uwa6xv_M05_HW

Alanna Hazlett

2024-02-29

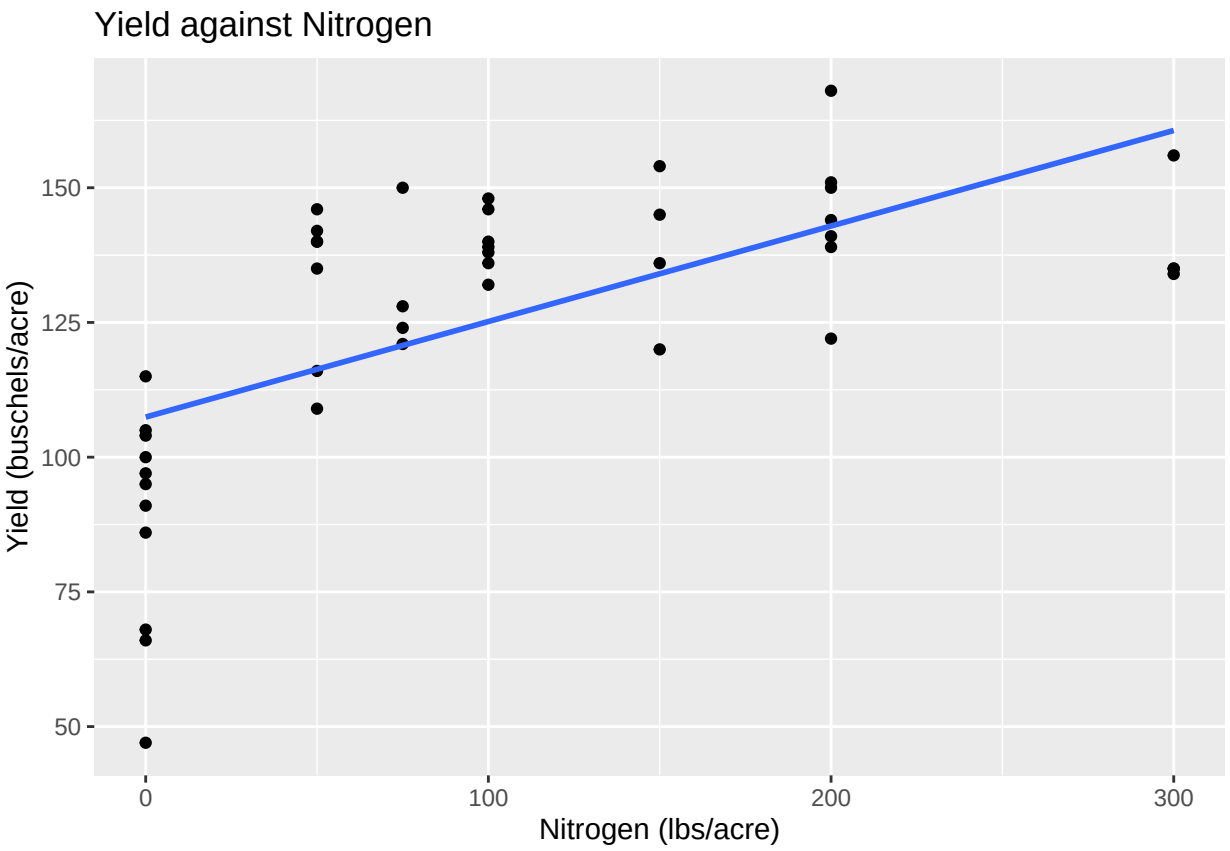
```
Data<-cornnit
```

Problem 1

(a)
What is the response variable and predictor for this study? Create a scatterplot of the data, and interpret the scatterplot.
The response variable is yield (bushels/acre) and the predictor variable is nitrogen (pounds/acre).

```
ggplot(Data, aes(x=nitrogen, y=yield))+
  geom_point()+
  geom_smooth(method='lm',se=FALSE)+
  labs(x="Nitrogen (lbs/acre)",y="Yield (buschels/acre)", title="Yield against Nitrogen")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



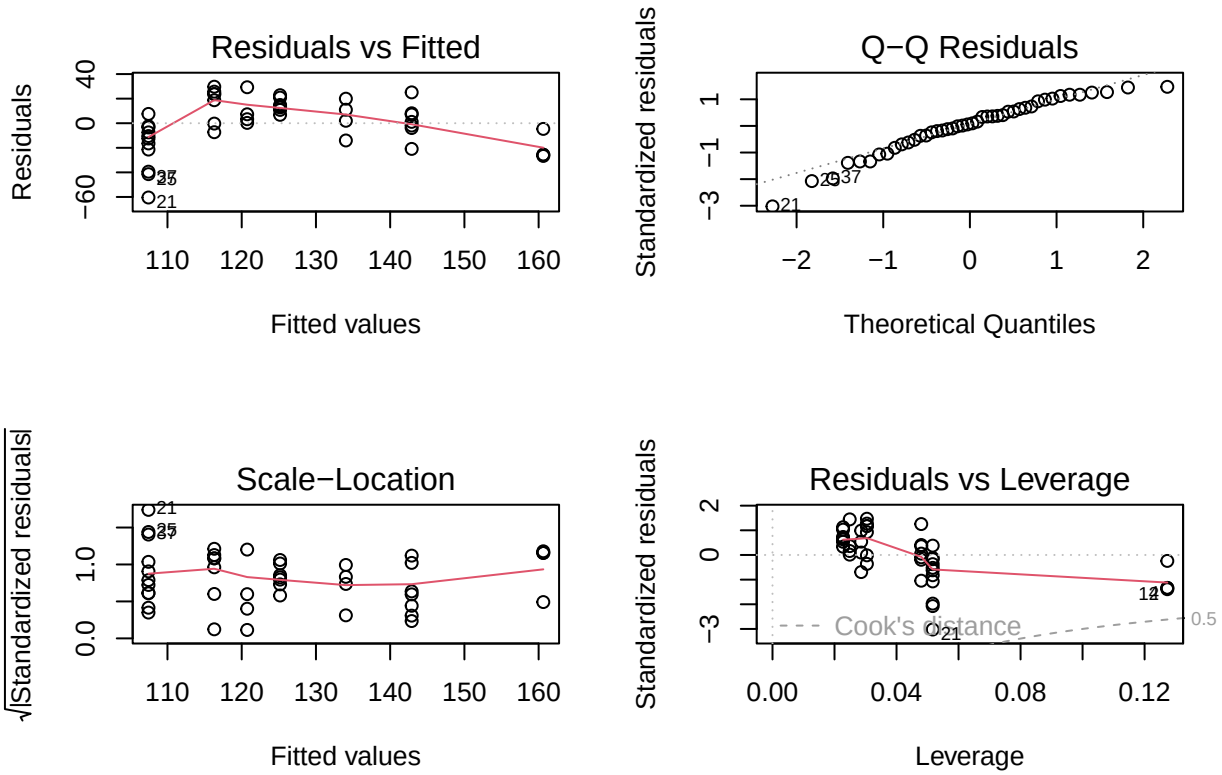
The data points do not seem to be fitting the linear regression line well. They seem to be starting at (0,0) and

increasing instanteously curving a bit and tapering off at the upper right, as if it is reaching an upper limit. The data points are not evenly spread on both sides of the line. At 0 on the x axis the majority of the data points are below the regression line. From about 50 to 150 on the x axis the data points are mainly above the regression line. The variance is large on the left, but then becomes rather consistent from 50 to 300, maybe slightly smaller at the very end.

(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)
par(mfrow=c(2,2))
plot(result)
```



