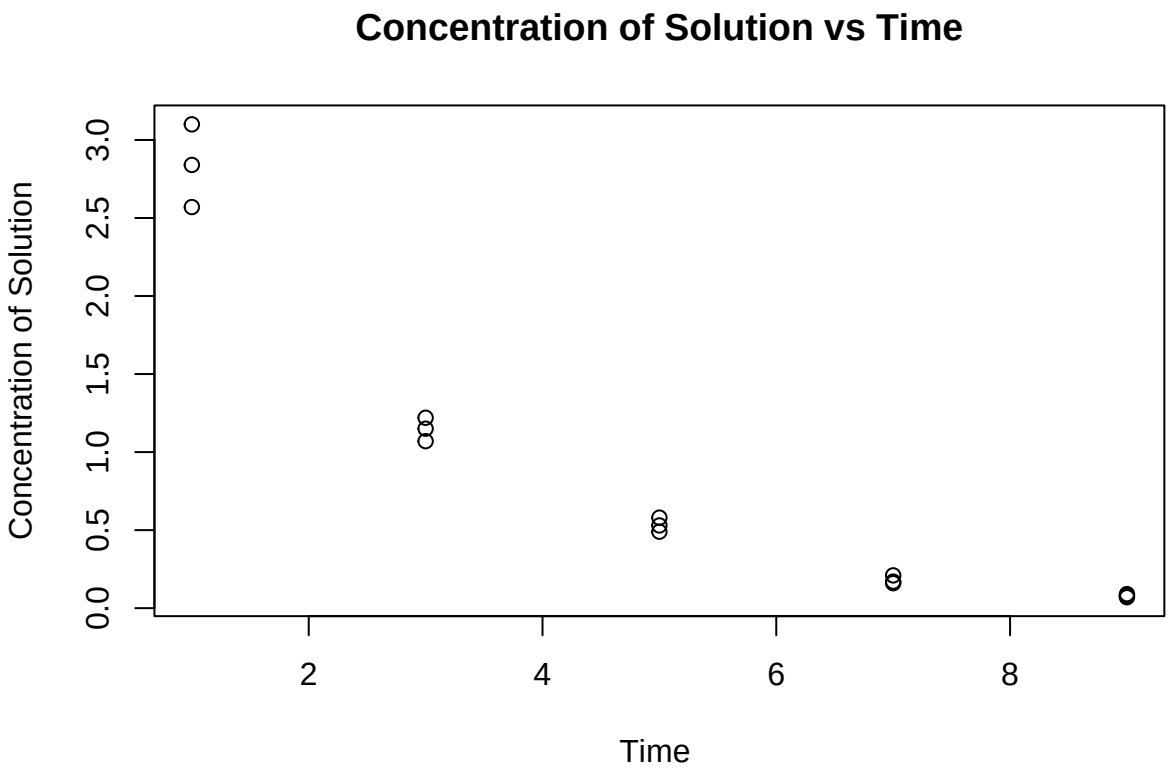
Nahom Ayele & Nick Huo

2022-09-30

```
# Import Data set
library(readr)
SolutionConcentration <- read_csv("SolutionConcentration.csv", show_col_types = FALSE)
```

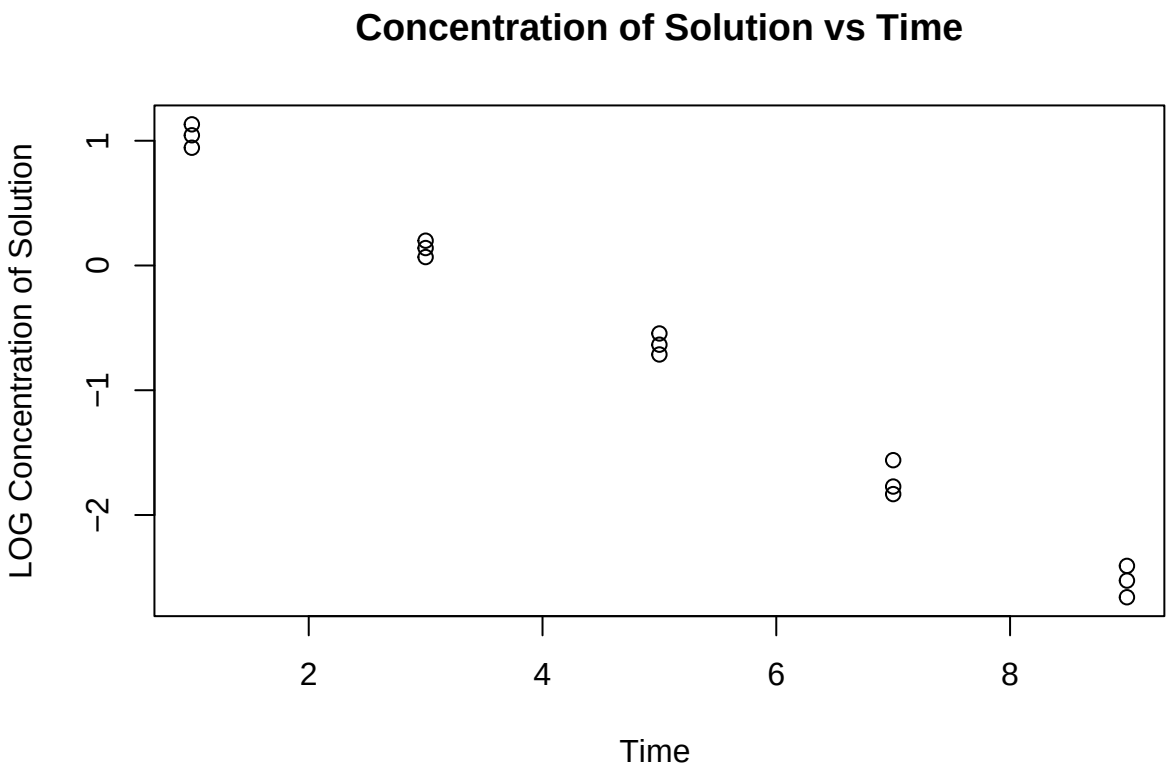
Question 1: Did you remember to use information labels and an informative title? What do you observe? What kind of a transformation do you think would be most appropriate? Why do you suggest this type?

```
# Set variables
xi_q1 <- SolutionConcentration$x
yi_q1<- SolutionConcentration$y


plot(xi_q1,yi_q1, xlab = "Time ", ylab = "Concentration of Solution", main = "Concentration of Solution
```

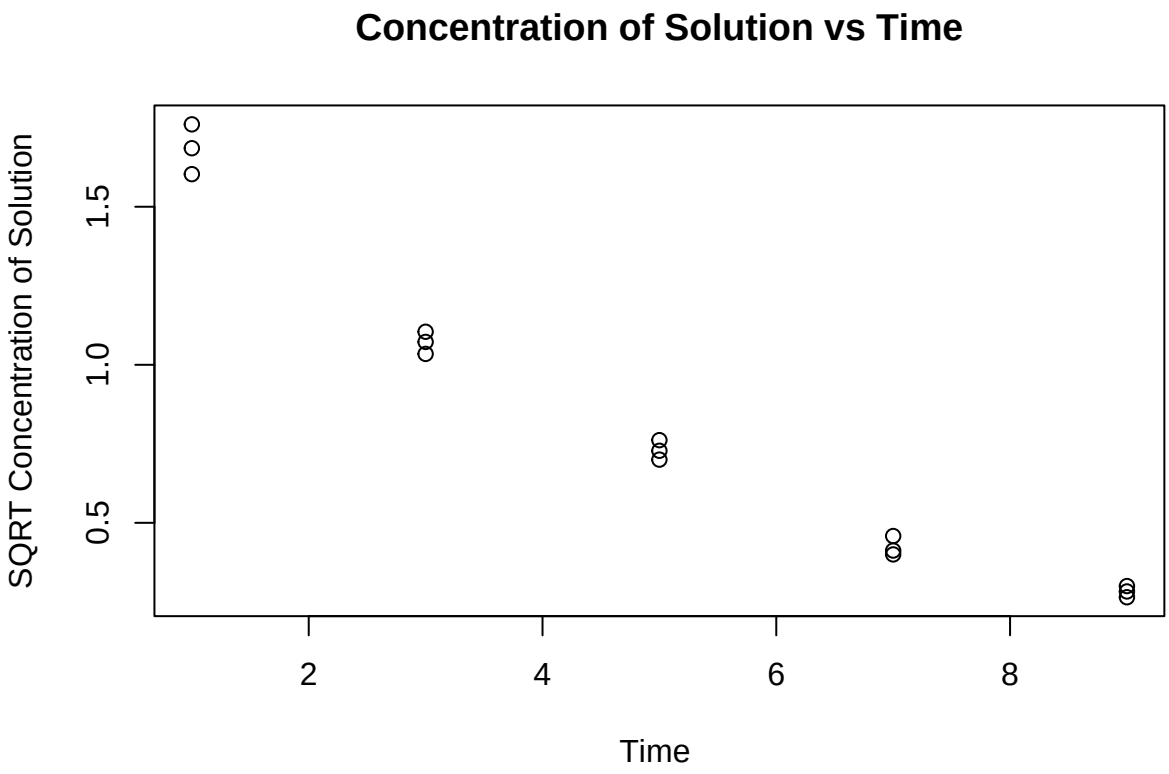


We appear to be looking at an exponential decay graph. We see that we have really high value for y and we want to transform this to a more linear model. So we suggest transforming this to a log.

```
## plotting Log transform
plot(SolutionConcentration$x,log(SolutionConcentration$y), xlab = "Time ", ylab = "LOG Concentration of
```

