Module03 Homework

Diana McSpadden

9/10/2020

Assignment: Stat 6021: Homework Set 3

Name: H. Diana McSpadden

UID: hdm5s

Attended a Study Group With: David Fuentes, Jing Huang, Caprill Wright, Arne Newman, Michael Kastanowski, Abby Bernhardt, Loren Bushkar

References

In addition to the textbook and Module 3, I also read:
https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

Question 1 (No R required)

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?"

Answer A lack of constant variance means that as x changes, the difference between The predicted y and actual y is increasing, decreasing, or displaying some other pattern, i.e. the variance of e|x changes as x changes. Transforming x would not have an affect on the lack of constant (yi - yi-hat) , because x would still be changing values even with the transformation, the variance is still affected.

I would also show them the formula for the variance of y given x:

Var(Y|x) = Var(B0 + B1(x) + e|x)

I would explain that both B0 and B1 are fixed values, and y's non-stable variance is based on x.

V(y|x) = Var(CONSTANT + e|x)

Var(y|x) = Var(e|x) == sigma^2

The only way to create a constant sigms^2 is to transform y.

Therefor, to fix issues with constant variance, we transform y to provide a constant/fixed sigma^2.

Transforming x will not actually transform the variance of y.

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.

```
library(faraway)
#summary(cornnit)

head(cornnit, 5)
```

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

```
attach(cornnit)

cornnit
```

```
##    yield nitrogen
## 1    115        0
## 2    128       75
## 3    136      150
## 4    135      300
## 5     97        0
## 6    150       75
## 7    154      150
## 8    156      300
## 9     95        0
## 10   121       75
## 11   120      150
## 12   134      300
## 13    91        0
## 14   124       75
## 15   145      150
## 16   135      300
## 17   105        0
## 18   140       50
## 19   138      100
## 20   139      200
## 21    47        0
## 22   140       50
## 23   132      100
## 24   151      200
```

```
## 25    66        0
## 26   109       50
## 27   136      100
## 28   144      200
## 29    86        0
## 30   135       50
## 31   139      100
## 32   150      200
## 33   100        0
## 34   146       50
## 35   148      100
## 36   168      200
## 37    68        0
## 38   116       50
## 39   146      100
## 40   122      200
## 41   104        0
## 42   142       50
## 43   140      100
## 44   141      200
```

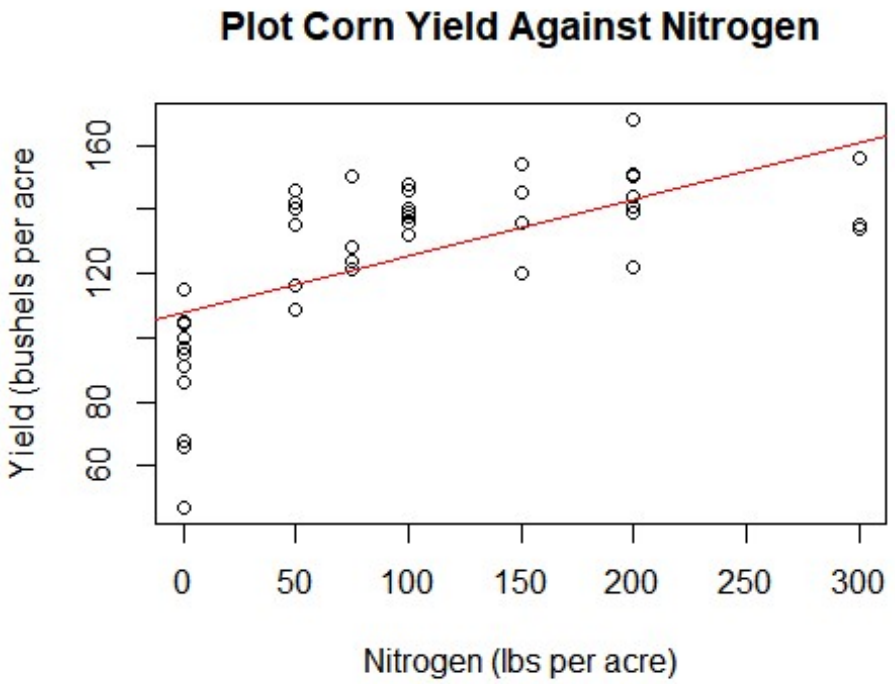
2a:What is the response variable and predictor for this study?

Answer The predictor variable is **nitrogen**

The response variable is **yield**

Create a **scatterplot** of the data, and interpret the scatterplot.

```
# create a model
cornModel <- lm(yield~nitrogen)

#plot the model and the linear model
plot(x=nitrogen, y=yield, main='Plot Corn Yield Against Nitrogen', xlab =
'Nitrogen (lbs per acre)', ylab = 'Yield (bushels per acre')
abline(cornModel,col="red")
```



