Homework Set 5

Hyun Suk (Max) Ryoo (hr2ee)

10/2/2021

Question 1

In your own words, try to explain the following question an undergraduate student asks you: "Why do we transform the response variable when the constant variance assumption is not met, instead of transforming the predictor variable?"

The constant variance assumption is that the spread of the errors is the same as the spread of the response variable. If the equality is not the same we need to transform the response variable since the response variable is what dicatates the variance. The predictor variable is constant or fixed and changing it won't really change the variance. Therfore, we should transform the response variable.

Question 2

For this question, we will use the cornnit data set from the faraway package. Be sure to install and load the faraway package first, and then load the data set. The data explore the relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application in a study carried out in Wisconsin.

(Prework)

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.0.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

```
## -- Conflicts -------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

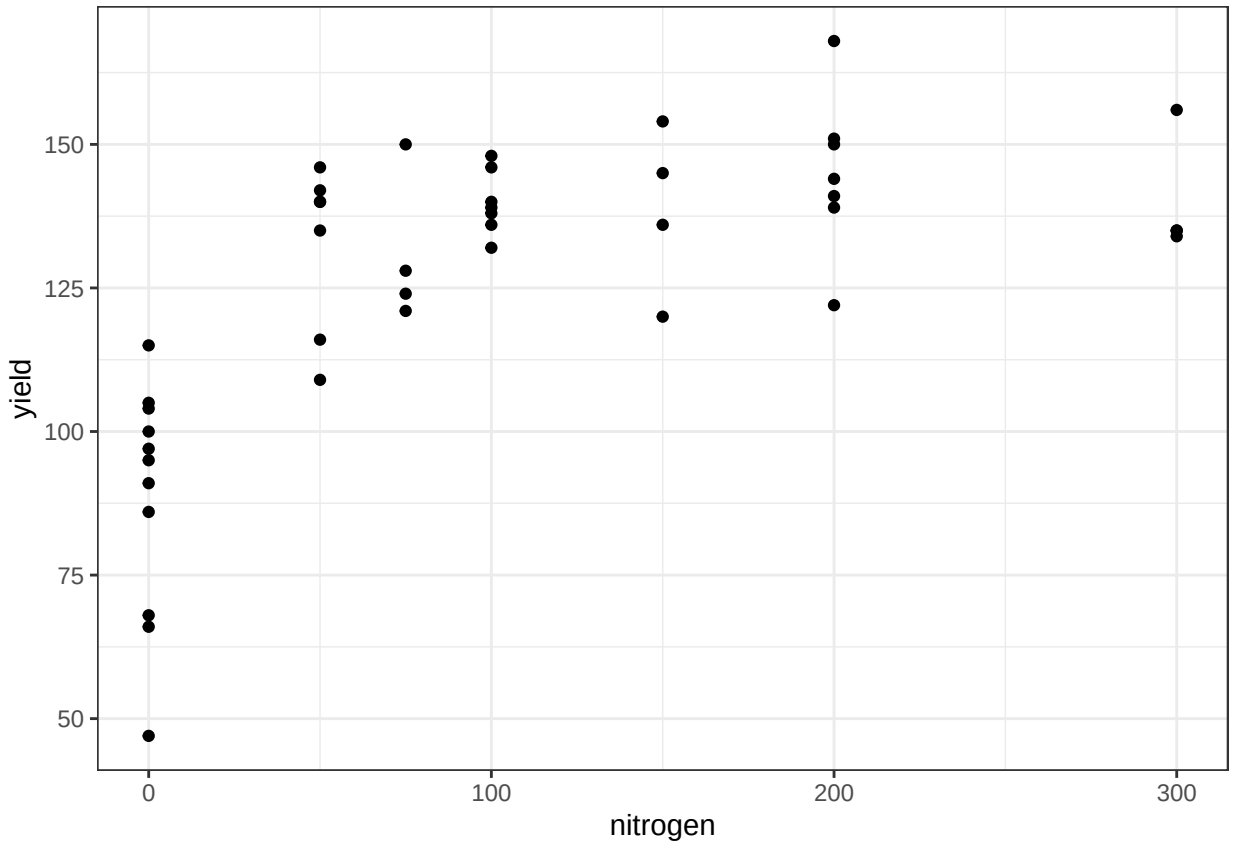
```
data <- cornnit
head(data)
```