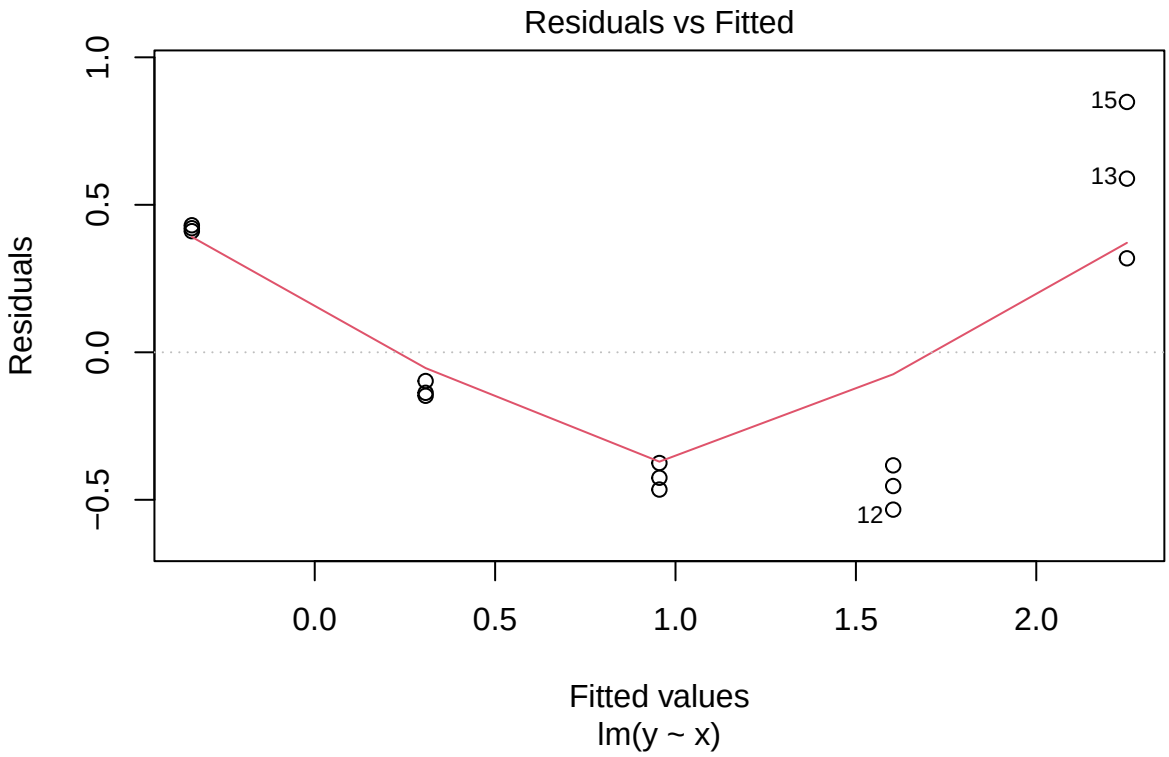


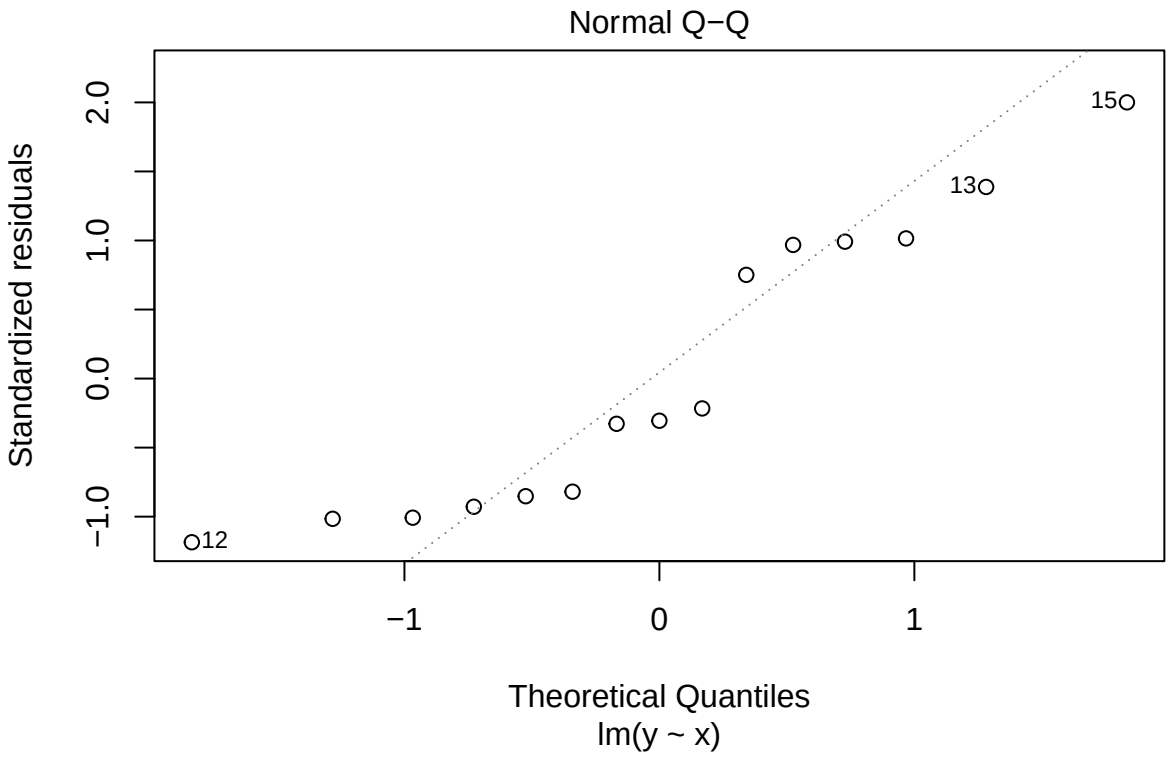

```
## plotting Square root transform
plot(SolutionConcentration$x,sqrt(SolutionConcentration$y), xlab = "Time ", ylab = "SQRT Concentration
```



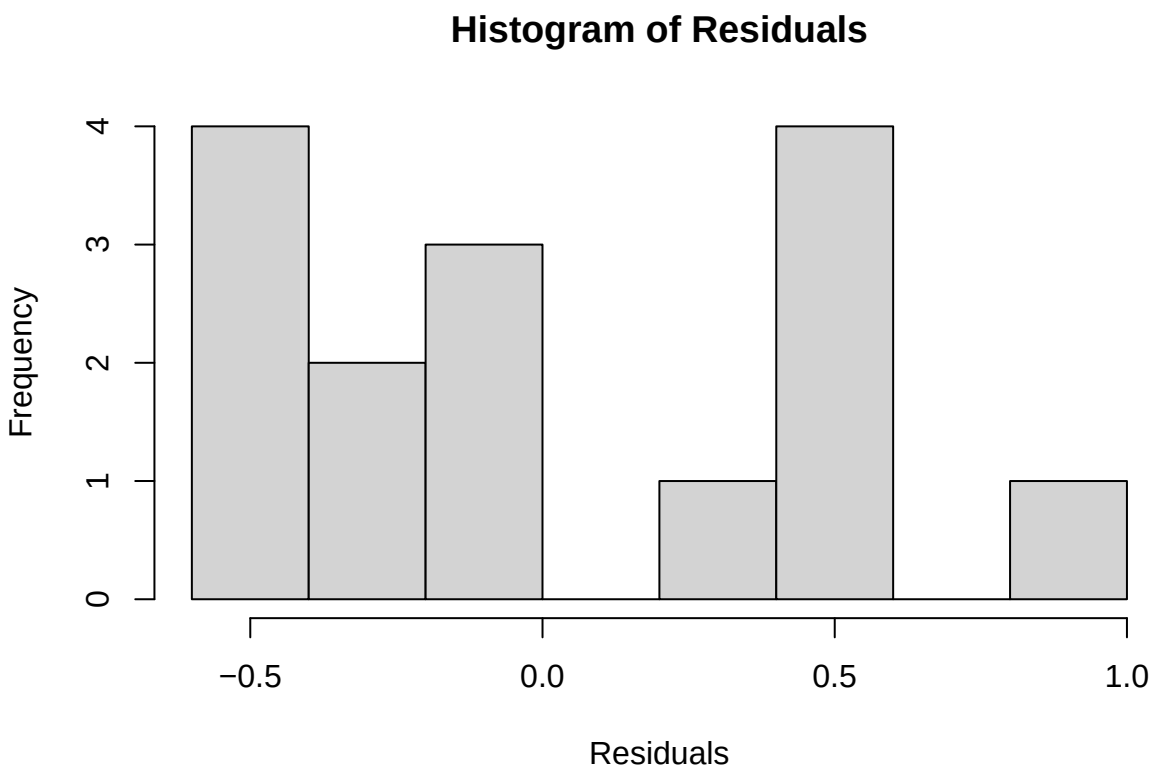
2 Fit a simple linear regression model to the data and obtain the residual plots (plot of residuals vs. fitted values, qqplot of residuals, and histogram of residuals). What do you conclude about the various conditions?

```
# Creates our SLR mode
model <- lm(y~x , data= SolutionConcentration)
# Auto creates many plots but we only want to see QQ and Residual
plot(model, which = c(1,2))
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
15
13
12
0.0
0.5
1.0
1.5
2.0
Fitted values
lm(y ~ x)




```
# Histogram of our Residuals
hist(model$residuals, main = "Histogram of Residuals", xlab= "Residuals")
```



Our Residuals vs Fitted values graph shows us that our errors appear to be scattered in a quadratic pattern. We needed to see a random spread of points on the chart however this quadratic shape means that not only did we fail our assumption of constant variance but possibly our linearity. When we look at our QQ plot to test normality, we can see that we have points above and below our line this tells us that we have heavy tails with a good majority of our observations being on the extreme. That along with our histogram not having a normal bell curve appearance tells us that we have also violated our normality assumption.

3. Write a function that can calculate the standardized observations Wi as defined above. This function should take two arguments: a vector of length n containing the Yi, and a scalar value of lambda The function should return a vector of length n which contains the W_i's.

```
# Creating a function for Wi

vectorWi<- NULL

wi <- function(yi, lambda) {

  n = length(yi)

  k2 = prod(yi)^(1/n)
  k1 = 1/(lambda*k2^(lambda-1))

  if (lambda == 0){
    vectorWi = k2*log(yi)
```

```
  }
  else{
  vectorWi = k1 * (yi^(lambda)-1)}
  return(vectorWi)
}
```

Question 4

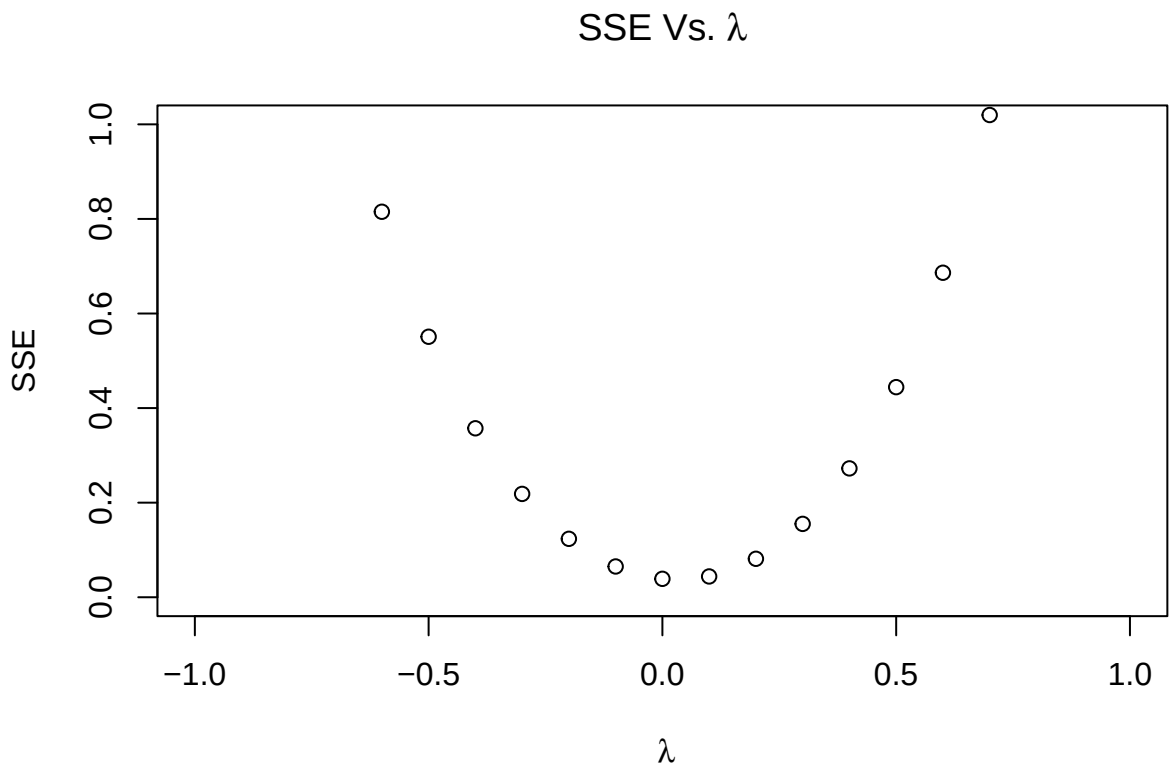
```
sse <- function(xi,wi){

  stdmodel <- lm(wi~xi)
  return (sum((wi- fitted(stdmodel))^2))
}
```

Question 5

```
lambdaset<- seq(from=-3, to=3, by=.1)

sse_calc<- rep(NA,length(lambdaset))

index <- 1

yi<-SolutionConcentration$y
xi<-SolutionConcentration$x

for(lda in seq(from=-3, to=3, by=.1)){
  Wi_calc <- wi(yi, lda)
  sse_calc[index] <- sse(xi,Wi_calc)
  index<- index+1
}
```

Question 6