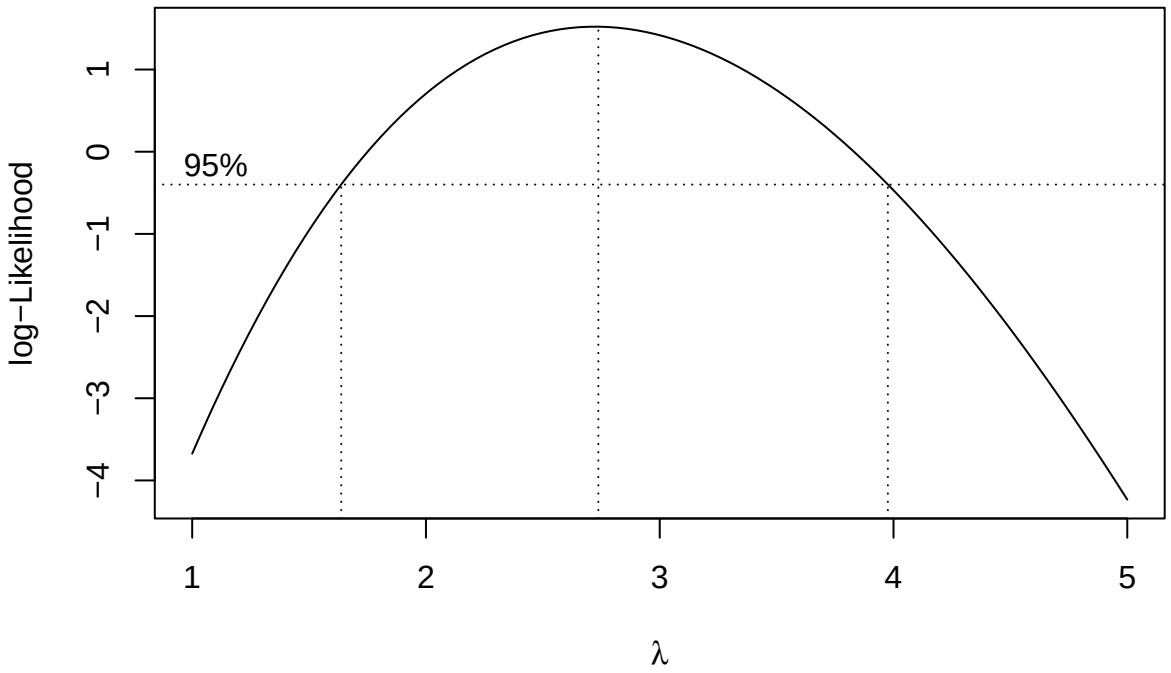
Based on the Residuals vs Fitted plot we can see that the errors mean is not equal to zero, so assumption one is not met. In this plot we can also see that the variance of the errors is not constant, from the left side it is large and going right it gets smaller, so assumption two is not met. Since the variance is getting smaller, we know that we will need a lambda with a value of greater than one.

The Q-Q Residuals plot gives us information about if the distribution of the variable is normal, the fourth assumption. This appears to be met.

Based on the fact that assumption one and two are not met we will be transforming our y variable, because this has an affect on both assumption one and two.

(c) Create a Box Cox plot for the profile log-likelihoods. How does this plot aid in your data transformation?