```
##   yield nitrogen
## 1   115        0
## 2   128       75
## 3   136      150
## 4   135      300
## 5    97        0
## 6   150       75
```

A) What is the response variable and predictor for this study? Create a scatterplot of the data, and interpret the scatterplot.

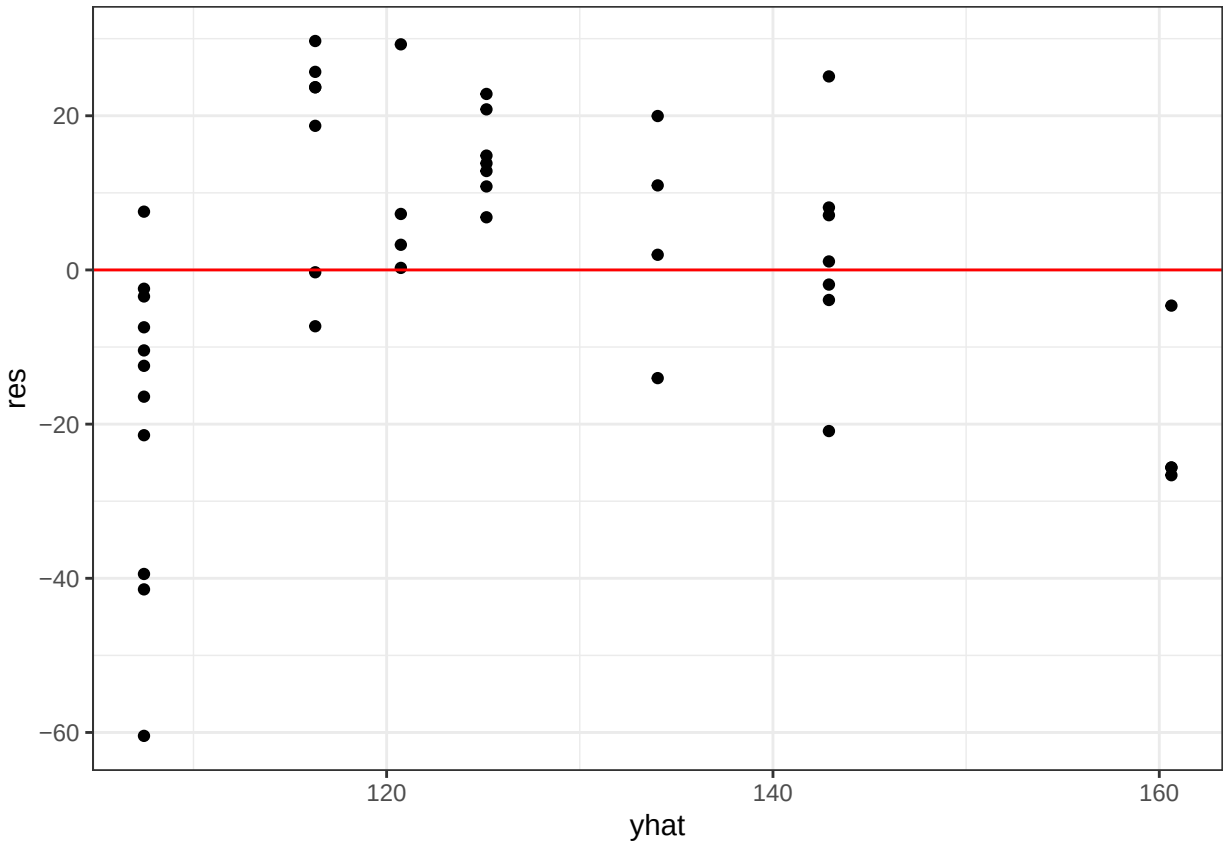
The response variable will be yield and the predictor will be nitrogen.