```
Q3_plot<- plot(lambdaset, sse_calc, xlab = expression(~lambda), ylab= "SSE", main = expression(paste("S
```



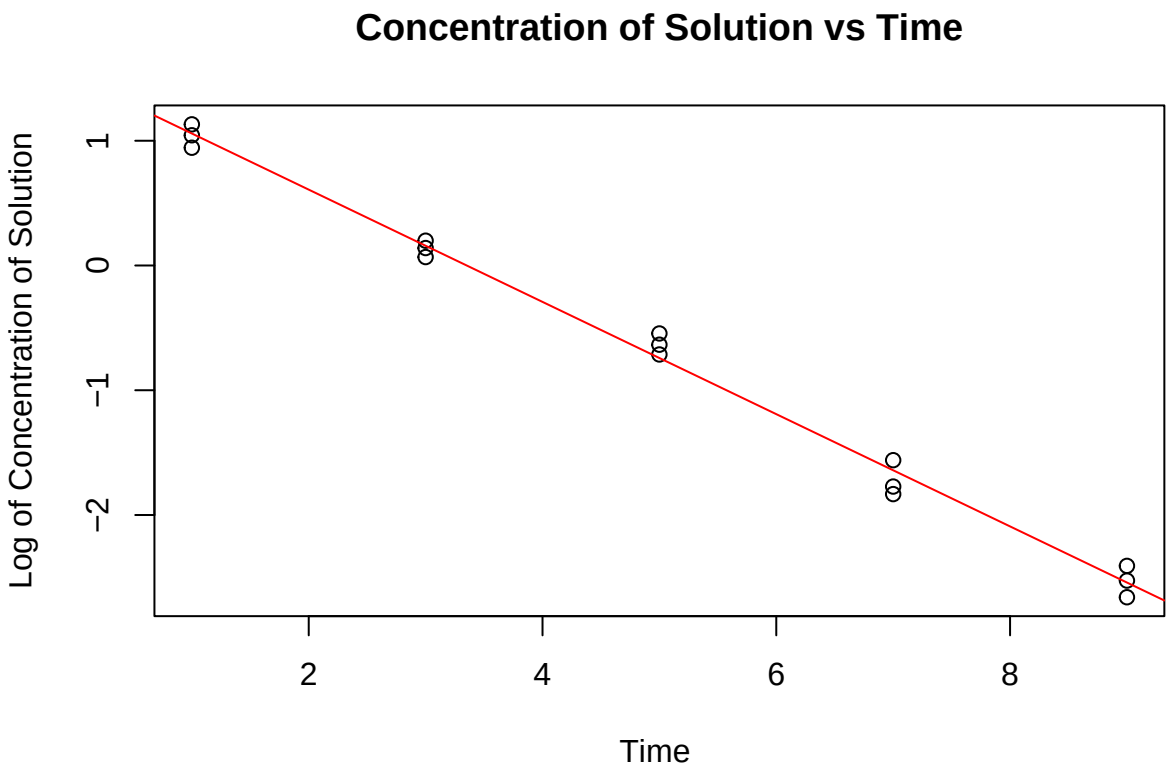

```
## Calculating the min lambda
lambdaset[which.min(sse_calc)]
```

```
## [1] 0
```

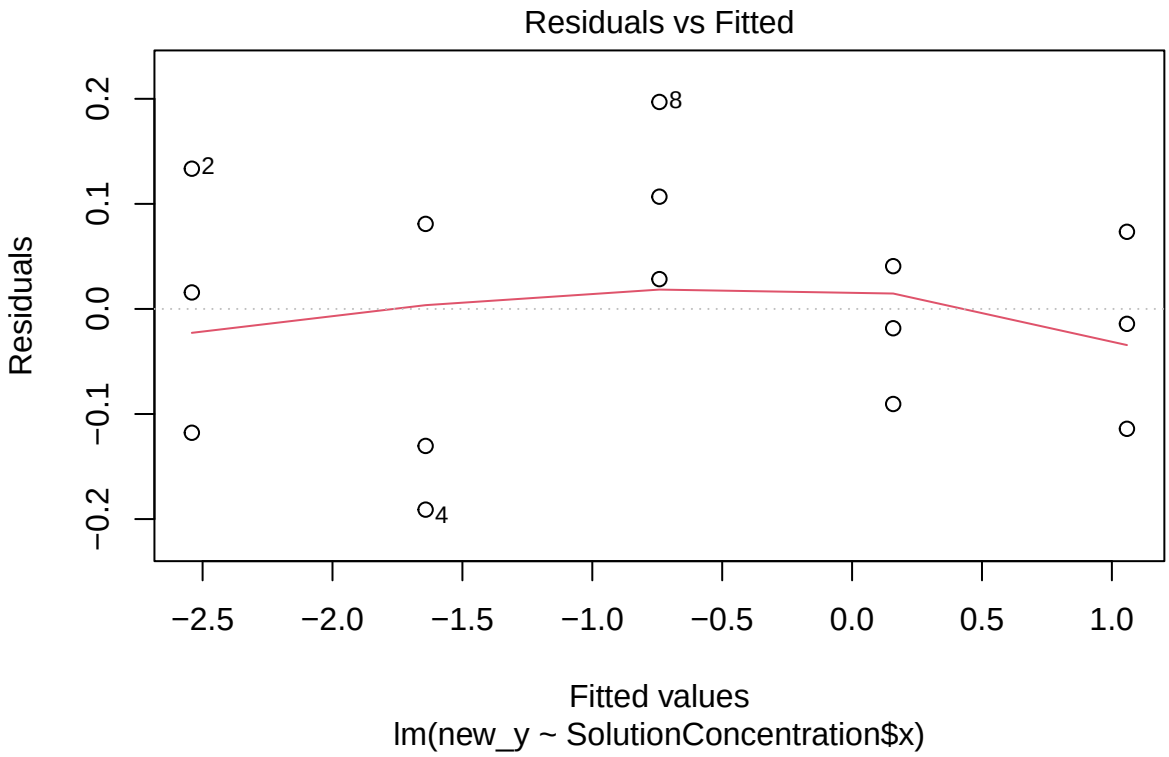
The value of lambda that minimizes our SSE is 0.

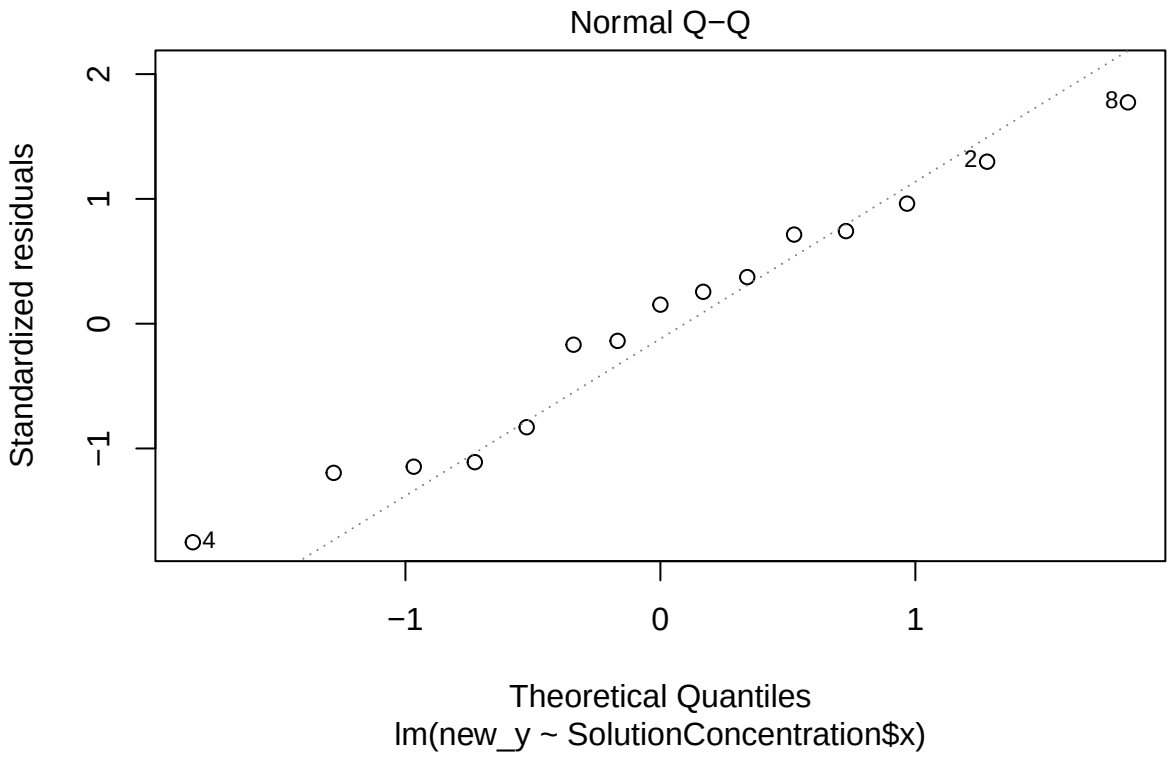
7 . Fit the model that is suggested by the Box-Cox transformation, and check the conditions for the model. If the conditions are satisfied, interpret the slope ofthe final model.

```
## plotting the Suggested Box-Cox transformation
new_y <-log((SolutionConcentration$y))
plot(SolutionConcentration$x,new_y, xlab = "Time ", ylab = "Log of Concentration of Solution", main = "
abline(TransformedYIModel<- lm(new_y~SolutionConcentration$x), col = "red")
```

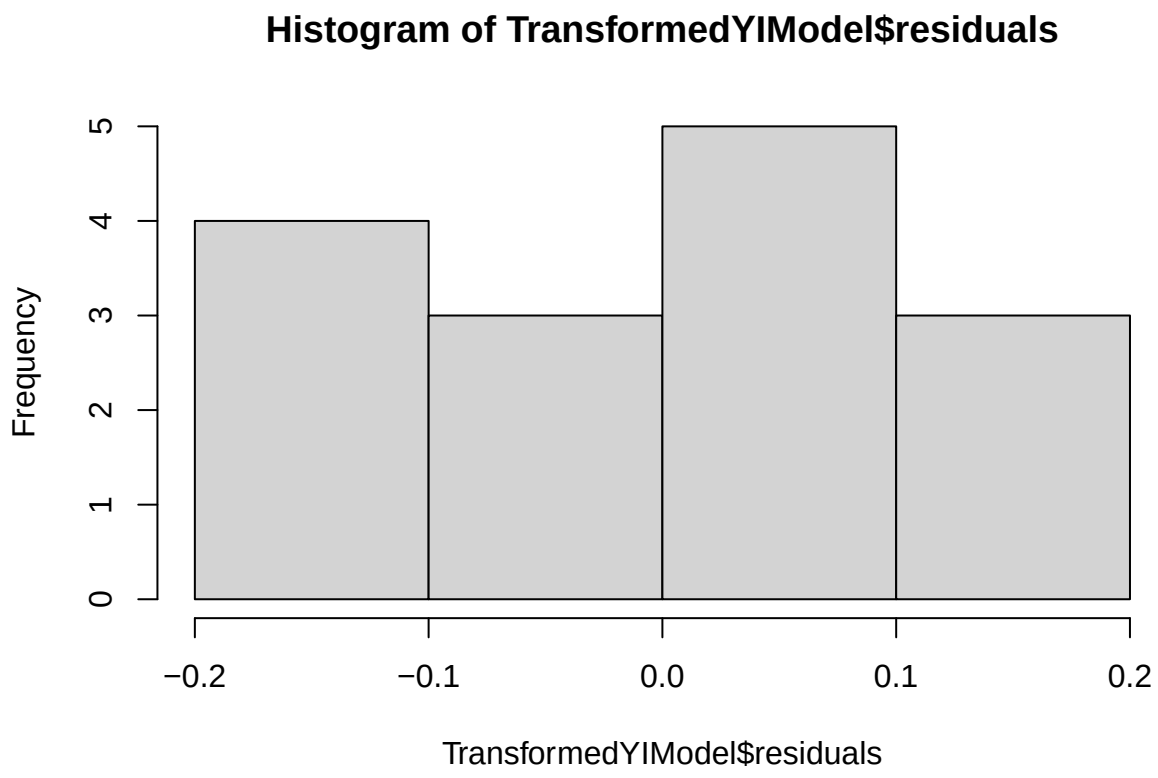



