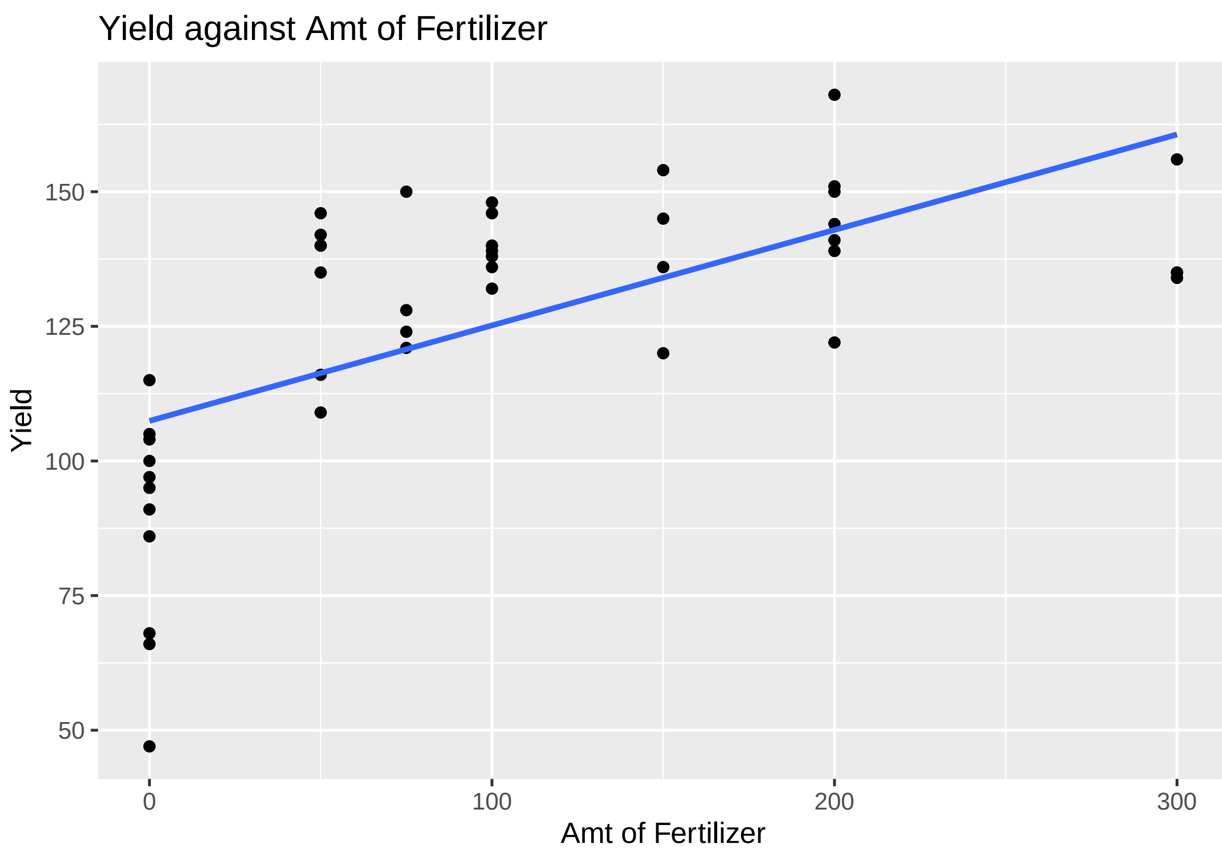
HW 5 Q 1 Solutions

```
library(faraway)
library(MASS)
library(tidyverse)
Data<-faraway::cornnit
```

(a)

The response is the yield of corn. The predictor is the amount of nitrogen fertilizer applied. The scatterplot is displayed below.

```
##scatter plot
ggplot2::ggplot(Data, aes(x=nitrogen,y=yield))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Amt of Fertilizer", y="Yield",
       title="Yield against Amt of Fertilizer")
```



