

Module03HW

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 $(y_i - \hat{y}_i)$, 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 :

$$\text{Var}(Y|x) = \text{Var}(B_0 + B_1(x) + e|x)$$

I would explain that both B_0 and B_1 are fixed values, and y 's non-stable variance is based on x .

$$V(y|x) = \text{Var}(\text{CONSTANT} + e|x)$$

$$\text{Var}(y|x) = \text{Var}(e|x) = \sigma^2$$

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

Therefor, to fix issues with constant variance, we transform y to provide a constant/fixed σ^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**

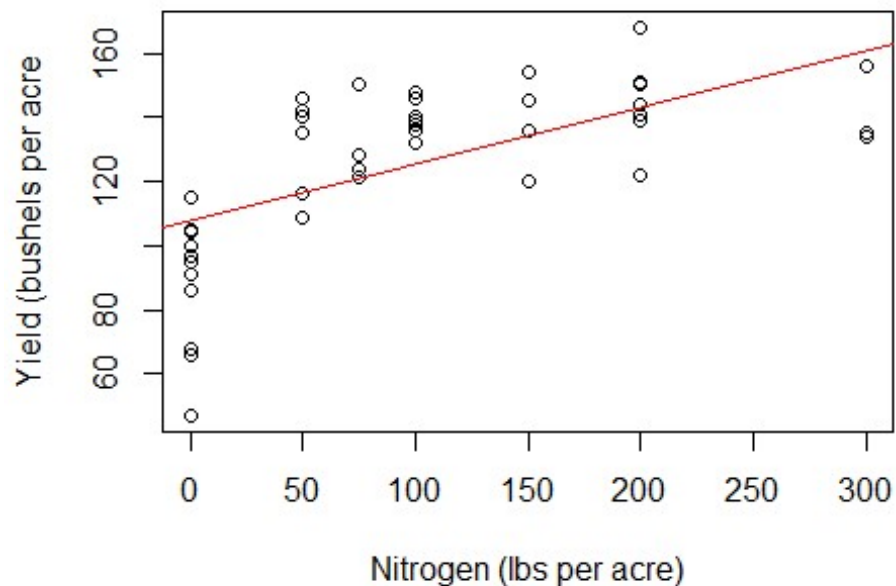
The response variable is **yield**

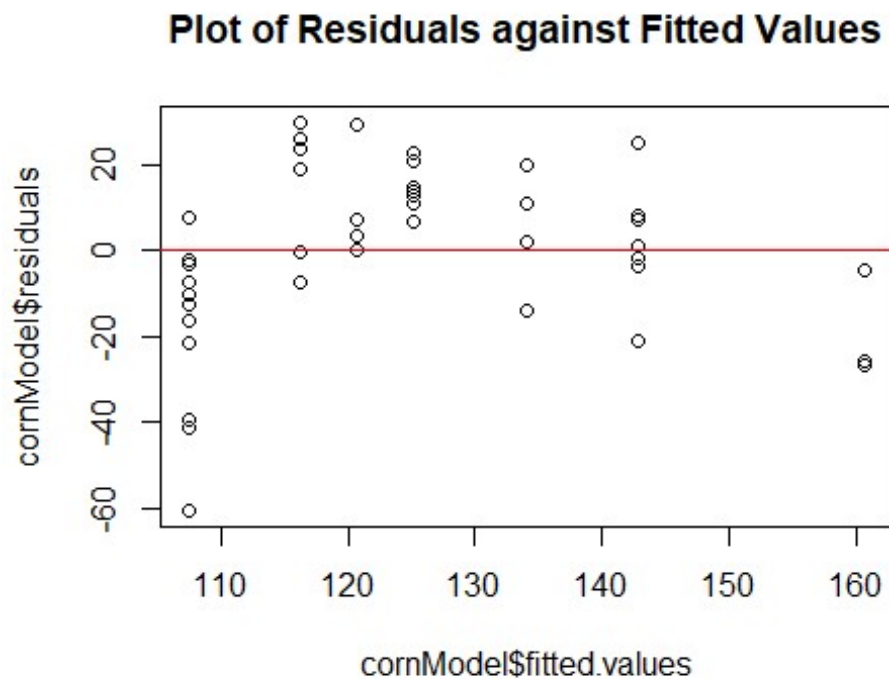
Create a **scatterplot** of the data, and interpret the scatterplot.

```
# 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")
```