```
## CHECKING CONDITIONS
plot(TransformedYIModel, which = c(1,2))
```


Residuals vs Fitted
Residuals
0.2
0.1
0.0
−0.1
−0.2
2
8
4
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
Fitted values
lm(new_y ~ SolutionConcentration$x)




```
hist(TransformedYIModel$residuals)
```




```
## Interpreting Our Coefficient
summary(TransformedYIModel)
```

```
##
## Call:
## lm(formula = new_y ~ SolutionConcentration$x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19102 -0.10228  0.01569  0.07716  0.19699
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.50792    0.06028   25.01 2.22e-12 ***
## SolutionConcentration$x -0.44993    0.01049  -42.88 2.19e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.115 on 13 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9924
## F-statistic:  1838 on 1 and 13 DF,  p-value: 2.188e-15
```

```
value<-((exp(TransformedYIModel$coefficients[2])) - 1)*100
```

Our Residual vs fitted values is now randomly dispersed allowing us to meet our constant variance assumption. Our QQ plot looks a lot more normal, with more of the points being on the line. Furthermore our

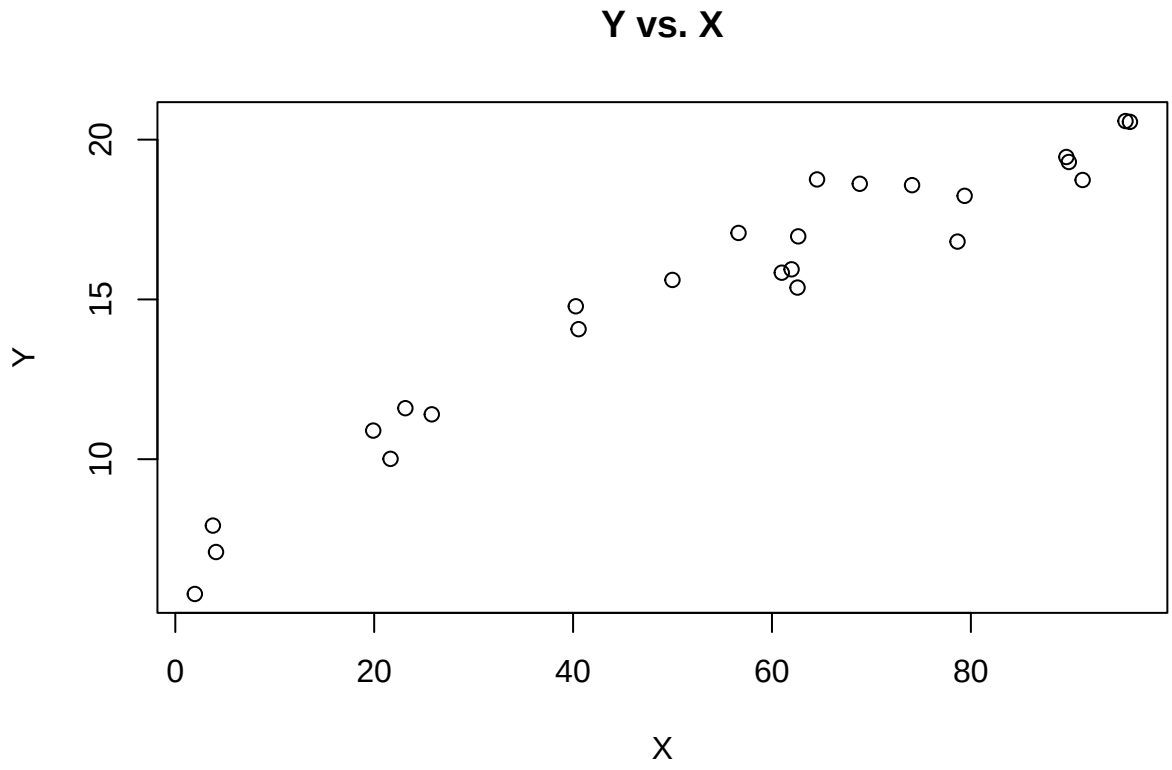
histogram now has that bell curve shape centered at 0 thus we have our normality assumption also met.

Now to interpret the slope, From our summary statistics we can see that our f test was significant at the 95% level and our B1 value has a pval that's <.05 making it also significant at the 95% level. Our coefficient after being exponentiated to deal with out transformation, tells us that for every 1 hour increase in time, our Concentration of solution decreases by 36.2 percent.

Section 2 - Transforming X - The Box-Tidwell Transformation

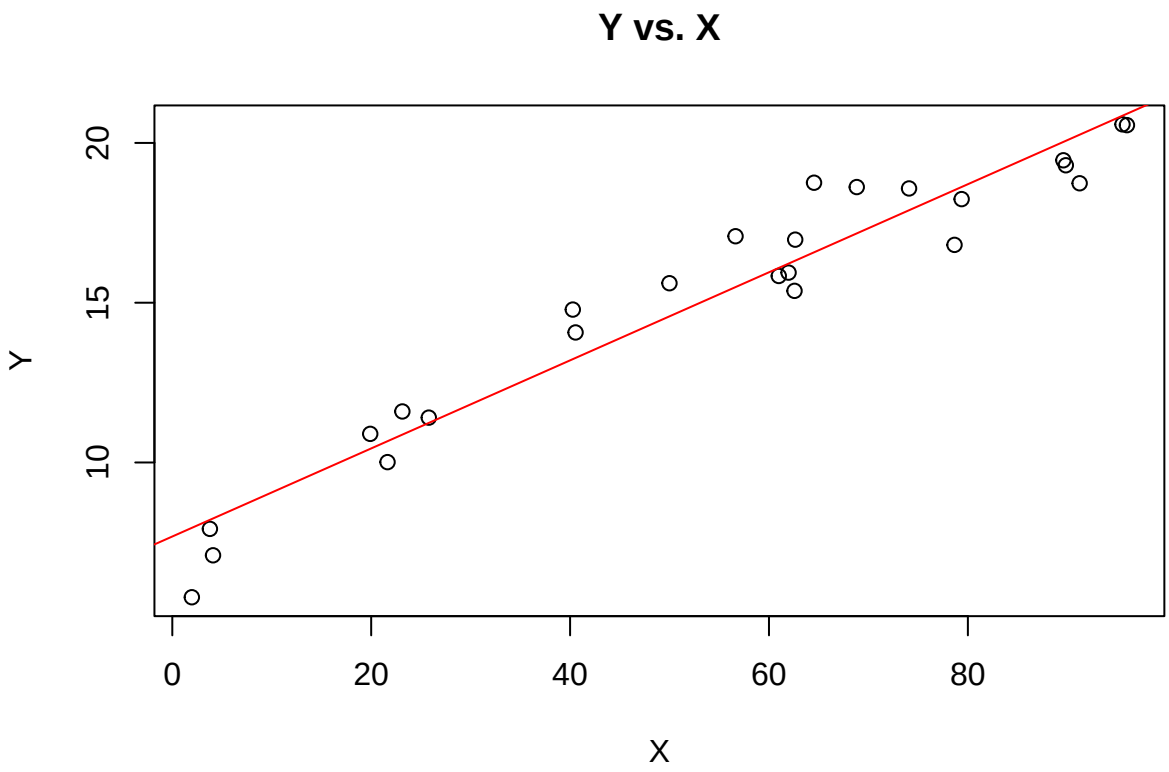
```
library(readr)
Lab4q2 <- read_csv("Lab4q2.csv", show_col_types = FALSE)
```