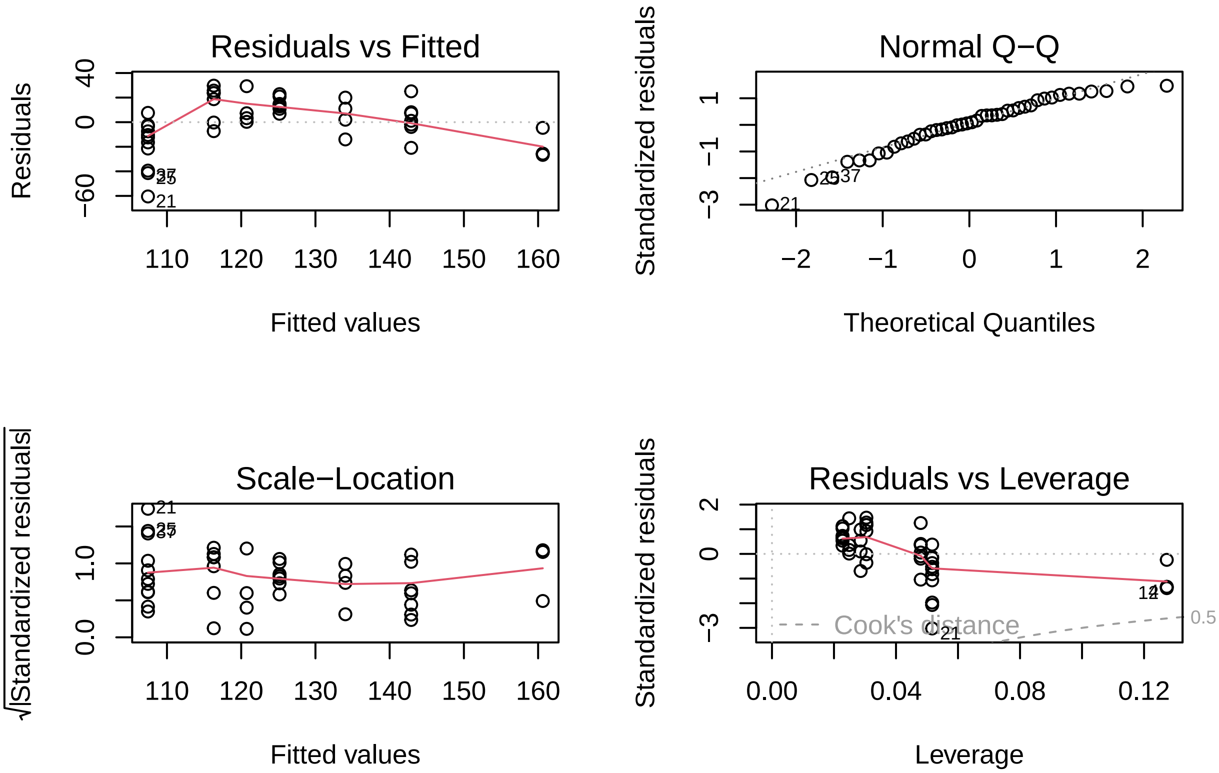
There appears to be a curved relationship between the amout of fertilizer applied and the yield of corn (not linear).

(b)

The residual plot without any data transformations is displayed below.

```
##fit regression
result<-lm(yield~nitrogen, data=Data)

##create residual plot
par(mfrow=c(2,2))
plot(result)
```



Two things to notice:

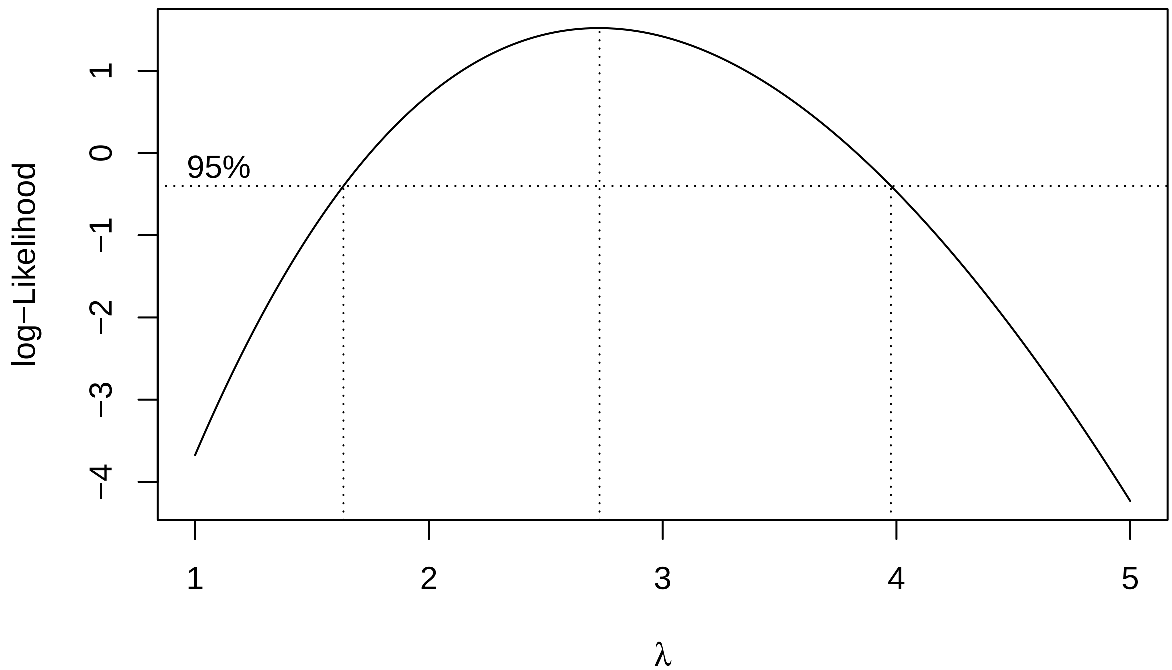
- The variance of the residuals is not constant. The variance appears to be decreasing for higher fitted values.
- The residuals are not evenly scattered across the horizontal axis, indicating a non-linear relationship between the variables.