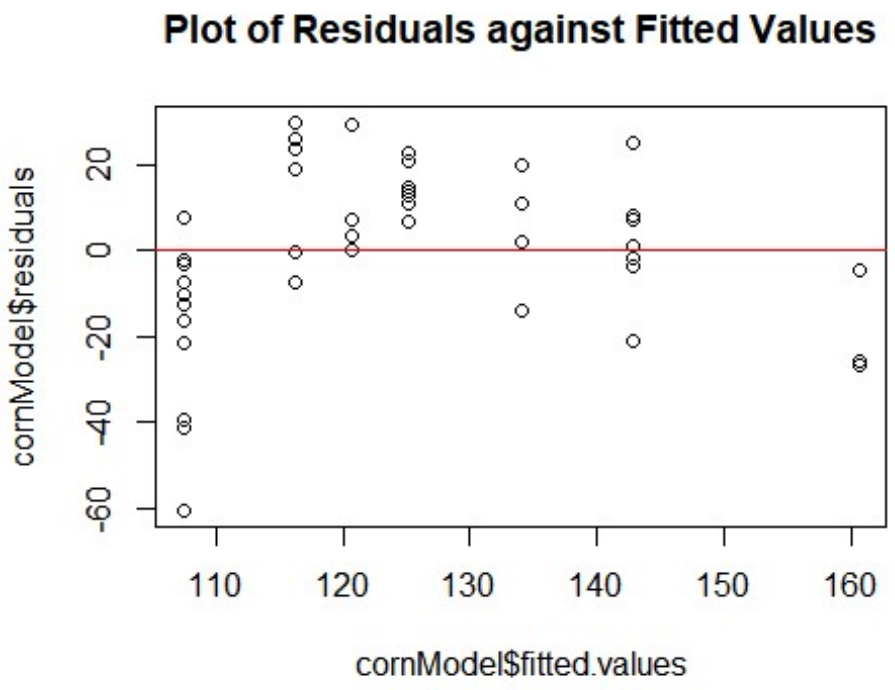
Interpretation

While the yield appears to increase as nitrogen increases, the points are not evenly spread above and below the linear model, and the variance appears smaller at the 100 and 300 nitrogen levels. I need to see more.

2b: Fit a linear regression without any transformations.

Create the corresponding residual plot.

```
plot(cornModel$fitted.values,cornModel$residuals, main="Plot of Residuals
against Fitted Values")
abline(h=0,col="red")
```



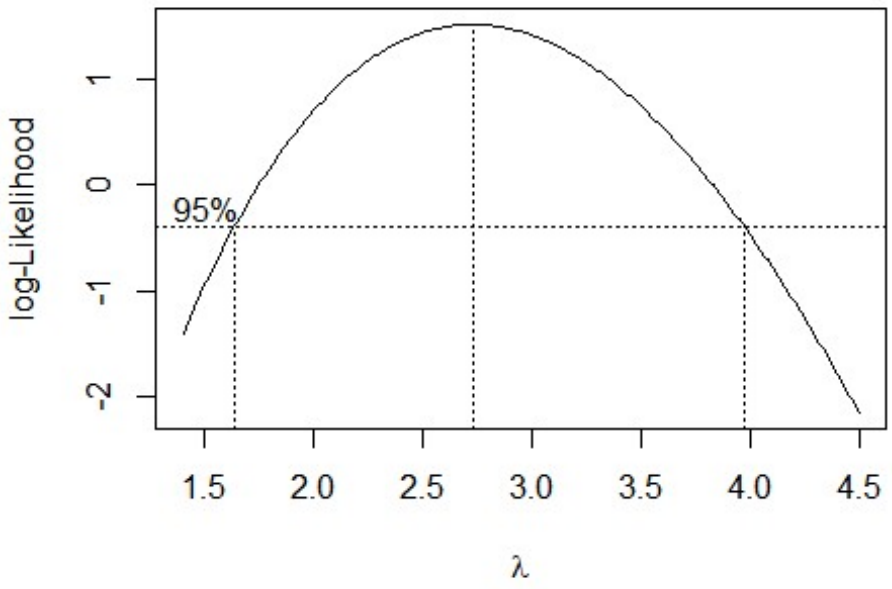
Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason. Answer

The residual plot both displays a lack of standard variance in the residuals and a non-0 mean of the residuals. I recommend a transformation on the response variable first, to attempt to address the variance issue, because it may also address the non-zero mean issue.

I recommend reviewing the Box-Cox to determine the correct y transformation.

2c: Create a Box Cox plot for the profile log likelihoods.

```
library(MASS)
#boxcox(cornModel)
boxcox(cornModel, lambda = seq(1.4, 4.5, 0.01))
```



How does this plot aid in your data transformation? 0 does not lie within the confidence interval, thus, we do not recommend a ln transformation.

1 does not lie within the confidence interval, thus I recommend applying a transformation to the response variable: yield.

2d: Perform the necessary transformation to the data.

Refit 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?