Plot Corn Yield Against Nitrogen





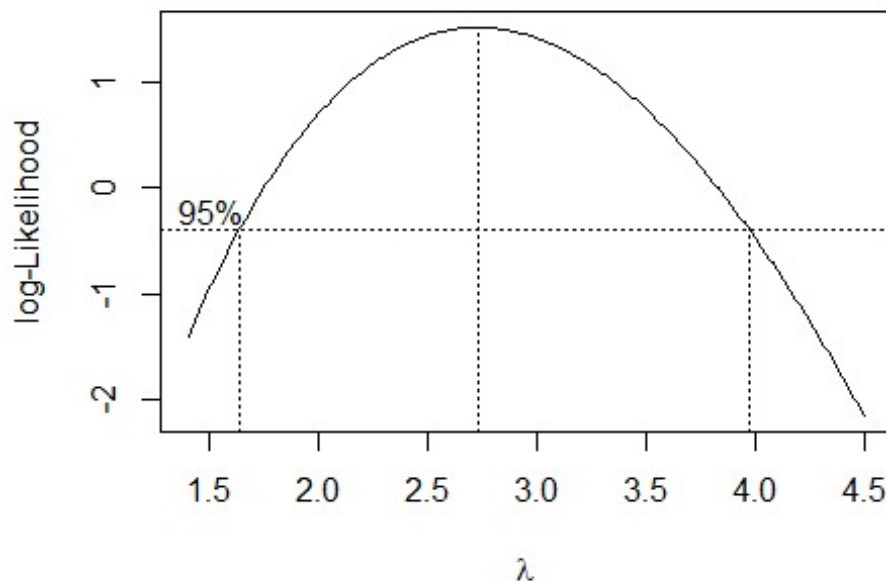
Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason. Answer

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 is 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)
```

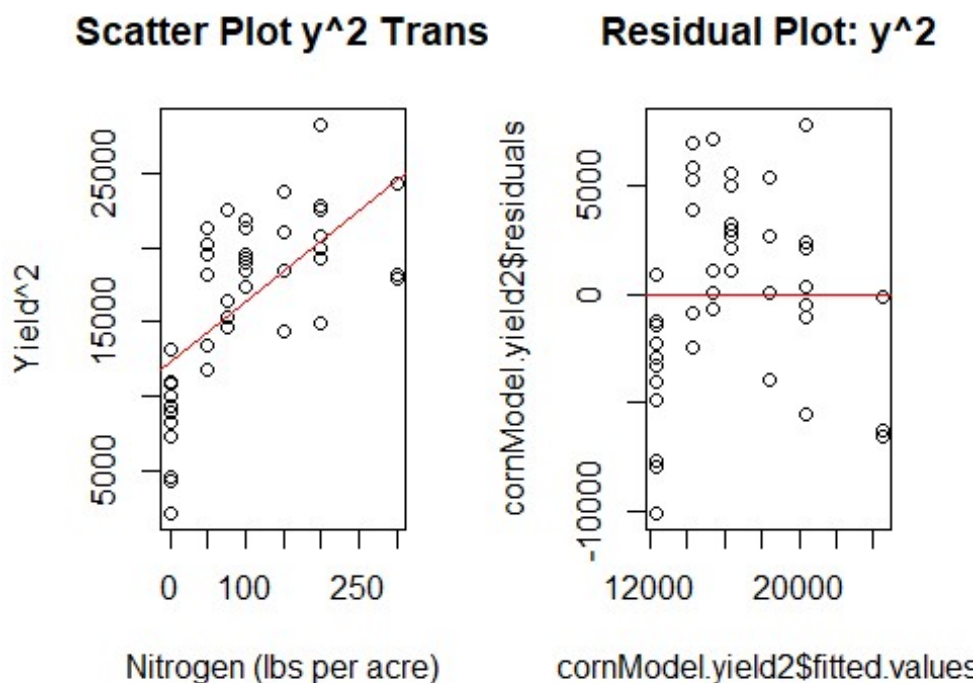
```
par(mfrow=c(1,2)) # 1 rows, 2 columns - there will be three panes to look through
```