```
data %>%
  ggplot(aes(nitrogen, yield)) + geom_point() + theme_bw()
```



B) Fit a linear regression without any transformations. Create the corresponding residual plot. Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.

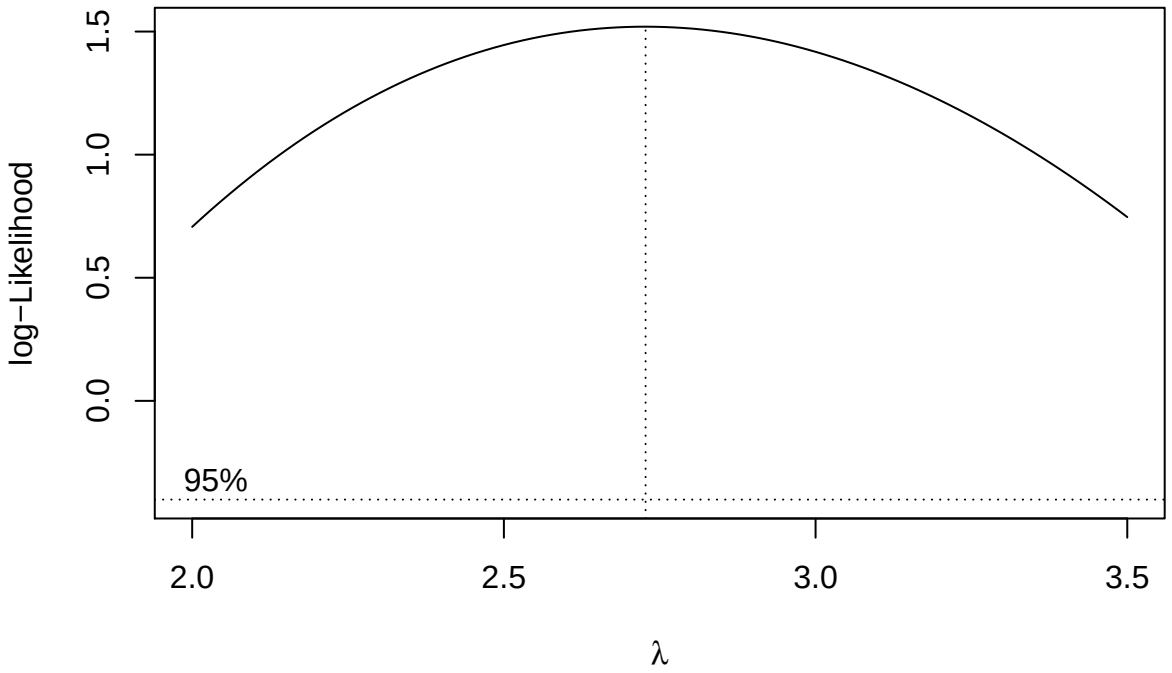
```
result <- lm(yield~nitrogen, data=data)
yhat = result$fitted.values
res = result$residuals
data %>%
  ggplot(aes(yhat, res)) + geom_point() + theme_bw() + geom_hline(yintercept=0, color="red")
```



The variance of the residuals is not equal through the x-axis. It seems like for variance decreases for higher values of fitted values. The trend doesn't seem to be equally divided across the x-axis, which means there needs to be transofmration. I will first try to transform the response variables due to the observations.