When both of these issues are present, we seek to stabilize the variance first by transforming the response variable first.

Note: There are a number of different transformations that will work. I am showing only one possibility. The most important thing is to provide a reason for each of your transformations.

(c)

```
MASS::boxcox(result,lambda=seq(1,5,by=0.01))
```



Based on the Box Cox plot, raising the response variable to power of a value between slightly less than 2 and 4 should work. To keep things simple, I will raise the response variable to the power of 2.

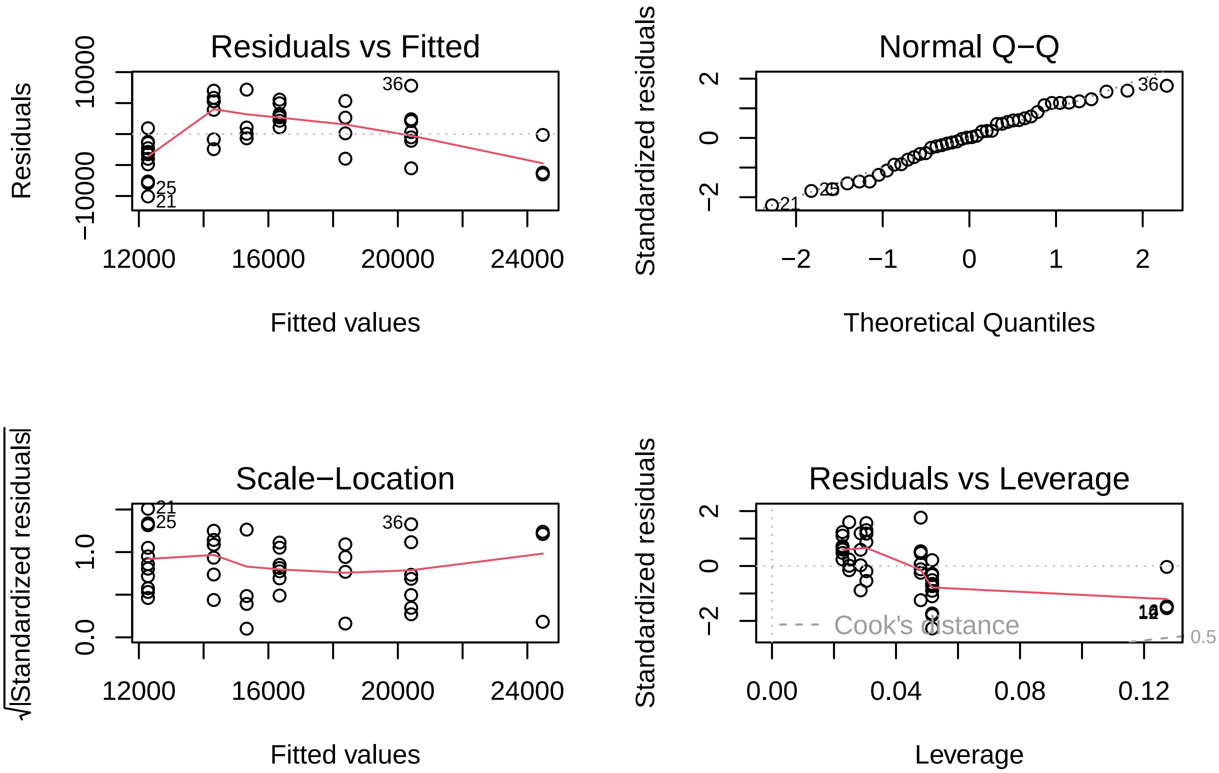
(d)