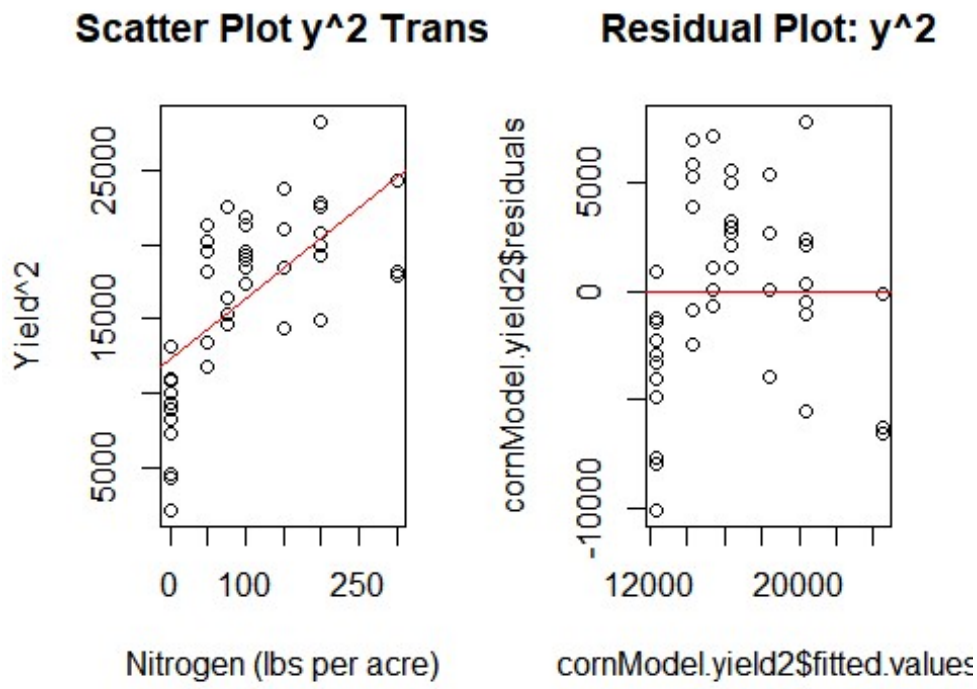
Note: in part 2d, there are a number of solutions that will work. You must clearly document your reasons for each of your transformations.

First, I am going to start with the variance issue, because it may address the non-zero mean issue.

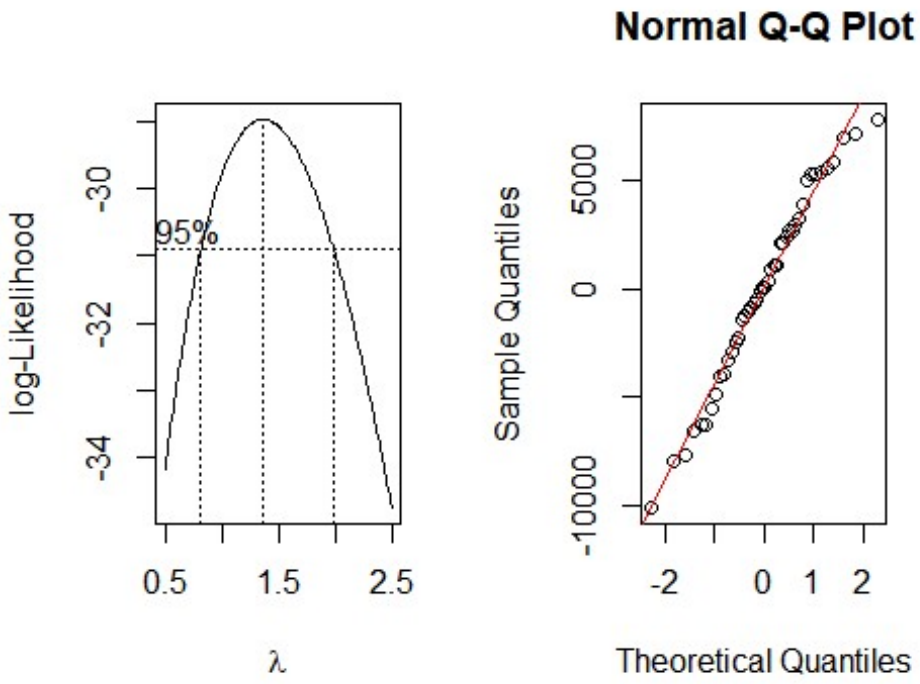
Based on the Box-Cox plot I will select a lambda of 2, which if within the 95% CI for lambda. I am using lambda 2 because it is easier for me to conceptualize than other choices within the CI.

```
# y' = y^2
yield2 <- yield^2
cornModel.yield2 <-lm(yield2~nitrogen)
```

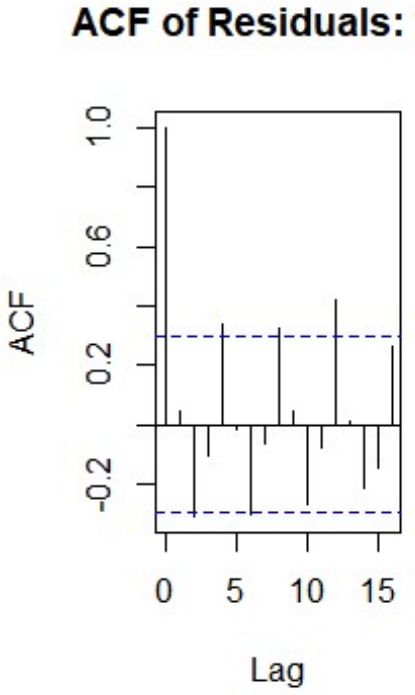
```r
par(mfrow=c(1,2)) # 1 rows, 2 columns - there will be three panes to look through

# scatter plot transformed y
plot(x=nitrogen, y=yield2, main='Scatter Plot y^2 Trans', xlab = 'Nitrogen (lbs per acre)', ylab = 'Yield^2')
abline(cornModel.yield2,col="red")

#residual plot of transformed y
plot(cornModel.yield2$fitted.values,cornModel.yield2$residuals,
main="Residual Plot: y^2")
abline(h=0,col="red")
```




```r
# boxcox of y^2 model
boxcox(cornModel.yield2, lambda = seq(0.5, 2.5, 0.01))

#QQ plot
qqnorm(cornModel.yield2$residuals)
qqline(cornModel.yield2$residuals, col="red")
```




```r
#ACF plot
acf(cornModel.yield2$residuals, main="ACF of Residuals: y^2")
```