C Create a Box Cox plot for the profile loglikelihoods. How does this plot aid in your data transformation?

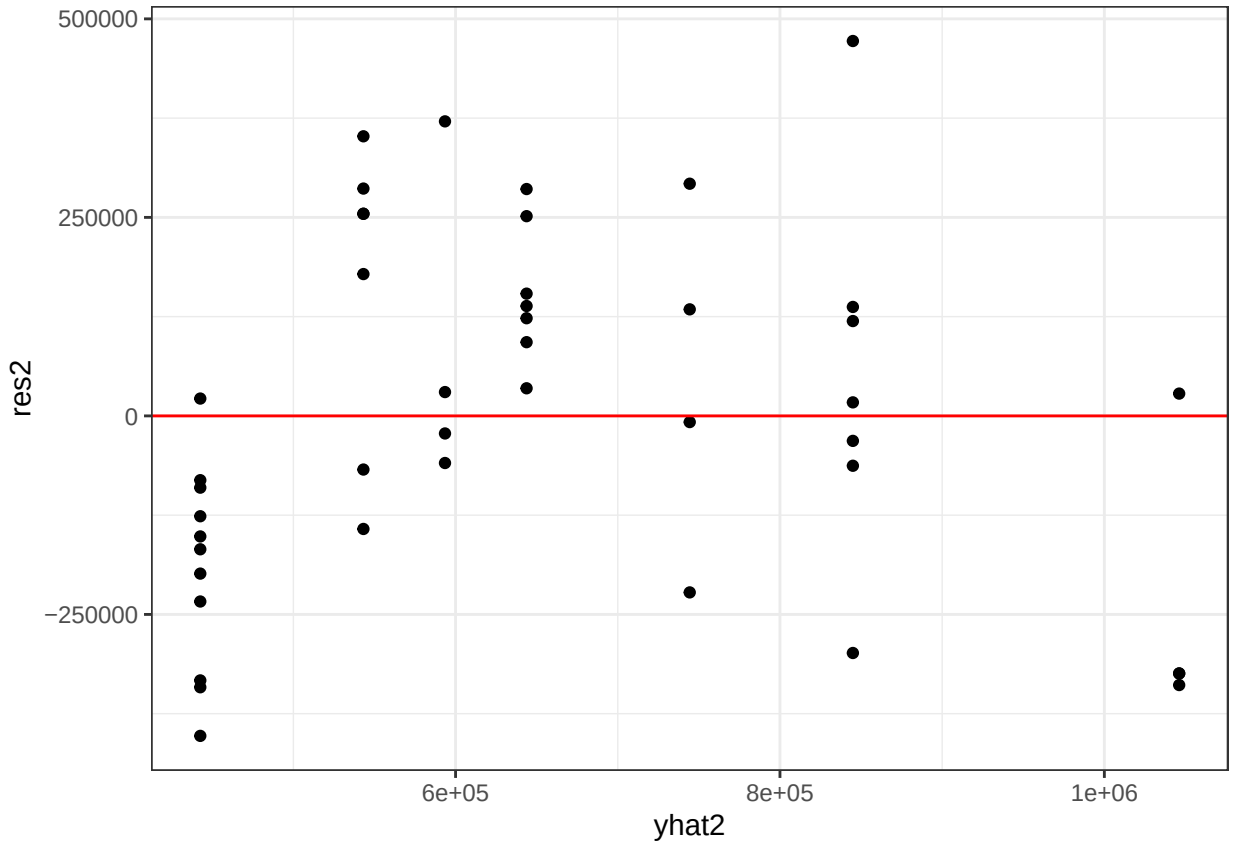
```
boxcox(result, lambda = seq(2, 3.5, 1/10))
```



From the boxcox plot above we can try a transformation lambda value between 2.5 and 3, which 2.75 will be selected for the transformation.

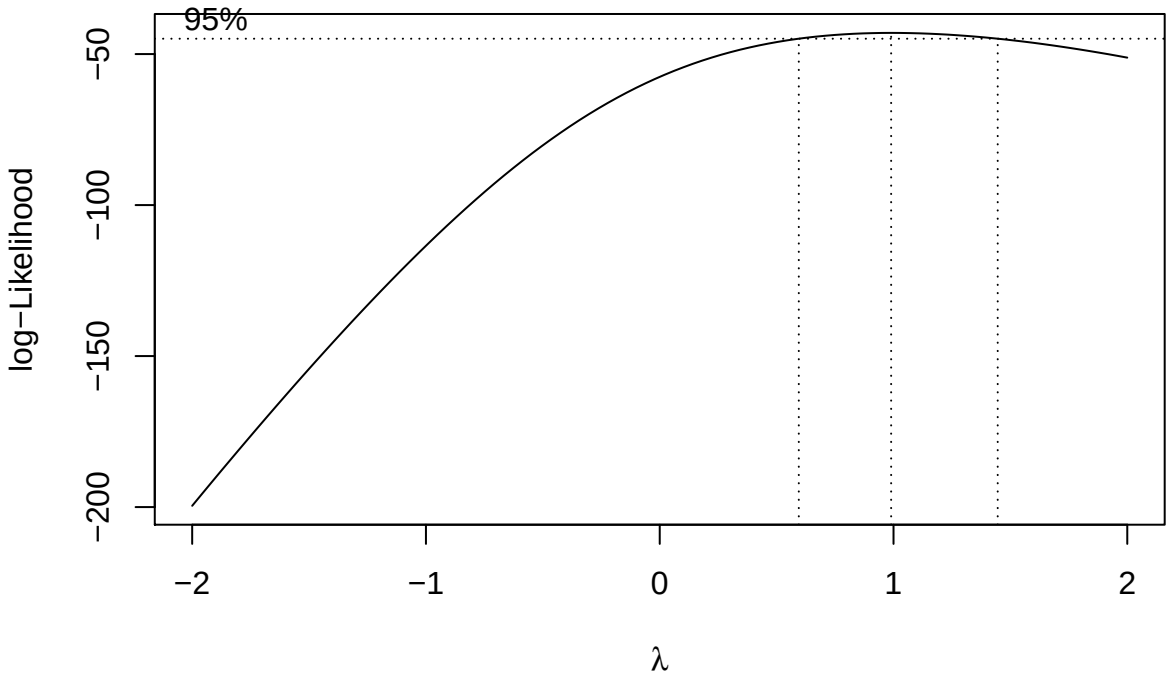
D Perform the necessary transformation to the data. Re fit the regression with the transformed variable(s) and assess the regression assumptions. You may have to apply transformations a number of times. Be sure to explain the reason behind each of your transformations. Perform the needed transformations until the regression assumptions are met. What is the regression equation that you will use?