```
# 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")
```

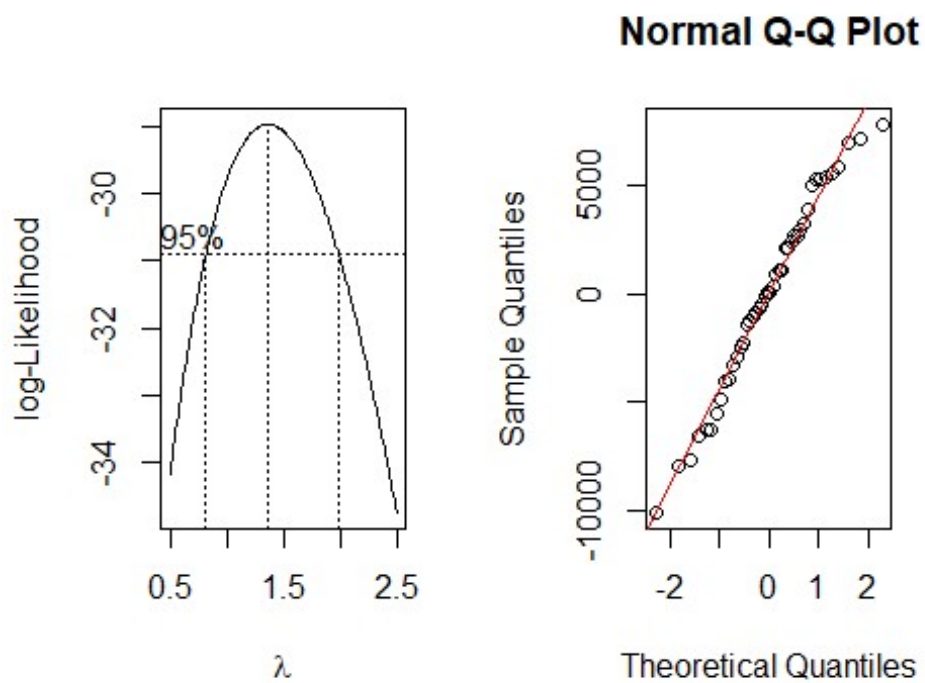


```
# boxcox of y^2 model
```

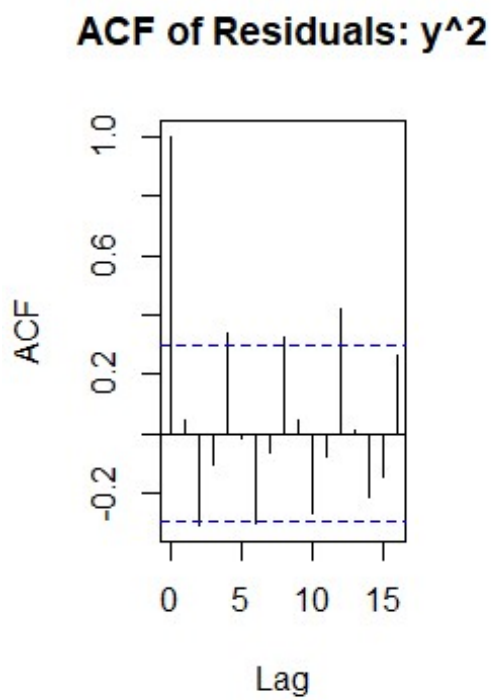
```
boxcox(cornModel.yield2, lambda = seq(0.5, 2.5, 0.01))
```

```
#QQ plot
```

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



```
#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)