```
#plotting the data
plot(Lab4q2$x,Lab4q2$y, xlab = "X", ylab = "Y", main = "Y vs. X")
```

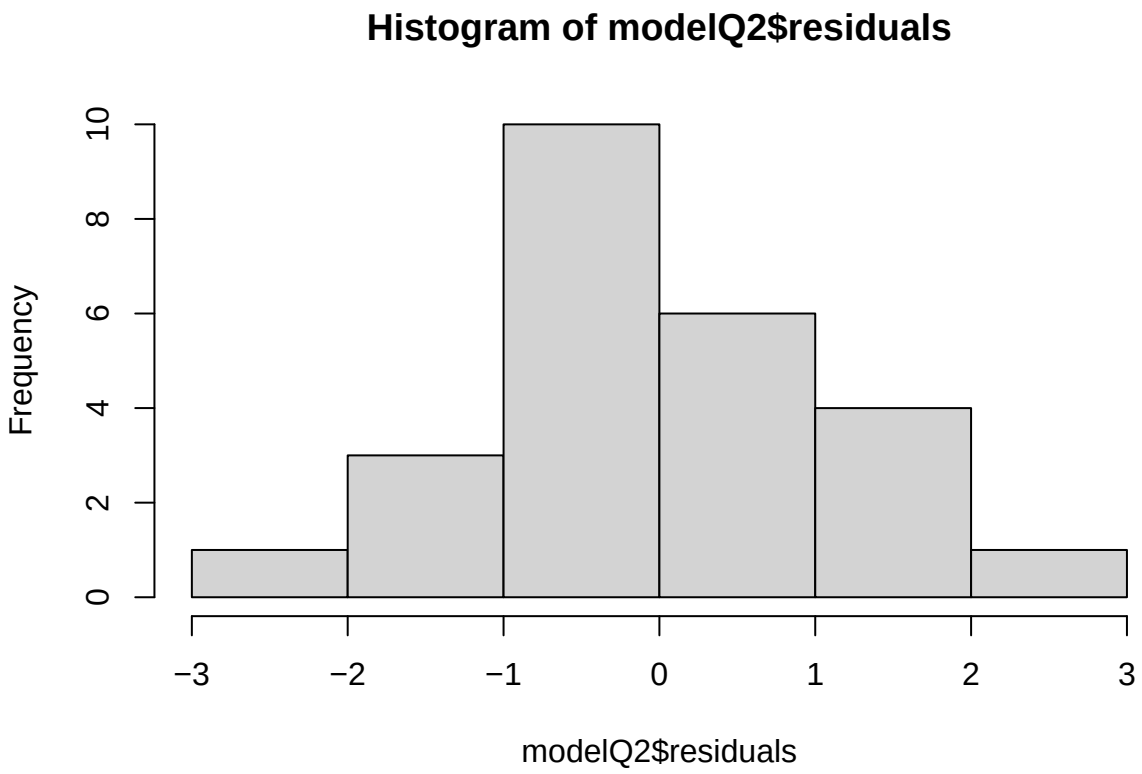


2. Fit a simple linear regression model and obtain the residual plots. What can you conclude about the conditions for SLR?

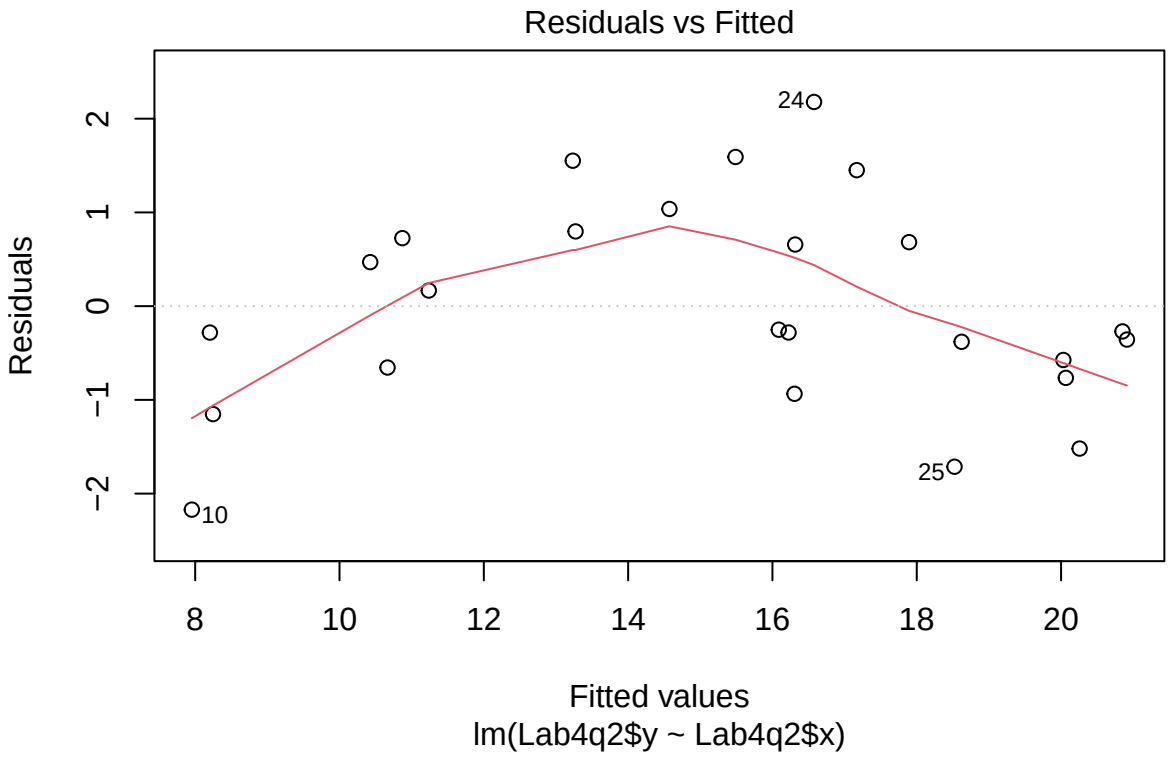
```
plot(Lab4q2$x,Lab4q2$y, xlab = "X", ylab = "Y", main = "Y vs. X")
abline(modelQ2<-lm(Lab4q2$y~Lab4q2$x), col = "red")
```

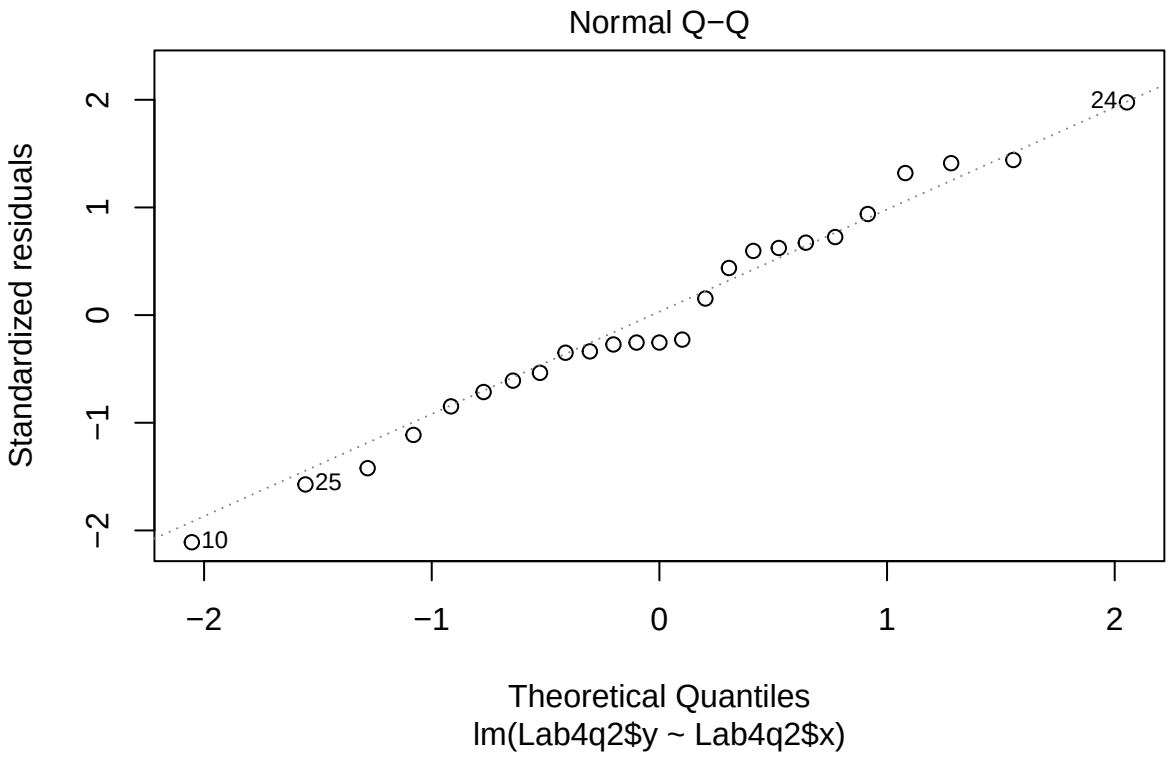



```
hist(modelQ2$residuals)
```




```
plot(modelQ2, which = c(1,2))
```


Residuals vs Fitted
Residuals
Fitted values
lm(Lab4q2$y ~ Lab4q2$x)
24
25
10
2
1
0
−1
−2
8
10
12
14
16
18
20



Our plot and regression originally appear to be a perfect linear fit but our residuals tell a different story. The normality condition is met as our qq plot appears to have all points on the line and our histogram displays a normal distribution. When we look at our residual vs fitted values plot, our assumption of constant variance is violated as there appears to be a parabolic arc to the behavior of the residuals.

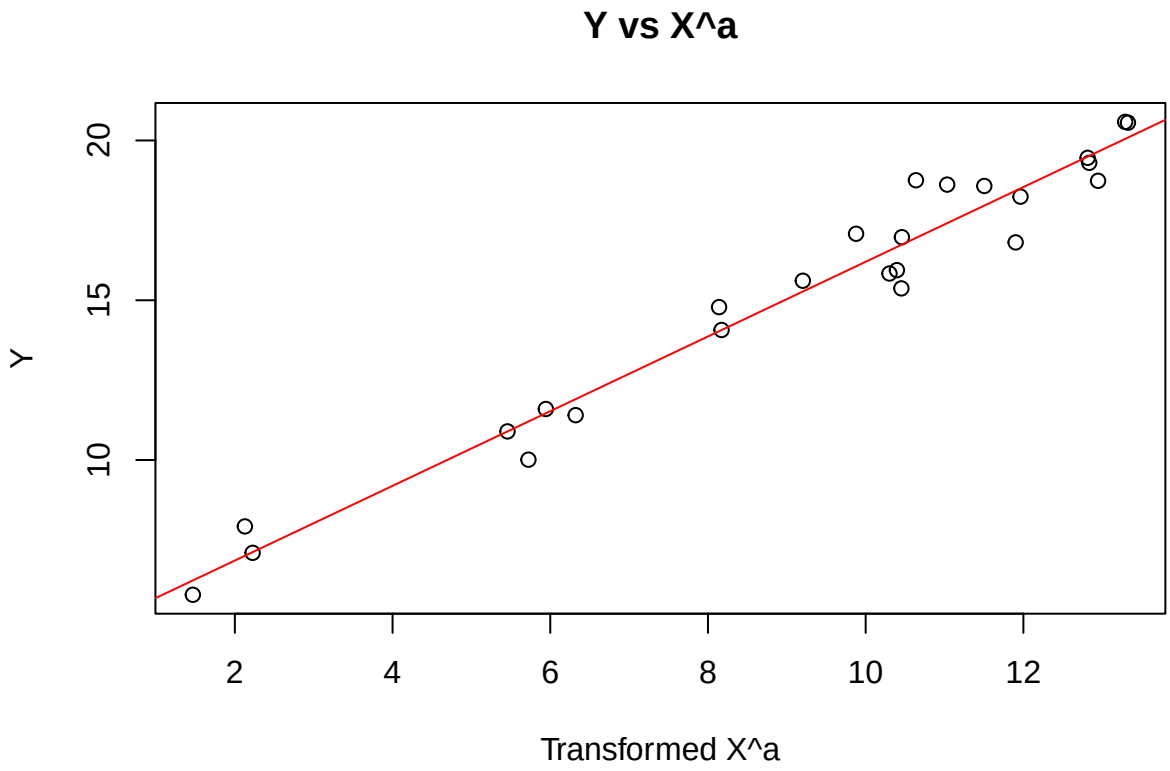
Question 3

```
a_calc <- function(x, y, a){
    X <- x^a
    XLOG<- X*log(x)



    model1<- lm(y~X )
    model2<- lm(y~ X + XLOG)
    summary(model2)
    a_new <- a + coefficients(model2)[3]/coefficients(model1)[2]
 return (a_new)
}
```

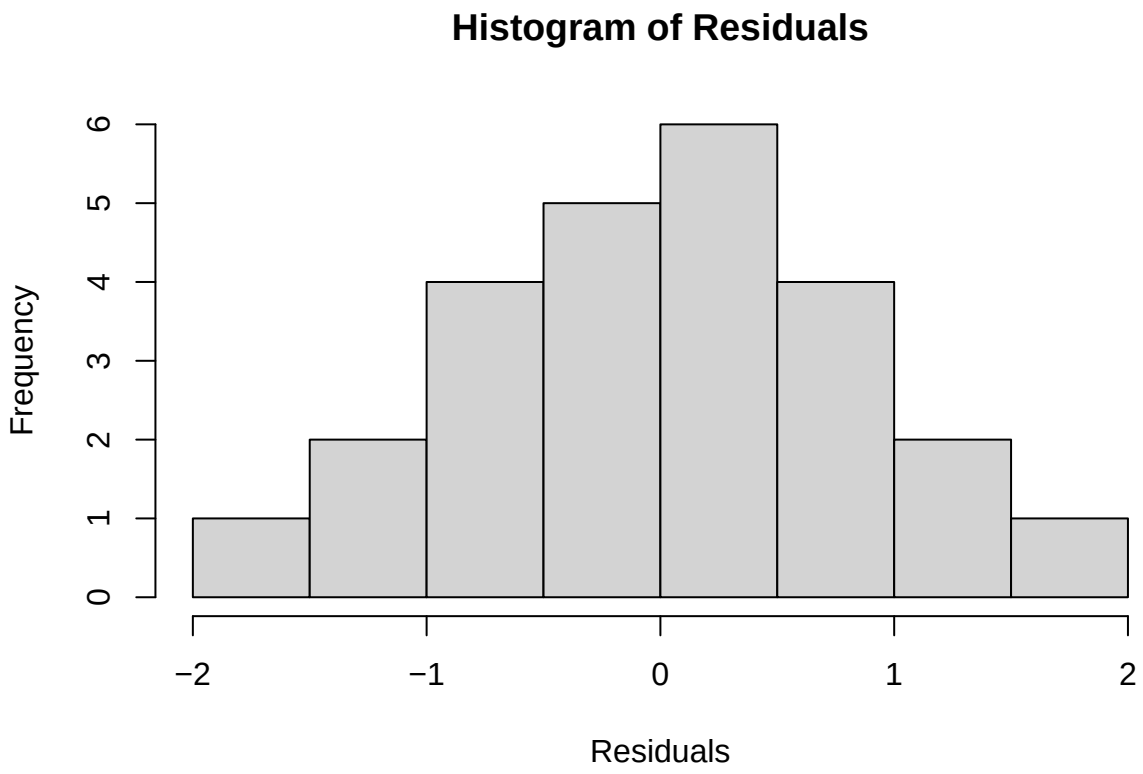
Question 4