```
temp <- data
temp$yield <- temp$yield^2.75
result2 <- lm(yield~nitrogen, temp)
yhat2 = result2$fitted.values
res2 = result2$residuals
data %>%
  ggplot(aes(yhat2, res2)) + geom_point() + theme_bw() + geom_hline(yintercept=0, color="red")
```



Lets double check the boxcox.

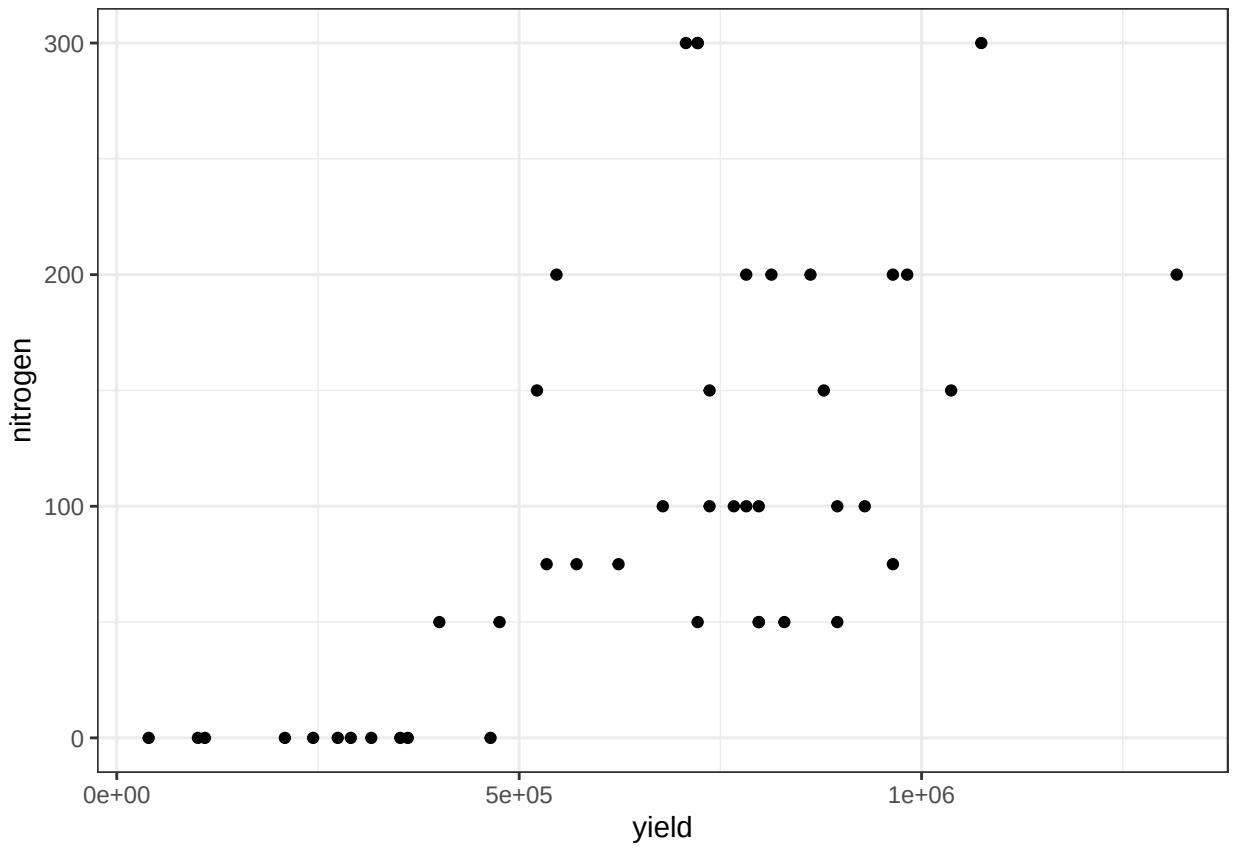
```
boxcox(result2, lambda = seq(-2,2, 1/10))
```



Based on the boxcox plot above, we do not need to tranform the response variable since 1 is between the 95% CI for the lambda value.

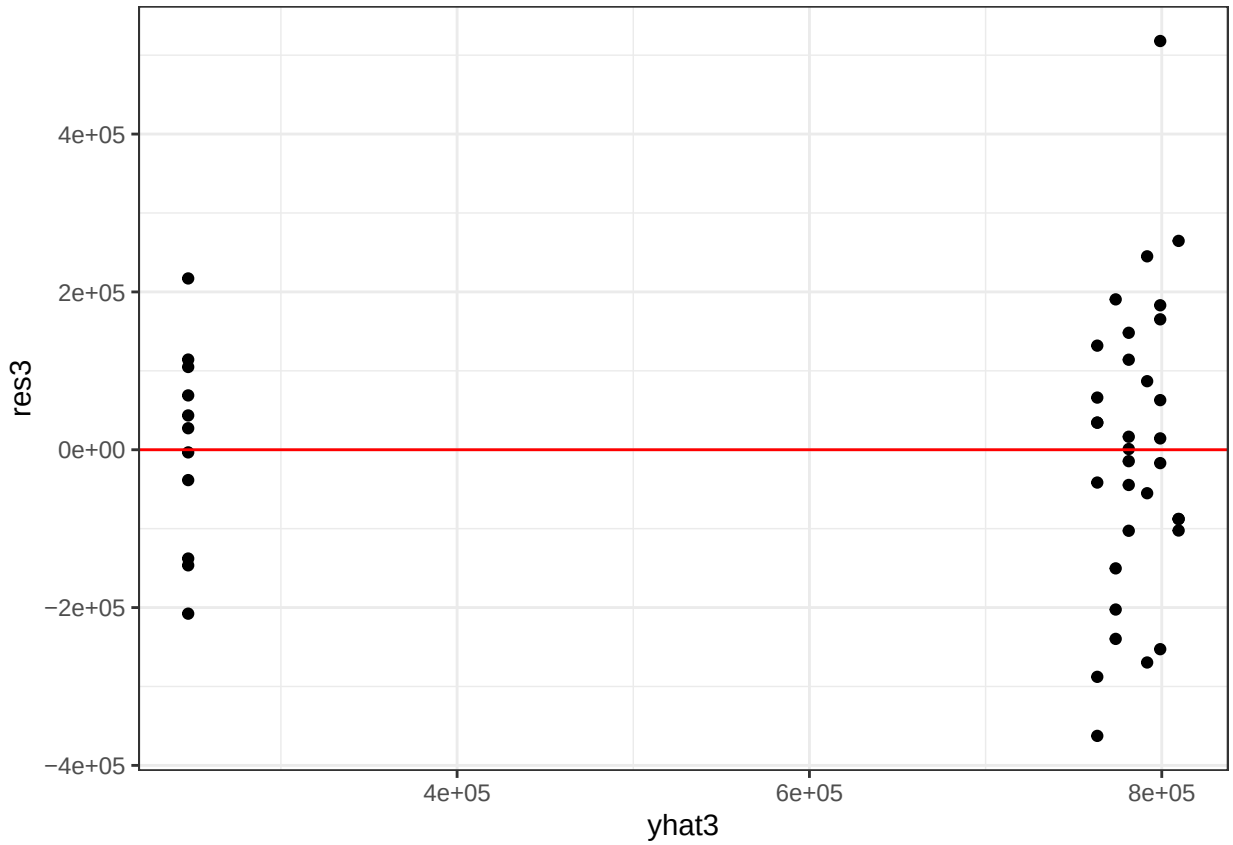
However, let's take a look at the scatter plot after the response variable transformation.