Analysis of the residual plot: I am still not thrilled with the variance, and the non-0 mean problem still exists.

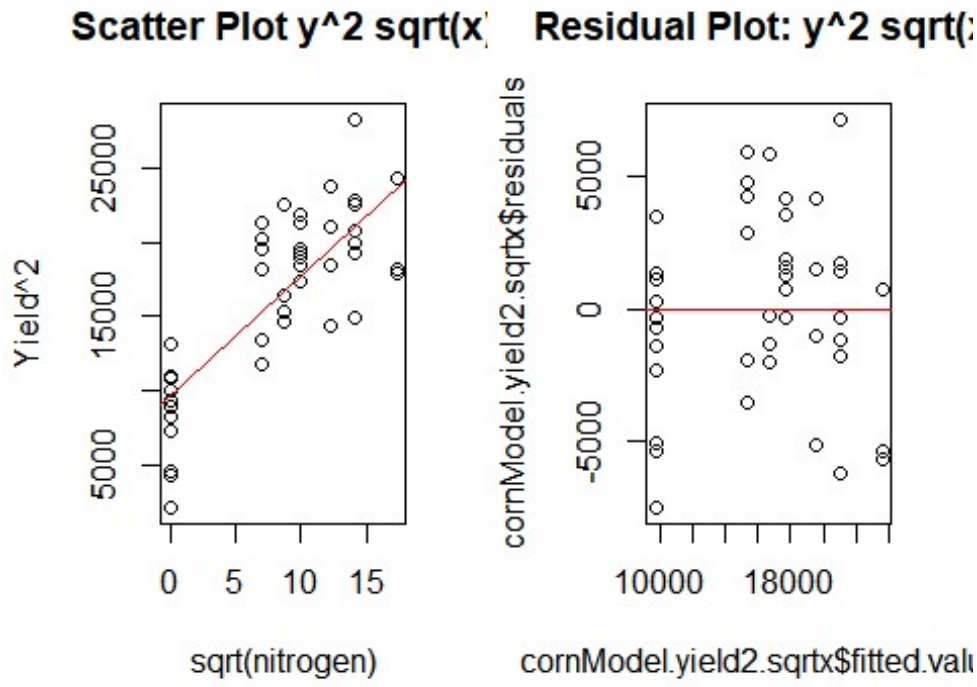
Analysis of the Box-Cox plot: 1 is within the CI for lambda so we don't have an obvious exponential transformation.

Analysis of the NOrmal Q-Q plot: the residuals do have a very nice normal distribution,

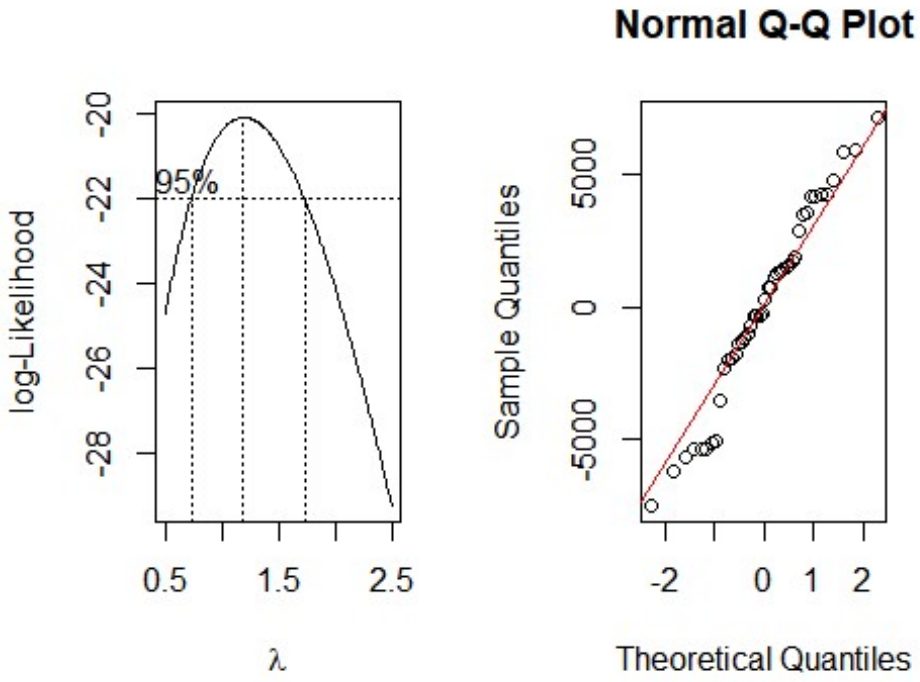
Analysis of the ACF plot: The residuals ARE correlated, this issue needs to be addressed.