```
# Calculates our Alpha vector
alpha<- NULL
alpha[1]<- 1
for( i in 1:10 ){
  alpha[i+1] <- a_calc(Lab4q2$x,Lab4q2$y,alpha[i])
}
```

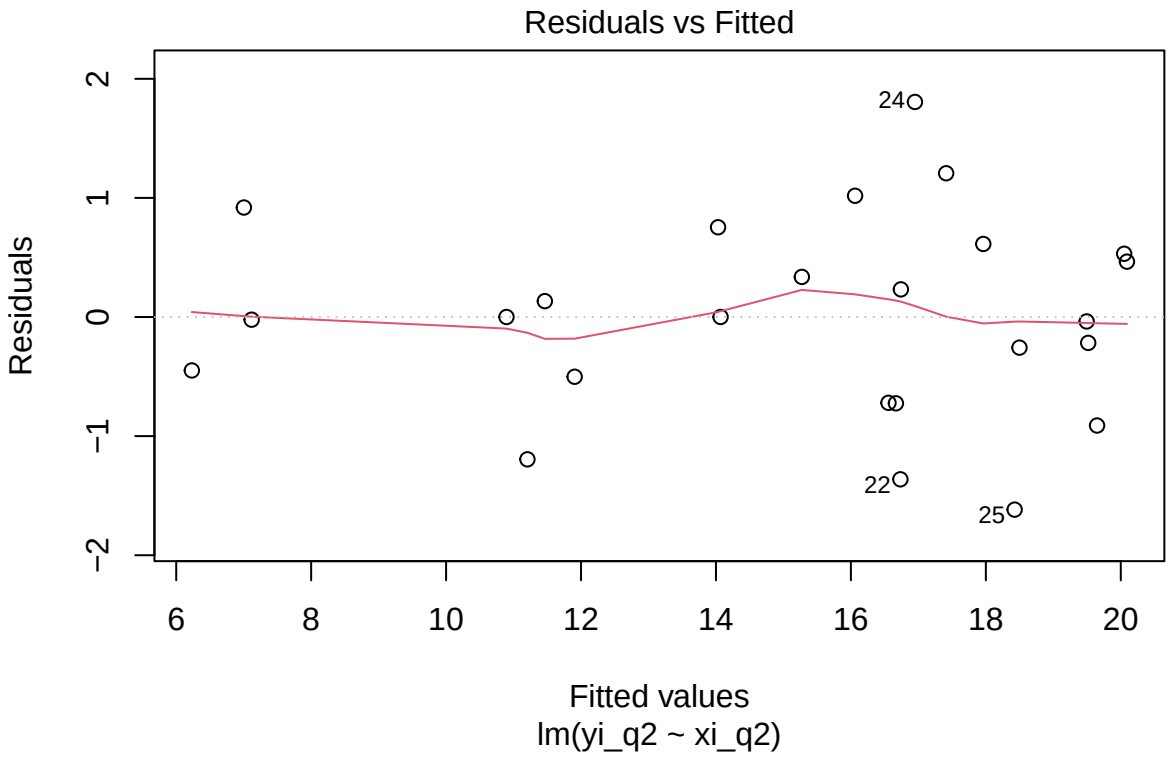
```
# Fitting the regression model using that alpha that converged
xi_q2<- Lab4q2$x^alpha[10]
yi_q2<- Lab4q2$y

plot(xi_q2,yi_q2, xlab = "Transformed X^a",ylab = "Y", main = "Y vs X^a" )
abline(TransformedXIModel<-lm(yi_q2~xi_q2), col = "red")
```

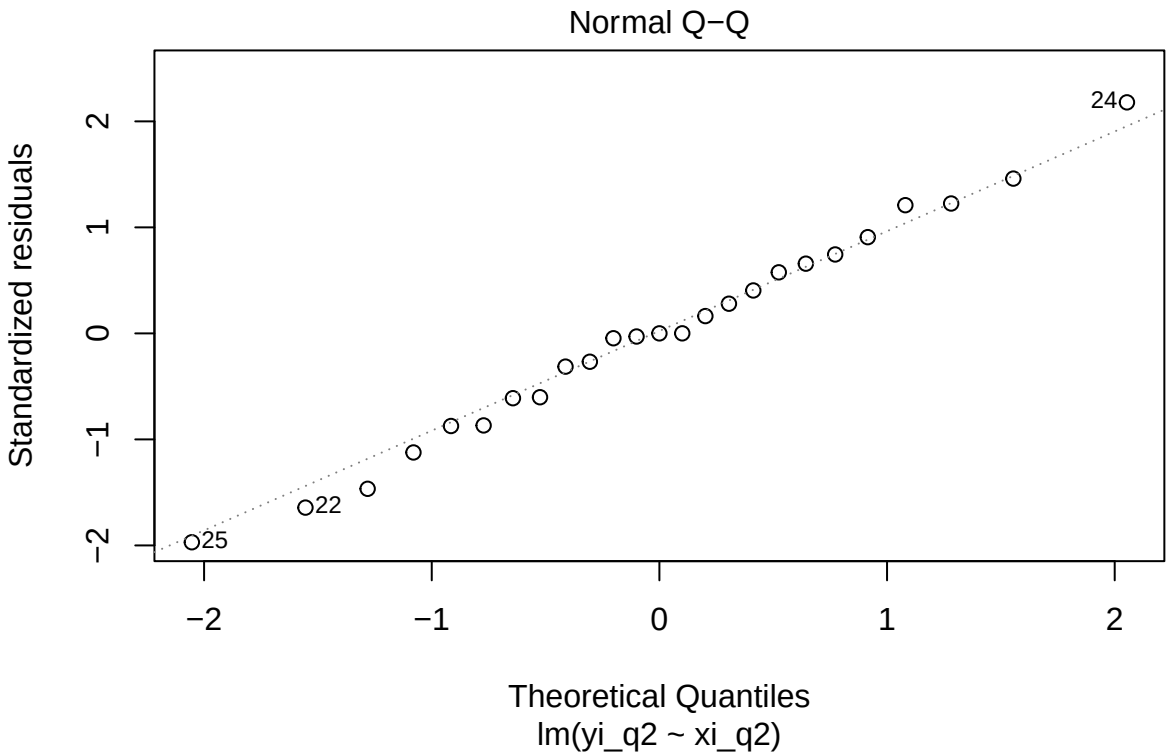



```
#histogram
hist(TransformedXIModel$residuals, main = "Histogram of Residuals", xlab= "Residuals")
```




```
#new regression analysis
plot(TransformedXIModel, which = c(1,2))
```


Residuals vs Fitted
Residuals
Fitted values
lm(yi_q2 ~ xi_q2)
24
22
25
2
1
0
−1
−2
6
8
10
12
14
16
18
20



Our transformation was done with hopes of fixing our lack of constant variance in our residuals, as we can see from our residuals vs fitted plots, we now have a random scatter of points indicating constant variance. This transformation also had the added benefit of giving us clearer representations of normality in the sense that our histogram of our residual looks more normal along with the points on our QQ plot fitting closer to the line.

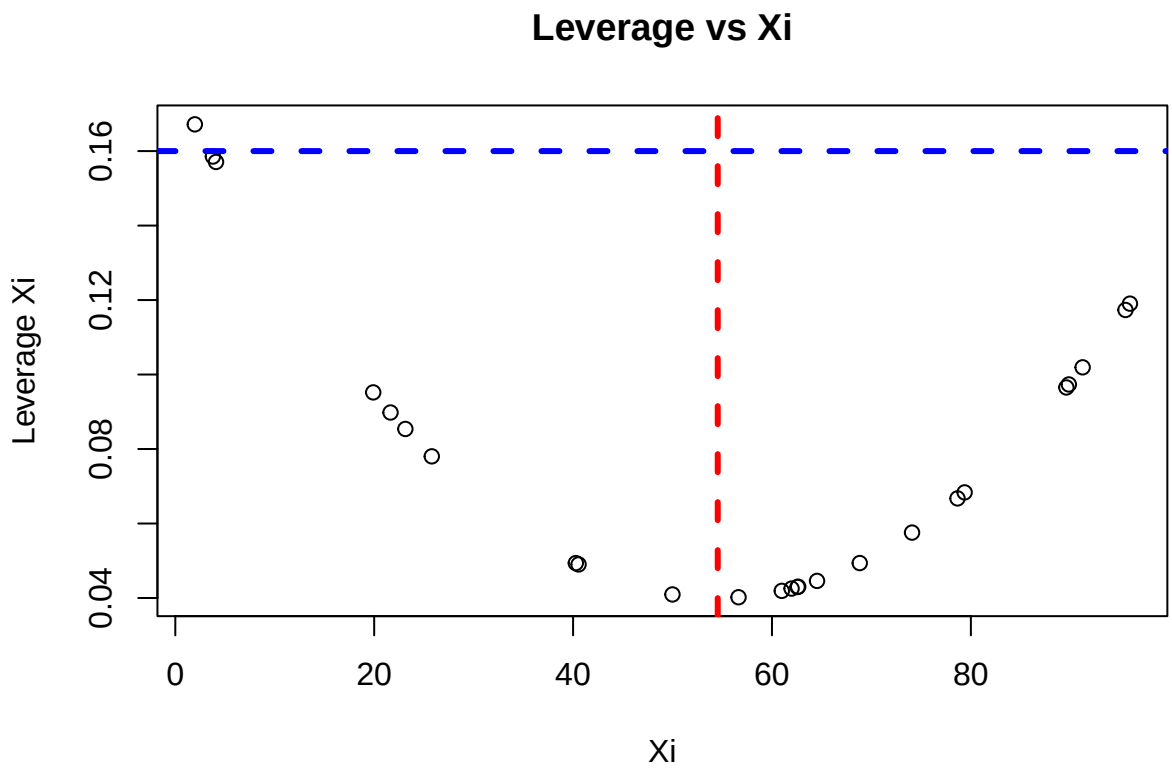
Section 3:Calculating Leverage and Externally Studentized Residuals

Question 1

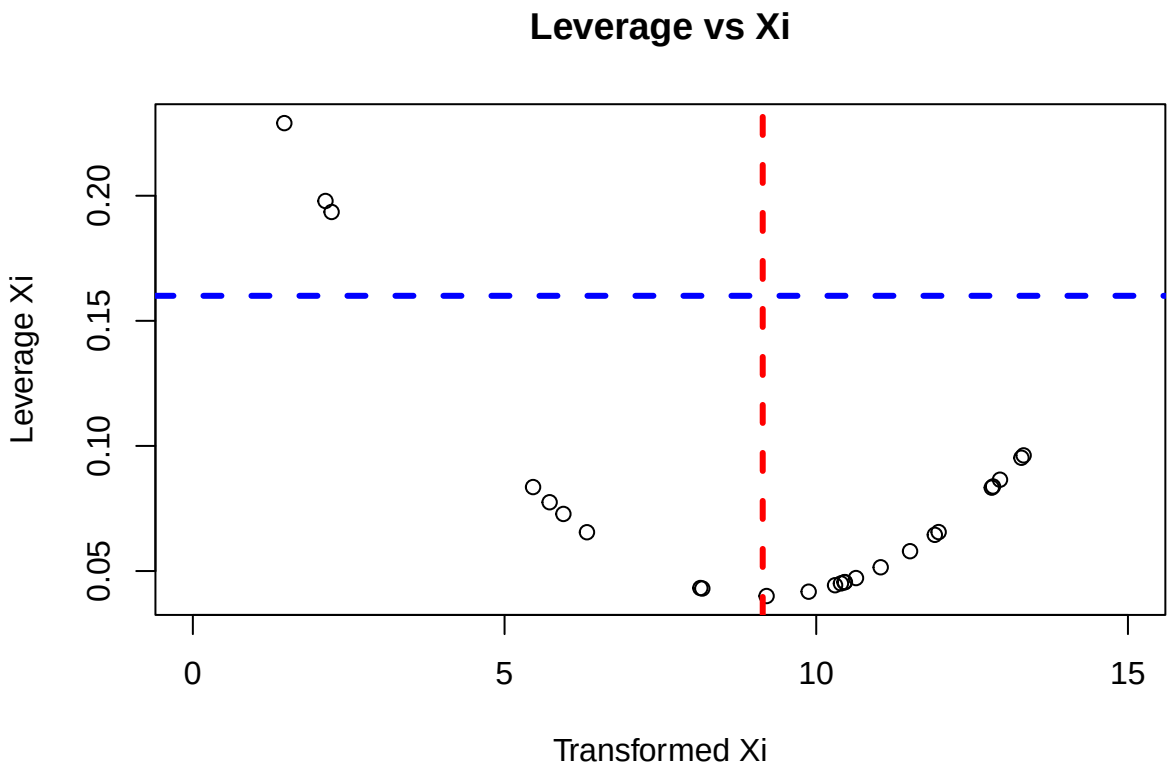
```
# function to calculate leverage
leverage <- function(xnew){
  n<- length(xnew)
  xbar<- mean(xnew)
  hii<- (1/n) + ((xnew-xbar)^2/ sum((xnew-xbar)^2)   )