```
temp %>%
  ggplot(aes(yield, nitrogen)) + geom_point() + theme_bw()
```



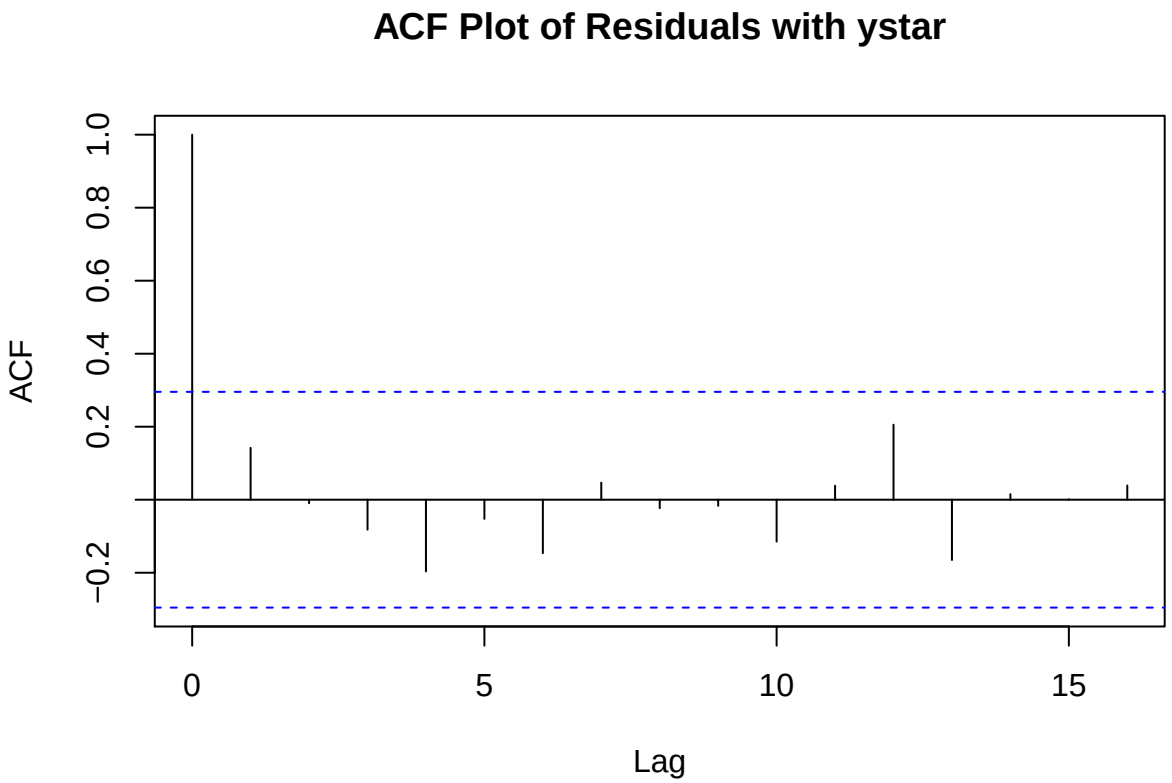
Based on this scatter plot in order for this to follow a linear relationship we can produce a transformation of the predictor variable. We will utilize a log transformation. However, since some values for the predictor are 0, log function will return undefined. We therefore, can add a very small constant to shift those values. Let us use 0.0000001, which is a small enough constant.