Regress y^2 against x, and create the corresponding residual plot.

```
##transform y
Data$newy<-(Data$yield)^2

##regress newy against x
result2<-lm(newy~nitrogen, data=Data)

##create residual plot
par(mfrow=c(2,2))
plot(result2)
```

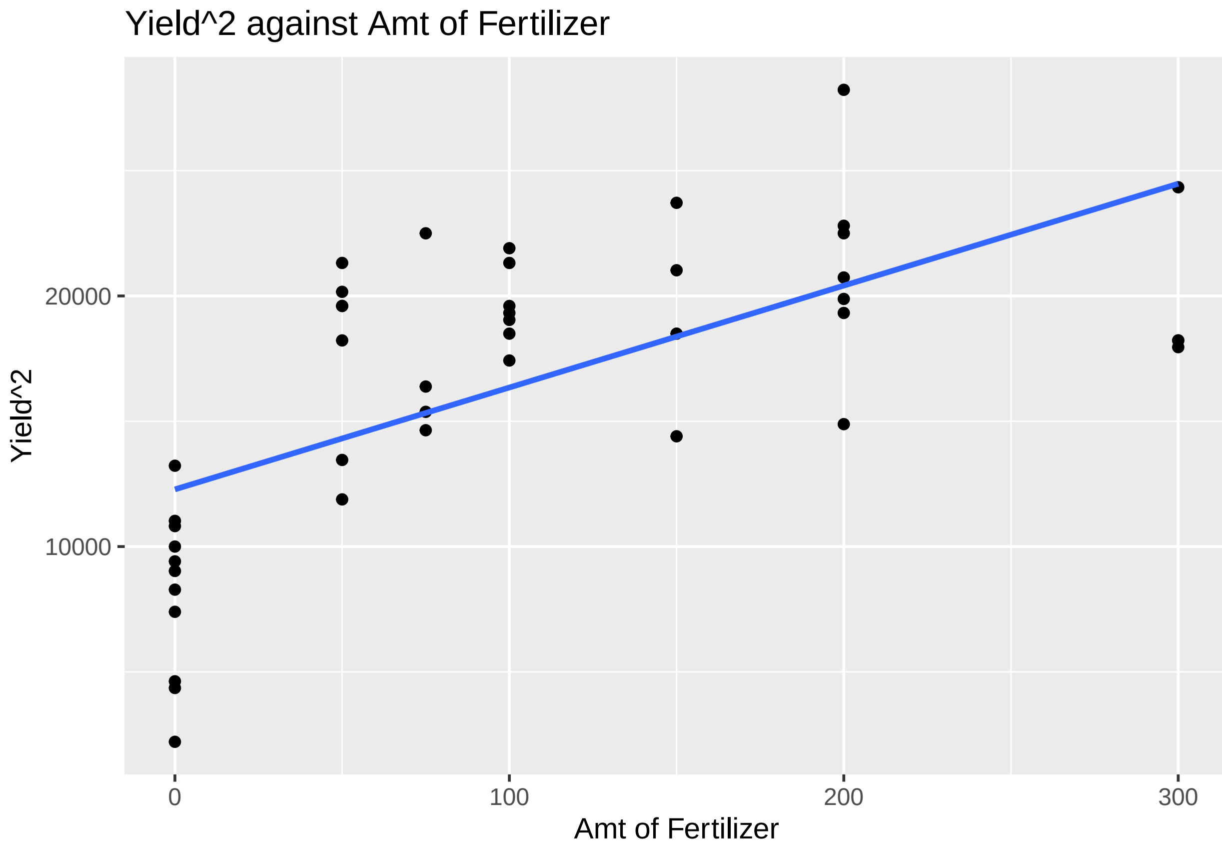


Two things to notice from the residual plot:

- The variance of the residuals is a lot more constant. The Box Cox plot indicates we no longer need to transform the response variable since 1 lies within the 95% CI.
- The residuals are not evenly scattered across the horizontal axis, indicating a non-linear relationship between the variables.