I want to try to address two issue: the non-0 mean, and the correlation of the residuals. My next transformation will be of the predictor variable. My transformation is based on the appearance of the scatter plot, which I believe appears to have decreasing parabola curve, so I will use a sqrt(x) transformation.

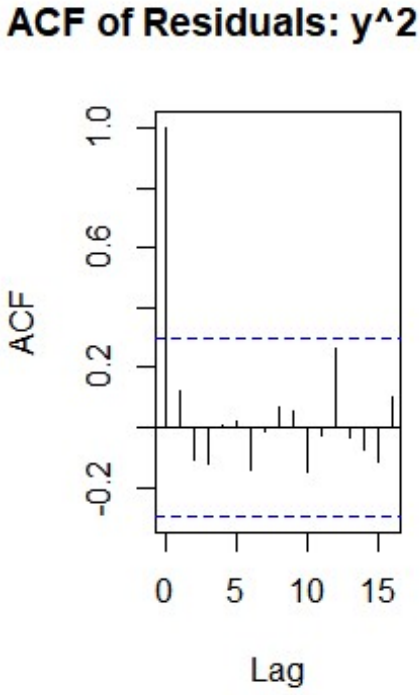
```
#x' = sqrt(x)

sqrt.nitrogen = sqrt(nitrogen)


cornModel.yield2.sqrtx <-lm(yield2~sqrt.nitrogen)


par(mfrow=c(1,2)) # 1 rows, 2 columns - there will be three panes to look
through

# scatter plot yield2, sqrt(x)
plot(x=sqrt.nitrogen, y=yield2, main='Scatter Plot y^2 sqrt(x)', xlab =
'sqrt(nitrogen)', ylab = 'Yield^2')
abline(cornModel.yield2.sqrtx,col="red")

#residual plot yield2, sqrt(x)
plot(cornModel.yield2.sqrtx$fitted.values,cornModel.yield2.sqrtx$residuals,
main="Residual Plot: y^2 sqrt(x)")
abline(h=0,col="red")
```




```r
# boxcox of y^2 model, sqrt(x)
boxcox(cornModel.yield2.sqrtx, lambda = seq(0.5, 2.5, 0.01))

#QQ plot of y^2 model, sqrt(x)
qqnorm(cornModel.yield2.sqrtx$residuals)
qqline(cornModel.yield2.sqrtx$residuals, col="red")
```




```
#ACF plot of y^2 model, sqrt(x)
acf(cornModel.yield2.sqrtx$residuals, main="ACF of Residuals: y^2 sqrt(x)")
```



The y^2 and sqrt(x) transformations successfully address the non-zero mean of the residuals, but also

results in two clusters. The clusters represent the data results where no nitrogen was applied to the field, and where nitrogen was applied to the field.

The variance of the residual values for the group where nitrogen was applied seems close to constant. The nitrogen=0 cluster has a different variance than the other cluster, which also leads me to want to filter the nitrogen = 0 values from the data set.

The Box-Cox plot still has 1 within the CI, indicating we do not need an exponential transformation of y.

The QQ plot shows normal distribution of residuals.

The sqrt(x) transformation addressed the correlation of residuals and they are now uncorrelated.

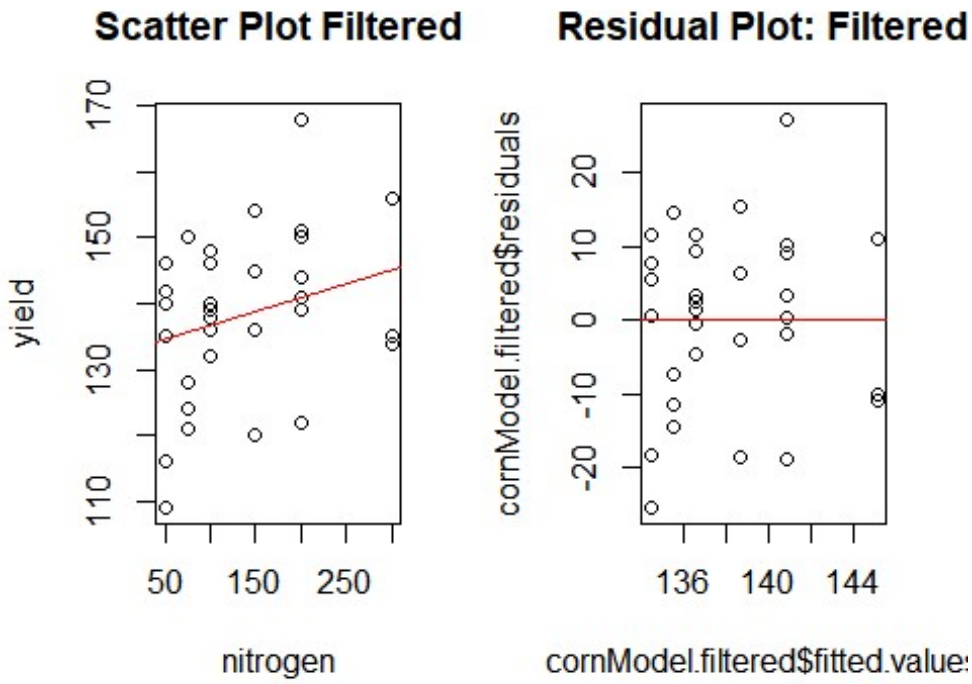
Other than the clustering I am pleased with the model.