```
temp <- data
temp$yield <- temp$yield^2.75
temp$nitrogen <- log(temp$nitrogen + 0.0000001)
result3 <- lm(yield~nitrogen, temp)
yhat3 = result3$fitted.values
res3 = result3$residuals
data %>%
  ggplot(aes(yhat3, res3)) + geom_point() + theme_bw() + geom_hline(yintercept=0, color="red")
```



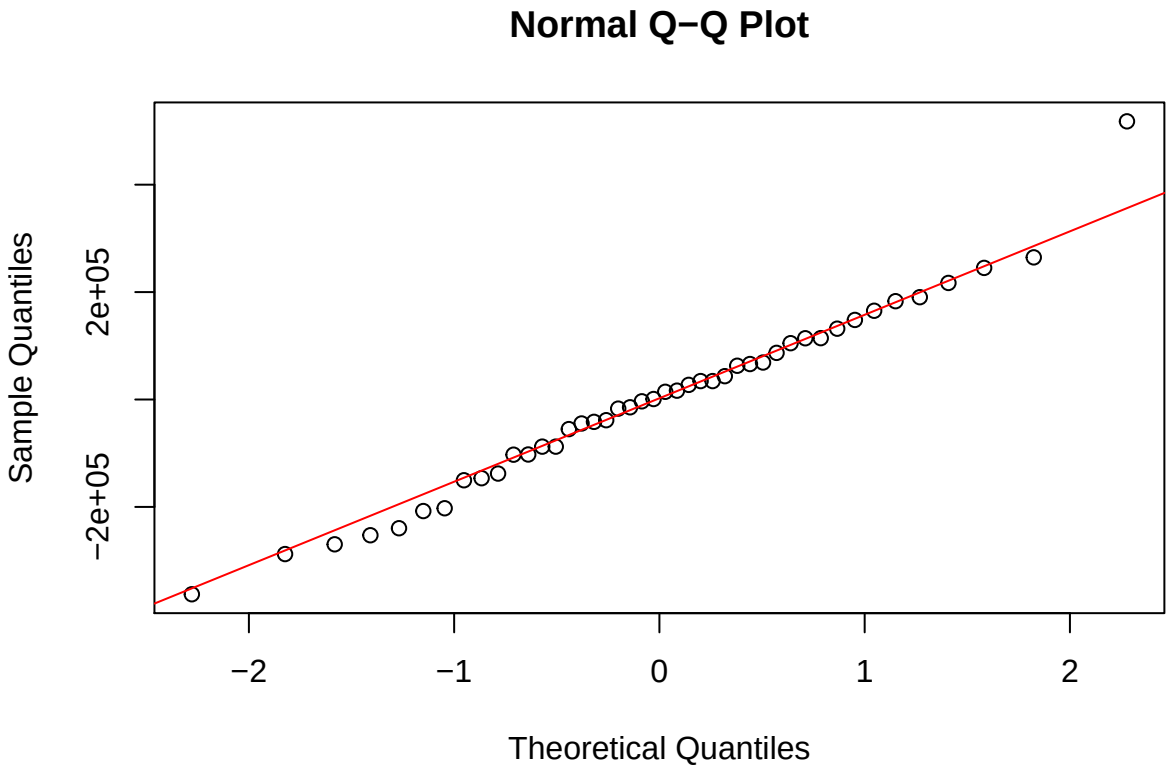
Considering the distribution along the x-axis, we can see that the above is more evenly scatter than the previous result plot with result2. Let's check the ACF and QQ plots for this last transformation