  return(hii)
}
```

Question 2

```
#Applying function to above
plot(Lab4q2$x,leverage(Lab4q2$x), xlab = "Xi", ylab = "Leverage Xi", main = "Leverage vs Xi")
abline(v = mean(Lab4q2$x), col="red", lwd=3, lty=2)

# Leverage line
abline(h = 4/length(Lab4q2$x), col="blue", lwd=3, lty=2)
```



```r
#Applying function to q2
plot(xi_q2,leverage(xi_q2), xlab = "Transformed Xi", ylab = "Leverage Xi", main = "Leverage vs Xi", xli
abline(v = mean(xi_q2), col="red", lwd=3, lty=2)
# Leverage line
abline(h = 4/length(xi_q2), col="blue", lwd=3, lty=2)
```



None of our observations appear to have high leverage. The way we reached this conclusion is by adding our leverage line (the blue horizontal line on our graph) which tells us the threshold for a leverage value that would be considered too large. Leverage is essentially the measurement of how much our observed values influence our predicted values. The general rule is that if hii $>$ 2 X (#of predictors) / n then its too large. So that concept is what allowed us to reach the conclusion of plotting it on our graph. The first Question had one observation that had a large leverage while the second section had 2.

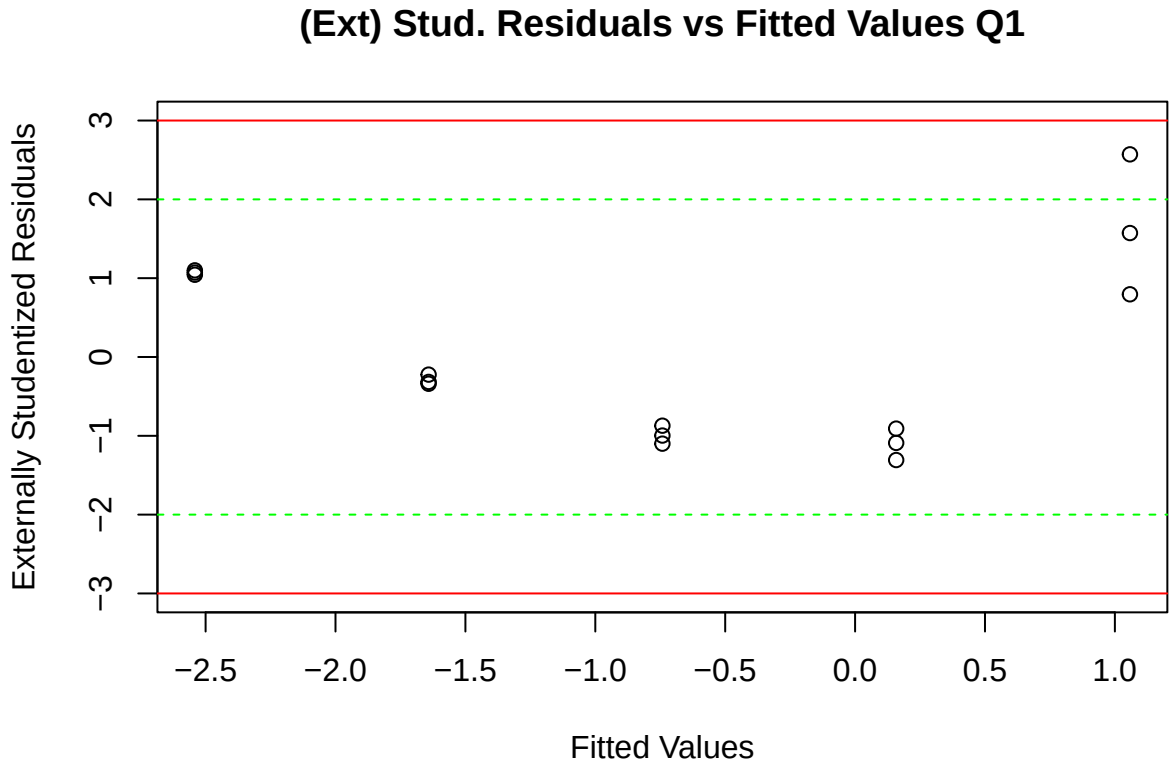
Question 3

```r
#function that calculates externally studentized residuals

ExtStudResid<- function(xi,yi){
  mod<-lm(yi~xi)
  ri<-mod$residual/sqrt(mean(mod$residuals^2) * (1-leverage(xi)))
  n<- length(xi)
  inner<- (n-3)/(n-2-ri^2)
  ti<- ri*sqrt(inner)
```

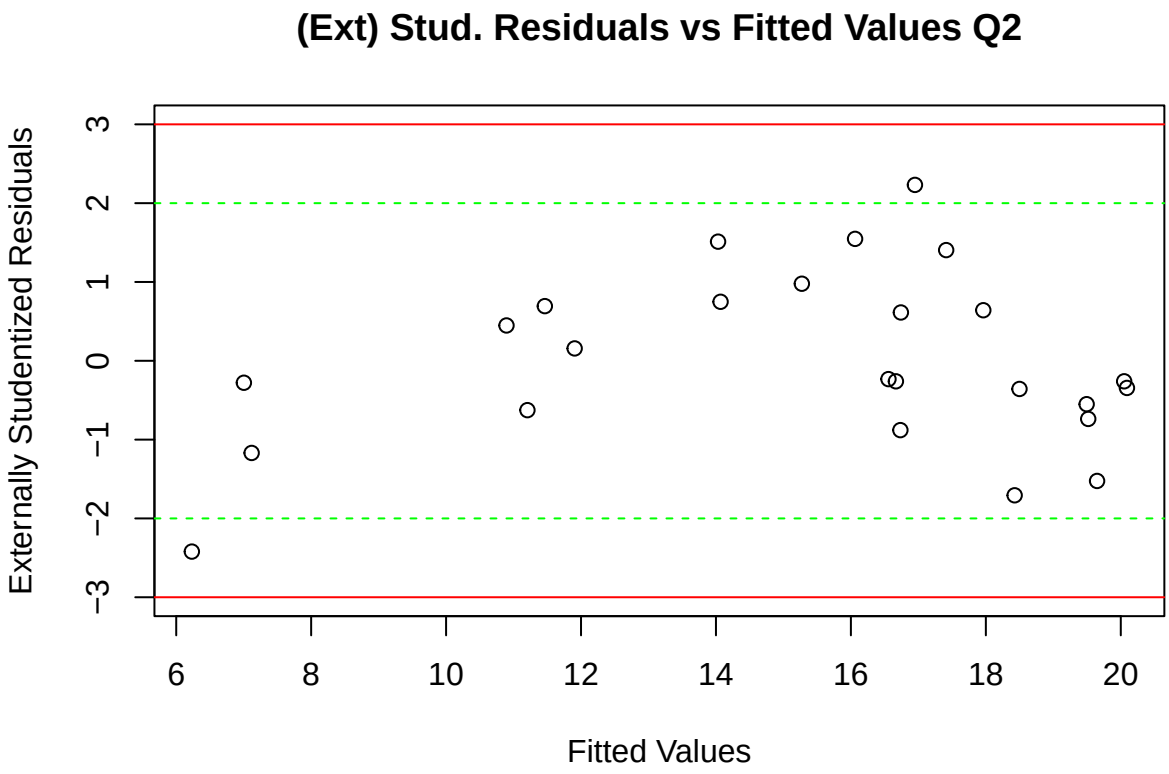
```
  return(ti)
}
```

Question 4

```
plot(TransformedYIModel$fitted.values,ExtStudResid(SolutionConcentration$x,SolutionConcentration$y), xla
abline(h= 3, col = "red")
abline(h= 2, col = "green", lty=2)
abline(h= -3, col = "red")
abline(h= -2, col = "green", lty=2)
```



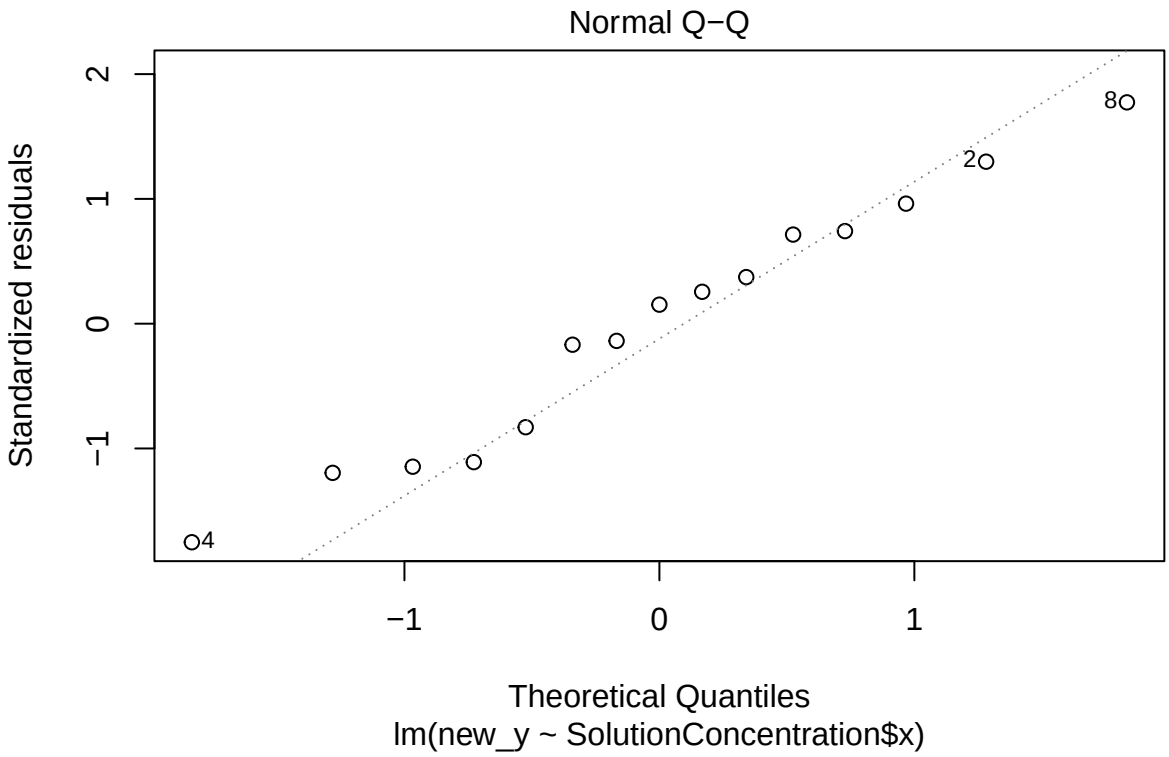