To address the clusters, first, create a new data set that has removed the x = 0 values.

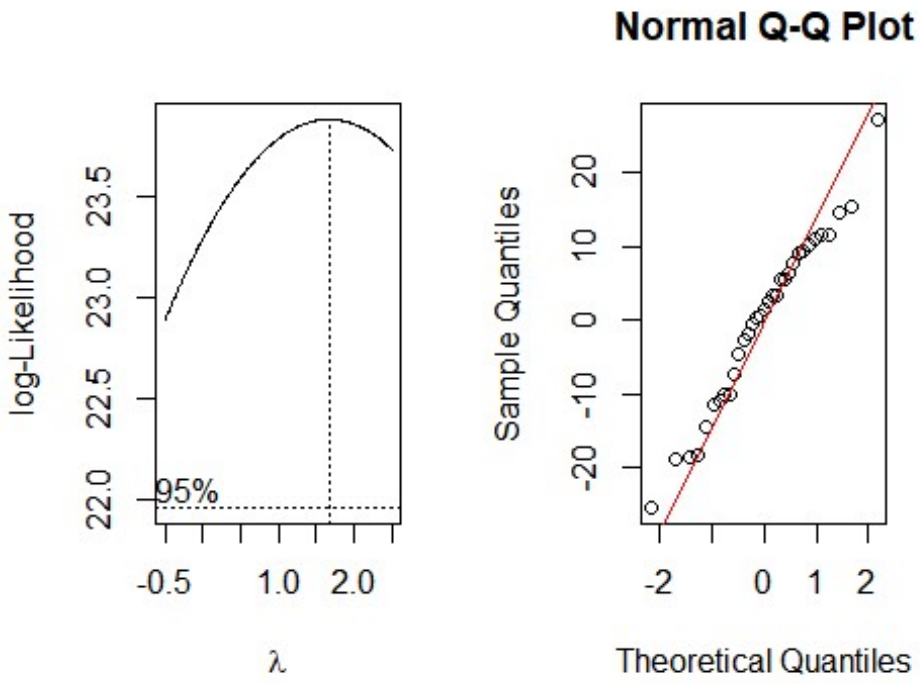
```
#head(cornnit)
cornnitFiltered = cornnit[cornnit[, "nitrogen"]>0, ]
```

First determine what transformations I would like to make on the cornnitFiltered data set by creating various plots

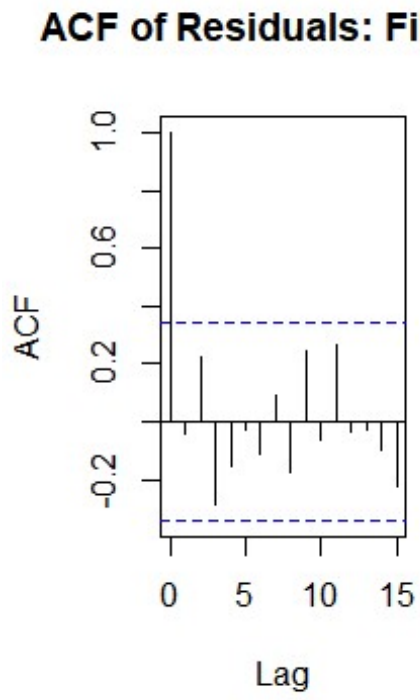
```
detach(cornnit)
attach(cornnitFiltered)

cornModel.filtered <-lm(yield~nitrogen)

par(mfrow=c(1,2)) # 1 rows, 2 columns - there will be three panes to look
through

# scatter plot filtered model
plot(x=nitrogen, y=yield, main='Scatter Plot Filtered', xlab = 'nitrogen',
ylab = 'yield')
abline(cornModel.filtered,col="red")

#residual plot filtered model
plot(cornModel.filtered$fitted.values,cornModel.filtered$residuals,
main="Residual Plot: Filtered")
abline(h=0,col="red")
```




```r
# boxcox of filtered model
boxcox(cornModel.filtered, lambda = seq(-0.5, 2.5, 0.01))

#QQ plot of filtered model
qqnorm(cornModel.filtered$residuals)
qqline(cornModel.filtered$residuals, col="red")
```




```r
#ACF plot
acf(cornModel.filtered$residuals, main="ACF of Residuals: Filtered")
```



Analysis of the filtered scatter and residual plot: The residual variance is not standard; however the zero mean of the residuals holds.

Analysis of the Box-Cox plot: 1 is within the CI for lambda so we don't have an obvious exponential transformation.

Analysis of the NOrmal Q-Q plot: the residuals do have a very nice normal distribution,

Analysis of the ACF plot: The residuals are uncorrelated.