```
acf(result3$residuals, main="ACF Plot of Residuals with ystar")
```



Based on the ACF plot, the residuals are uncorrelated, so we don't have evidence that the errors are dependent.

```
{
qqnorm(result3$residuals)
qqline(result3$residuals, col="red")
}
```



The plots fall closely to the line, so the residuals follow a normal distribution.

```
result3
```

```
##
## Call:
## lm(formula = yield ~ nitrogen, data = temp)
##
## Coefficients:
## (Intercept)     nitrogen
##      662613        25763
```

The final equation is as follows.

$y^{2.75} = 25763 * log(x + 0.0000001) + 662613$

Question 3

A) Based only on Figure 1, would you recommend transforming the predictor, x, or the response, y, first? Briefly explain your choice.

Well, we can see through the scatter plot that the relationship doesn't seem to be linear. Also, from the residual plot we can see that the constant variance assumption is not met. Because of these reasons, transformation the response variable will be first. If after transforming the response y we still cannot meet the assumptions and the relationship still seems to be non-linear then at that point we can perform a transformation of the x variable.

B) The profile log-likelihoods for the parameter, λ, of the Box-Cox power transformation, is shown in Figure 2. Your classmate says that you should apply a log transformation to the response variable first. Do you agree with your classmate? Be sure to justify your answer.

I agree, basedon the Boxcox plot we can see that the lambda's 95% CI is between ~0.1 and 0.1. THe middle seems to be near 0.0, which means a log transformation is appropriate for the response variable.

C) Your classmate is adament on applying the log transformation to the response variable, and fits the regression model. The R output is shown in Figure 3. Write

down the estimated regression equation for this model. How do we interpret the regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_0$ in context?

The regression equation will be $log(y) = -0.44993x + 1.50792$

The $\hat{\beta}_1$ means that for every increase of time the concentration increases by $\epsilon^{-0.449933}$. The $\hat{\beta}_0$ means taht when the time is 0, the predicted concentration is $\epsilon^{1.5079}$