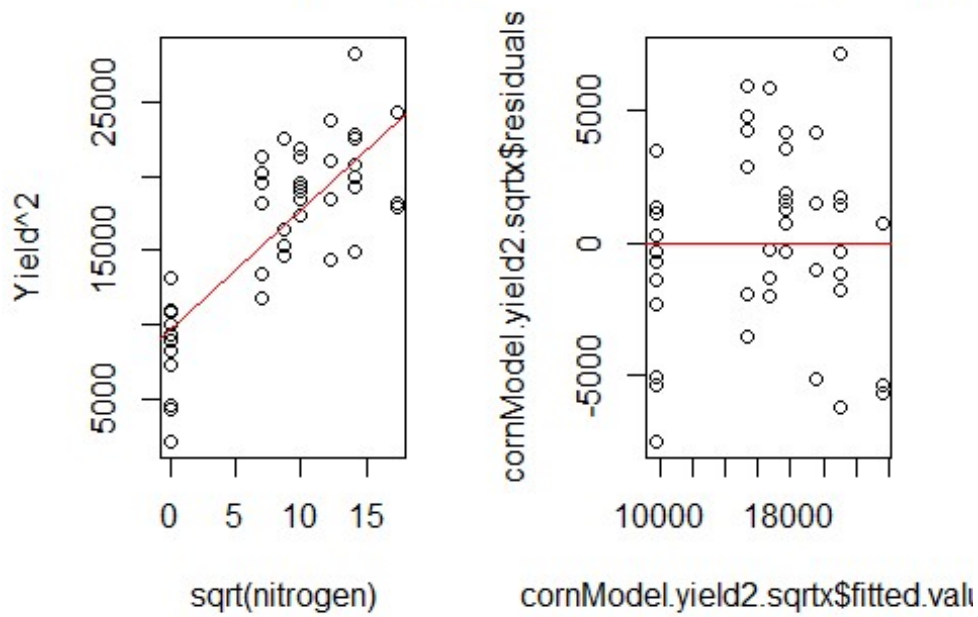
cornModel.yield2.sqrtn <- lm(yield2~sqrt.nitrogen)

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.sqrtn,col="red")

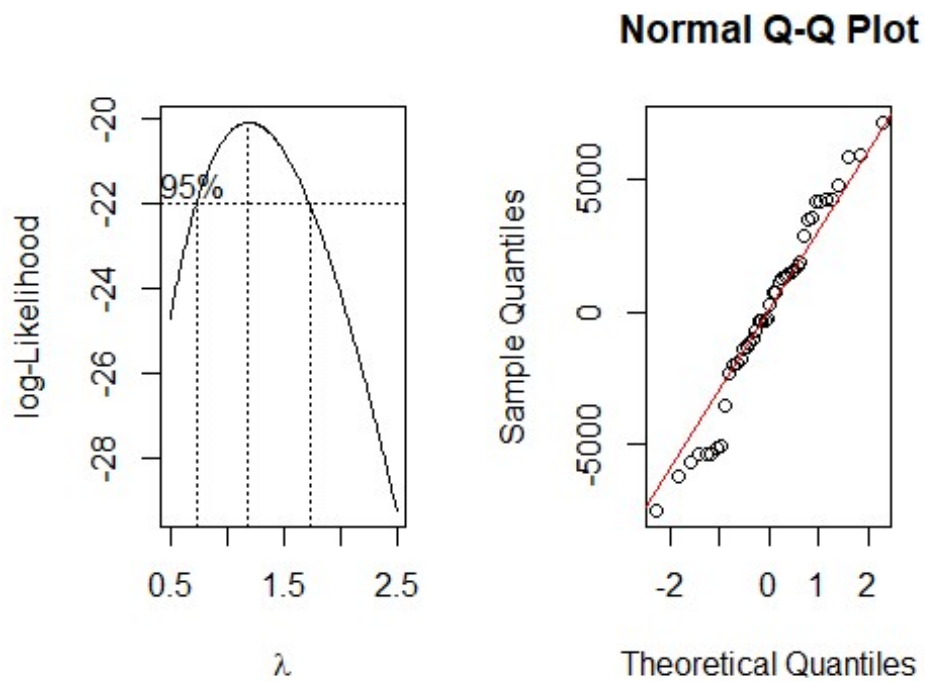
#residual plot yield2, sqrt(x)
plot(cornModel.yield2.sqrtn$fitted.values,cornModel.yield2.sqrtn$residuals,
main="Residual Plot: y^2 sqrt(x)")
abline(h=0,col="red")
```

Scatter Plot $y^2 \sqrt{x}$ **Residual Plot: $y^2 \sqrt{x}$**



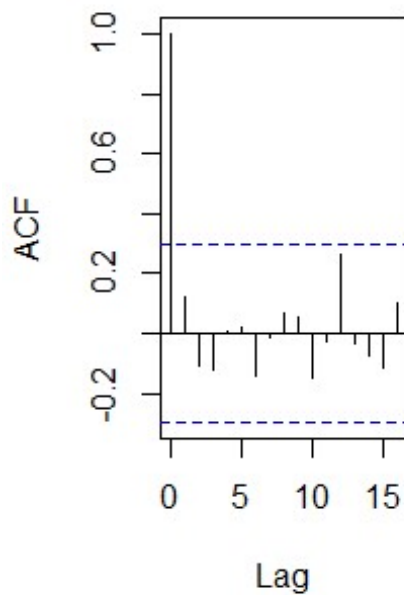
```
# boxcox of  $y^2$  model,  $\sqrt{x}$ 
boxcox(cornModel.yield2.sqrtr, lambda = seq(0.5, 2.5, 0.01))