What does the ANOVA table tell us about these data?

```
anova(cornModel.filtered)
```

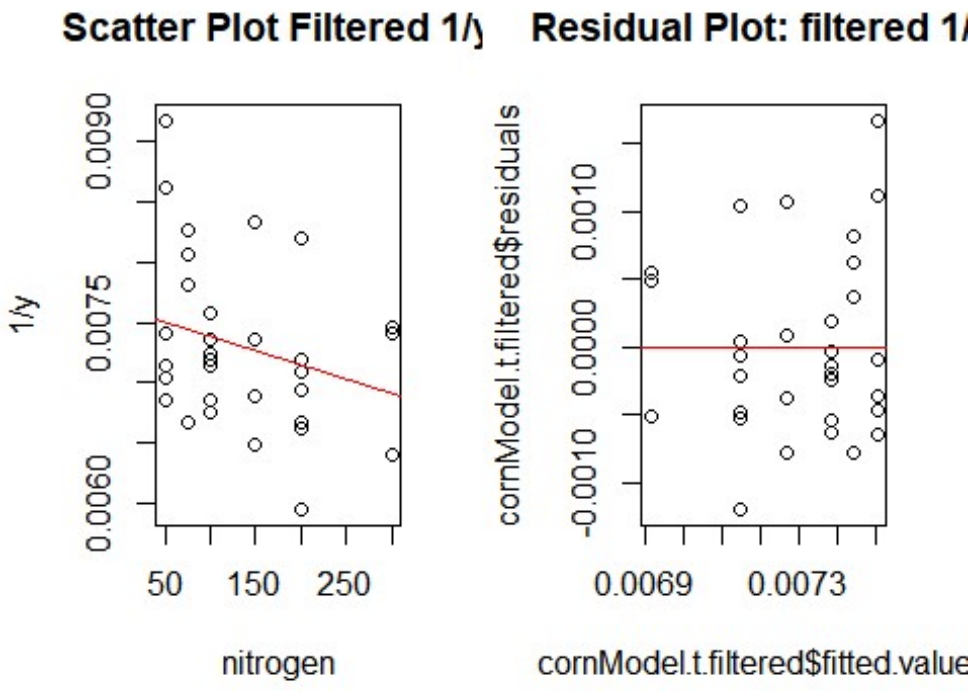
```
## Analysis of Variance Table
##
## Response: yield
##           Df Sum Sq Mean Sq F value Pr(>F)
## nitrogen   1    385  384.96  2.6638 0.1128
## Residuals 31   4480  144.51
```

Our p-value is not significant from the ANOVA table. The y needs a transformation to correct the variance, but because 1 is within the CI of lambda there is not an obvious exponential transformation.

x does not need a transformation because the mean of the residuals is 0.

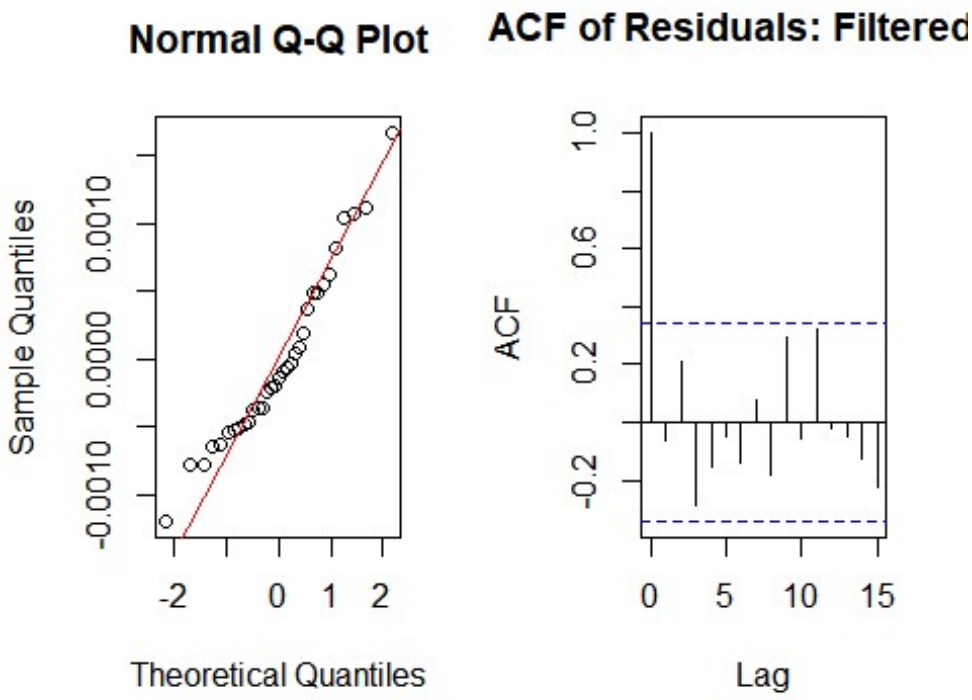
As a final effort to create a meaningful linear model from the filtered data I will apply an inverse transformation to y because in the original plot there seemed to be a plateau curve, but without a y limit.

```
yield.t.filtered = 1/yield

cornModel.t.filtered <-lm(yield.t.filtered~nitrogen)

par(mfrow=c(1,2)) # 1 rows, 2 columns - there will be three panes to look through

# scatter plot filtered transformed
plot(x=nitrogen, y=yield.t.filtered, main='Scatter Plot Filtered 1/y', xlab = 'nitrogen', ylab = '1/y')
abline(cornModel.t.filtered,col="red")

#residual plot filtered transformed
plot(cornModel.t.filtered$fitted.values,cornModel.t.filtered$residuals, main="Residual Plot: filtered 1/y")
abline(h=0,col="red")
```




```
#QQ plot of filtered transformed
qqnorm(cornModel.t.filtered$residuals)
qqline(cornModel.t.filtered$residuals, col="red")
```

```r
#ACF plot
acf(cornModel.t.filtered$residuals, main="ACF of Residuals: Filtered 1/y")
```




```r
anova(cornModel.t.filtered)
```

```
## Analysis of Variance Table
##
## Response: yield.t.filtered
##           Df     Sum Sq    Mean Sq F value Pr(>F)
## nitrogen   1 1.1748e-06 1.1748e-06  2.6793 0.1118
## Residuals 31 1.3593e-05 4.3847e-07
```