We now need to consider a transformation to the predictor. To decide how to transform the predictor, we create a scatterplot of y^2 against x.

```
##scatter plot
ggplot2::ggplot(Data, aes(x=nitrogen,y=newy))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Amt of Fertilizer", y="Yield^2",
       title="Yield^2 against Amt of Fertilizer")
```



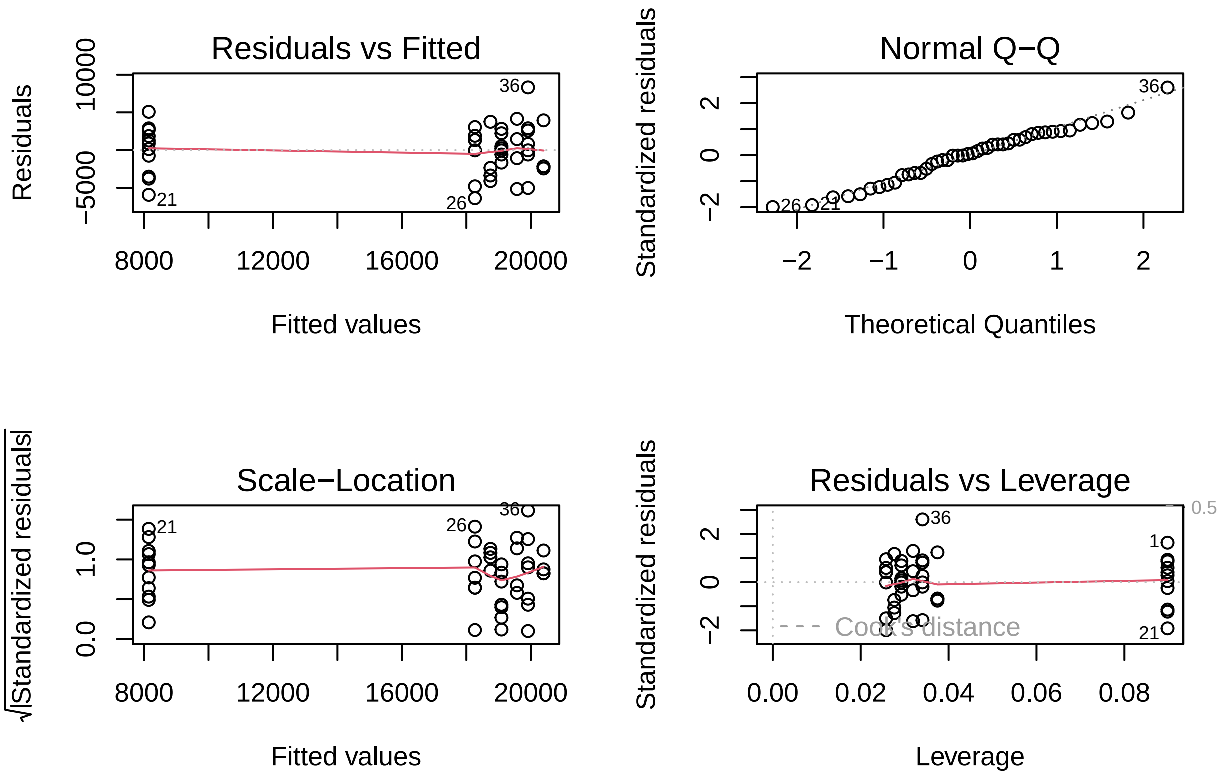
Based on the shape of the scatterplot, we consider a log transformation to the predictor. However, we note that some observations have a value of 0 for the predictor, so we add a small constant, 0.01, first to the predictor and then apply a log transformation to the predictor. So let $x^* = \log(x + 0.01)$.

Next, we regress y^2 against x^* and create the corresponding residual plot.

```
Data$newx<-log(Data$nitrogen+0.01)
##add a small constant since some values of x are 0.

result3<-lm(newy~newx, data=Data)

##create residual plot
par(mfrow=c(2,2))
plot(result3)
```



Two things to notice from the residual plot:

- The variance of the residuals is a lot more constant. The Box Cox plot indicates we no longer need to transform the response variable since 1 lies within the 95% CI.
- The residuals are evenly scattered across the horizontal axis.

We no longer need to transform the variables.

The assumptions are met after letting $y^* = y^2$ and $x^* = \log(x + 0.01)$.

```
summary(result3)
```

```
##
## Call:
## lm(formula = newy ~ newx, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6385.2 -2218.6   183.6  2316.4  8310.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13617.4      568.5  23.954  < 2e-16 ***
## newx          1188.3      119.8   9.918 1.44e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3247 on 42 degrees of freedom
## Multiple R-squared:  0.7008, Adjusted R-squared:  0.6936
## F-statistic: 98.36 on 1 and 42 DF,  p-value: 1.435e-12
```

The regression equation is now

$$y^* = 13617.4 + 1188.3x^*$$

where $y^* = y^2$ and $x^* = \log(x + 0.01)$.