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



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

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, ]
```

Next, apply the y² transformation and the sqrt(x) transformation, create the model, and plot the data.

```
detach(cornnit)
attach(cornnitFiltered)

yield2.filtered = yield^2 # square the filtered y
sqrt.nitrogen.filtered = sqrt(nitrogen)

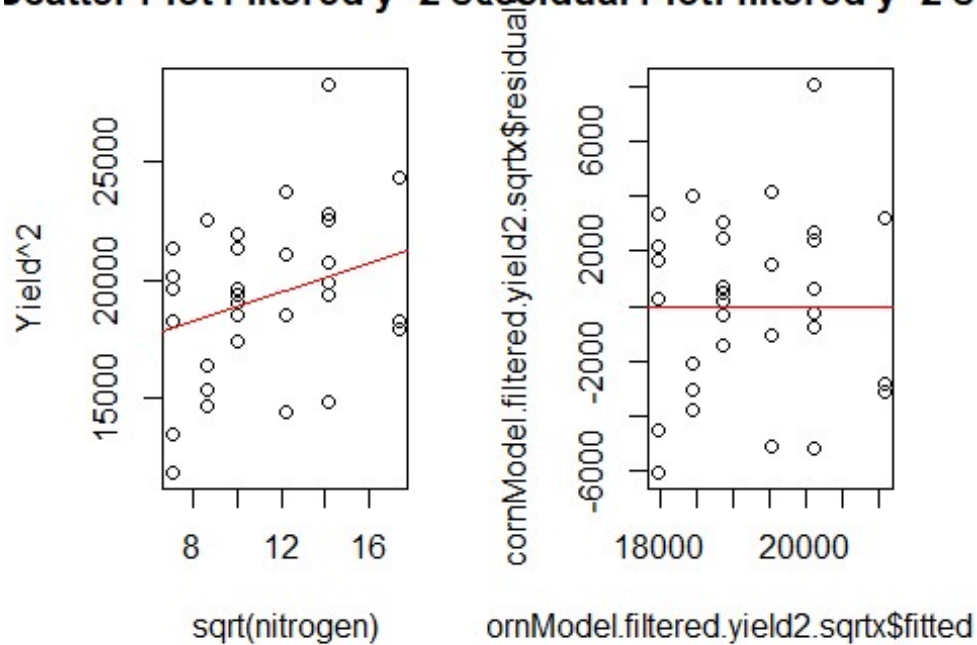
cornModel.filtered.yield2.sqrtn <- lm(yield2.filtered~sqrt.nitrogen.filtered)