After trying several transformations on yield (above I show the inverse y transformation results), the model displays more variance in the residuals than I would expect; however, the mean of residuals issue has been addressed. There are no longer two clusters. The residuals are uncorrelated.

I will check the summary and anova tables for the filtered transformed:

```r
summary(cornModel.t.filtered)
```

```
##
## Call:
## lm(formula = yield.t.filtered ~ nitrogen)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0011967 -0.0004825 -0.0001378  0.0004933  0.0016726
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.619e-03  2.291e-04  33.258   <2e-16 ***
## nitrogen    -2.350e-06  1.436e-06  -1.637    0.112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0006622 on 31 degrees of freedom
## Multiple R-squared:  0.07955,    Adjusted R-squared:  0.04986
## F-statistic: 2.679 on 1 and 31 DF,  p-value: 0.1118
```

```
anova(cornModel.t.filtered)
```

```
## Analysis of Variance Table
##
## Response: yield.t.filtered
##           Df     Sum Sq    Mean Sq F value Pr(>F)
## nitrogen   1 1.1748e-06 1.1748e-06  2.6793 0.1118
## Residuals 31 1.3593e-05 4.3847e-07
```

I also want to review the anova table for the unfiltered y^2 sqrt(x) model:

```
summary(cornModel.yield2.sqrtx)
```

```
##
## Call:
## lm(formula = yield2 ~ sqrt.nitrogen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7497.3 -1951.6     8.3  2107.3  7160.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9706.27     993.28   9.772 2.22e-12 ***
## sqrt.nitrogen   803.07      97.68   8.222 2.75e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3674 on 42 degrees of freedom
## Multiple R-squared:  0.6168, Adjusted R-squared:  0.6076
## F-statistic:  67.6 on 1 and 42 DF,  p-value: 2.75e-10
```

```
anova(cornModel.yield2.sqrtx)
```

```
## Analysis of Variance Table
##
## Response: yield2
##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## sqrt.nitrogen  1 912533289 912533289  67.596 2.75e-10 ***
```

```
## Residuals     42 566993931  13499856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

Surprisingly, the unfiltered y^2 sqrt(x) model has passes the hypothesis test. This model meets the assumptions necessary for a linear model. However, this model is difficult to express in a linearized function.

If we change the hypothesis to whether there is a linear relationship between yield and nitrogen applied, only when a non-zero amount of nitrogen is applied, then the 0-filtered model with the y^2 sqrt(x) transformations still displays excess variance in the residuals, and the F-statistic does not pass the hypothesis test.

With multiple transformations (y' = y^2 and x = sqrt(x)) I was able to create a model that meets the linear regression assumptions.

Question 3 (No R required)

A chemist studied the concentration of a solution, y, over time, x, by fitting a simple linear regression. The scatterplot of the dataset, and the residual plot from the regression model are shown in Figure 1.

3a: Based only on Figure 1, would you recommend transforming the predictor, x, or the response, y,first?

Briefy explain your choice.

Answer I recommend transforming the response variable first because by addressing the lack of constant variance of the residuals which I do see in the scatter plot. By transforming the response variable, we may be able to address the non-zero mean of the residuals.

3b: The profile log-likelihoods for the parameter, gamma, 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.

Answer Yes, I do agree with my classmate, because 0 is within the 95% CI for gamma, which, based on Module 3 instruction, indicates a natural log transformation will be helpful to reduce residual variance. Response variable variance was apparent in the residual plot in Figure 1.

3c: 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 B1-hat and B0-hat in context?

Answer After transformation, the regression coefficients are: B0-hat: 1.50792

B1-hat: -0.44993

The chemist is studying the concentration of a solution over time.

A ln transformation was applied to the response variable, y.

Interpretation of B1 after an ln transformation on y For each change in x, y changes by the multiple exp(-0.44993).

If we want to present this as a percentage:

y, the solution concentration, decreases by (exp(-0.44993) - 1) * 100 % == **decreases by 36.2327%** for every unit increase in x: time.

```
(exp(-0.44993) - 1) * 100
```

```
## [1] -36.23272
```

Interpretation of B0 after an ln transformation on y The y intercept is ln(B0). In our case, the y intercept is exp(1.50792) (e^1.50792), or 4.5173.

```
exp(1.50792)
```

```
## [1] 4.517325
```

When time is 0, the concentration is 4.5173.