```
plot(TransformedXIModel$fitted.values,ExtStudResid(Lab4q2$x,Lab4q2$y), xlab = "Fitted Values", ylab = "
abline(h= 3, col = "red")
abline(h= 2, col = "green", lty=2)
abline(h= -3, col = "red")
abline(h= -2, col = "green", lty=2)
```



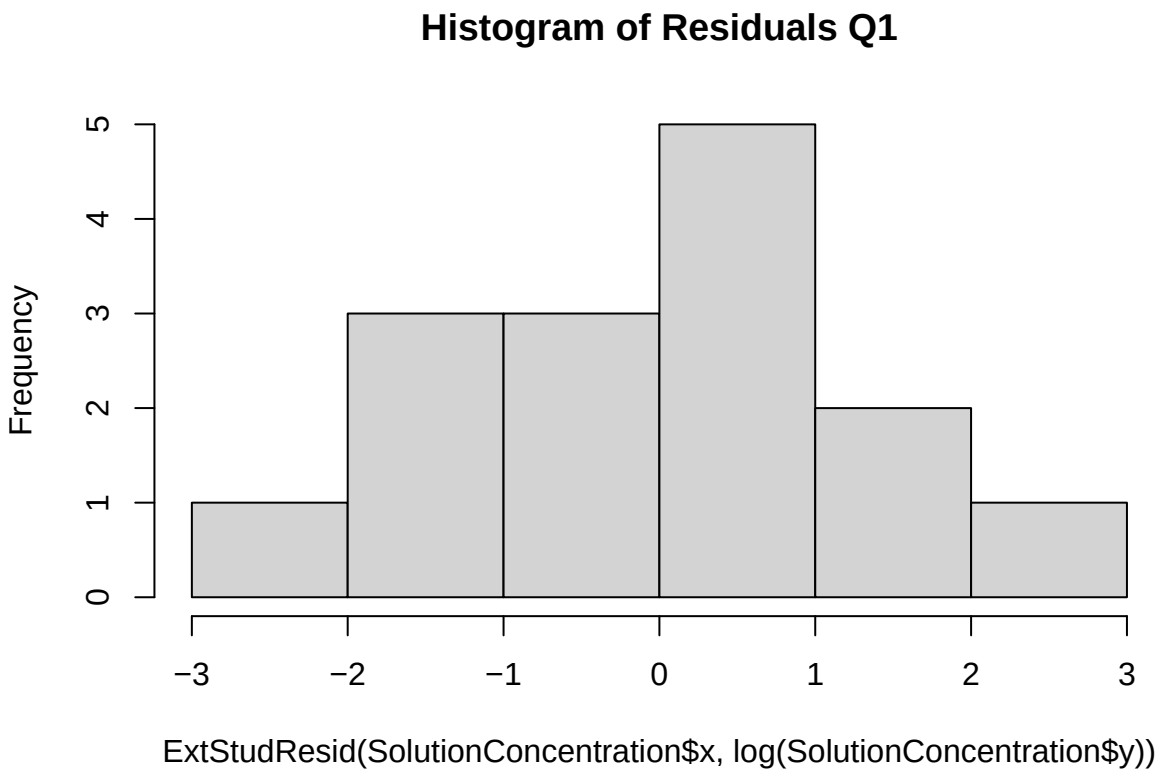
When we think outlier, we are referring to the specific y value relative to the overall bivariate relationship. The way we achieve this is through our studentized residuals. Usually, values greater than 3 or less than -3 be our obvious outliers but our band could be tightened if we wanted to clean up our data more. When making the band from -2, 2 we can see that in the first plot from the first section there's only one observation that breaks our boundaries while in the second question there are 2.The problem with outliers is that they are highly influential and they can impact the slope (aka relationship of our variables with one another) which could have negative implications when we're trying to make sense of our output.

Question 5

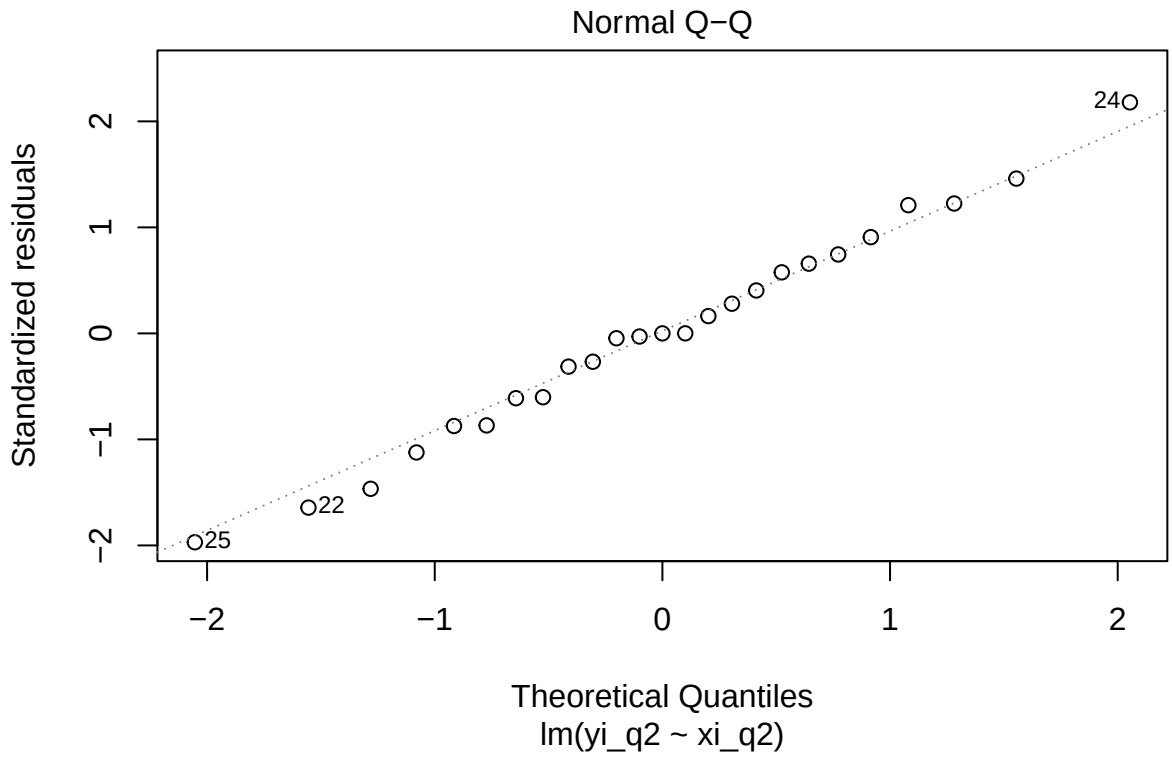
```
plot(TransformedYIModel, which = 2)
```



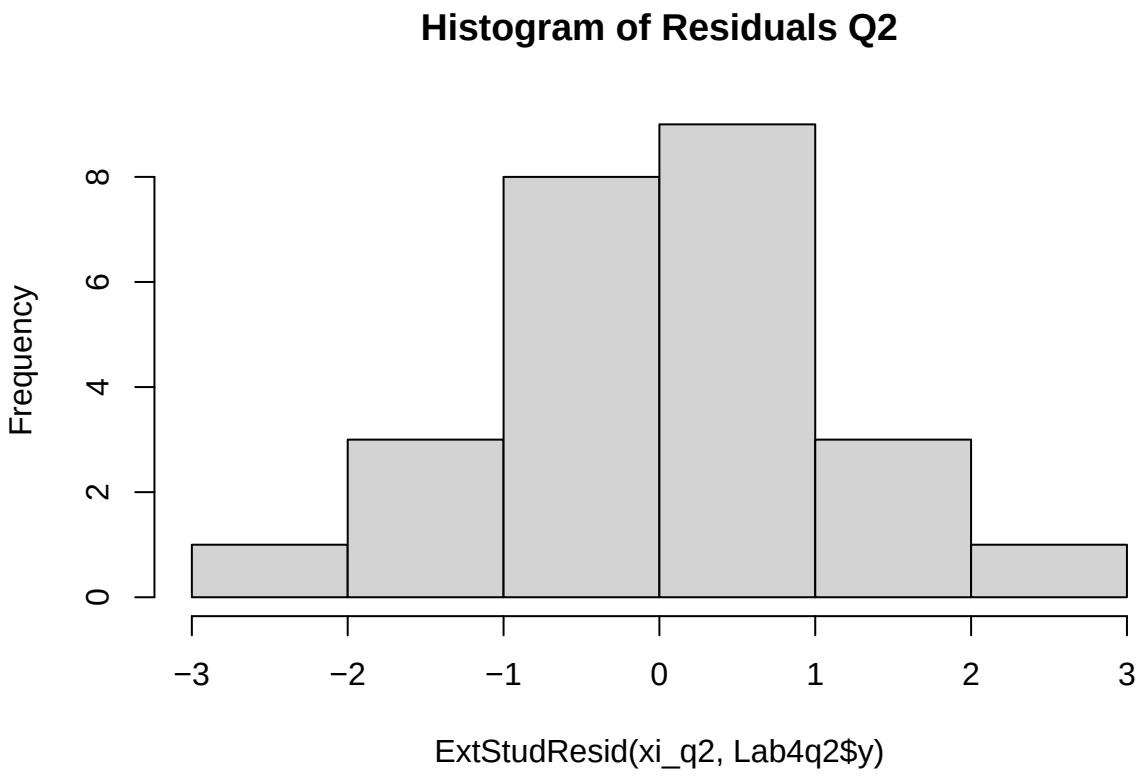

```
hist(ExtStudResid(SolutionConcentration$x,log(SolutionConcentration$y)), main = "Histogram of Residuals
```




```
plot(TransformedXIModel, which = 2)
```




```
hist(ExtStudResid(xi_q2,Lab4q2$y), main = "Histogram of Residuals Q2")
```



Our QQ plot has a majority of our points fitted to the line for both sections, the second section does display a better fitting QQ plot. When it comes to the residuals of each section, they both have normally distrubuted standardized residuals. This allows us to meet our normality assumption for linear regression.