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

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

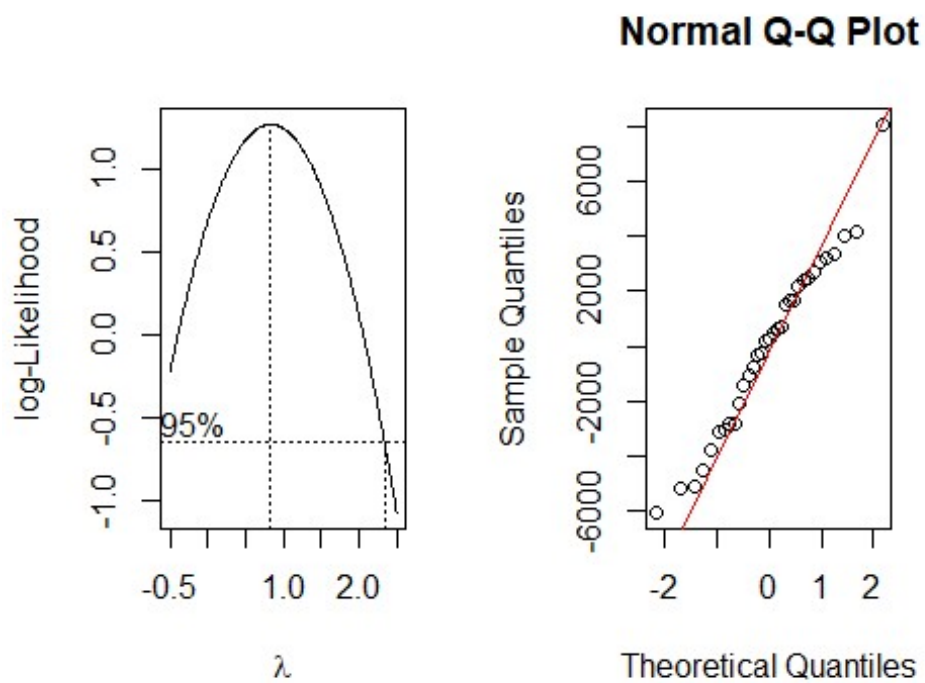
#residual plot filtered yield2, invx
plot(cornModel.filtered.yield2.sqrtn$fitted.values,cornModel.filtered.yield2.sqrtn$residuals, main="Residual Plot: filtered y^2 sqrt(x)")
abline(h=0,col="red")
```

Scatter Plot Filtered y^2 s residual Plot: filtered y^2 s



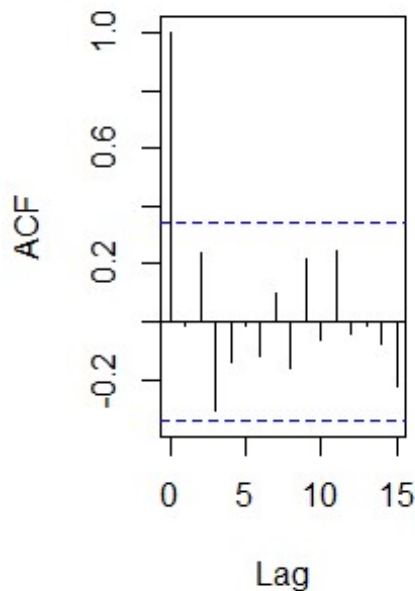
```
# boxcox of filtered  $y^2$  sqrt(x) model
boxcox(cornModel.filtered.yield2.sqrtr, lambda = seq(-0.5, 2.5, 0.01))

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



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

ACF of Residuals: Filtered y^2



The model still has 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 $y^2 \sqrt{x}$ model:

```
summary(cornModel.filtered.yield2.sqrtn)

##
## Call:
## lm(formula = yield2.filtered ~ sqrt.nitrogen.filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6088.6 -2860.9   255.4  2380.5  8104.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15819.7     2000.0   7.910 6.27e-09 ***
## sqrt.nitrogen.filtered    304.0       170.3   1.785   0.084 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3276 on 31 degrees of freedom
## Multiple R-squared:  0.09321,    Adjusted R-squared:  0.06396
## F-statistic: 3.187 on 1 and 31 DF,  p-value: 0.08403

anova(cornModel.filtered.yield2.sqrtn)
```

```
## Analysis of Variance Table
##
## Response: yield2.filtered
##              Df      Sum Sq  Mean Sq F value   Pr(>F)
## sqrt.nitrogen.filtered  1  34198368 34198368   3.1866 0.08403 .
## Residuals              31 332691362 10731979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I also want to review the anova table for the unfiltered $y^2 \sqrt{x}$ model:

```
summary(cornModel.yield2.sqrtn)

##
## 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.sqrtn)

## 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?

Briefly 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, γ , 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 γ , 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 adamant 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 B_1 -hat and B_0 -hat in context?

Answer After transformation, the regression coefficients are: B_0 -hat: 1.50792

B_1 -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 B_1 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.