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

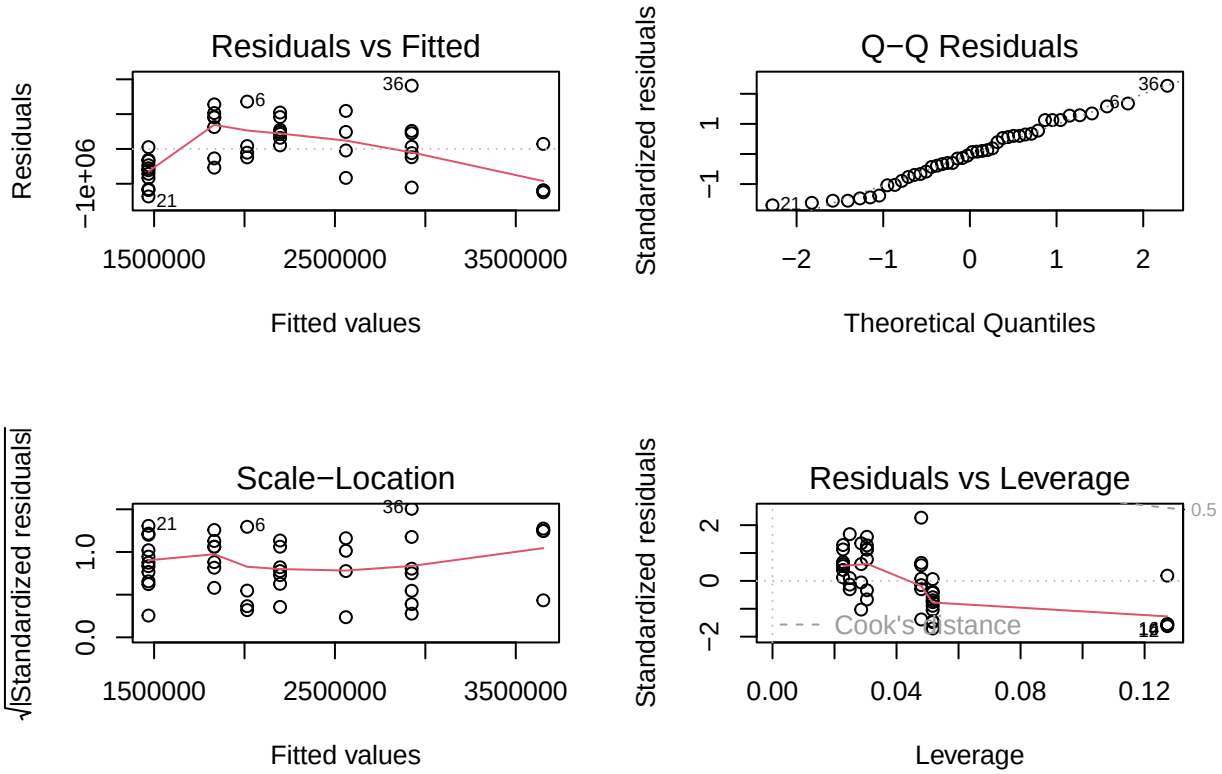


Our 95% confidence interval for lambda lies roughly between 1.75 and just under 4. This allows us to determine what value of lambda to use to transform our y variable, as it gives a lower limit and upper limit for our confidence interval. It also states what the optimal value of lambda should be with the centered vertical line.

(d)

Perform the necessary transformation to the data. Re-fit the regression with the transformed variable(s) and assess the regression assumptions.

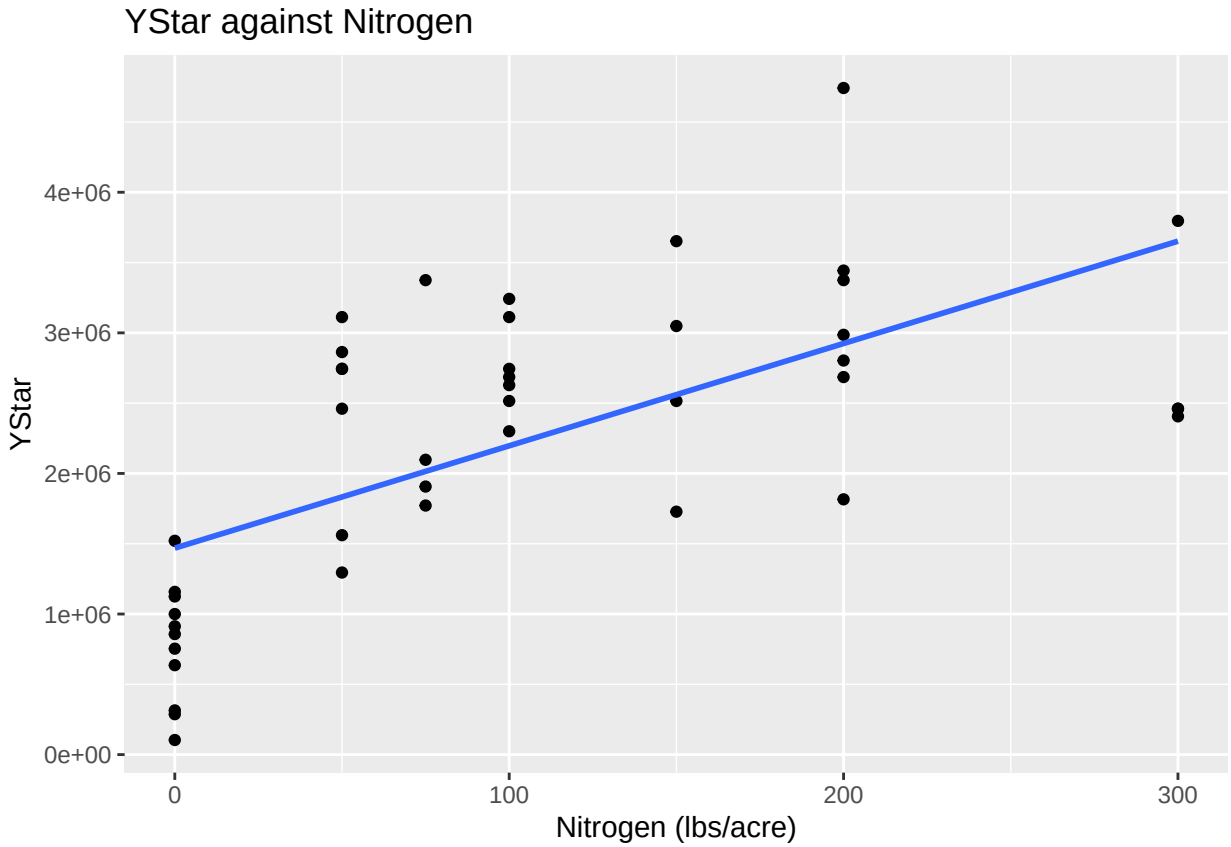
```
ystar<-(Data$yield) ** 3
Data<-data.frame(Data,ystar)
ystar_result<-lm(ystar~nitrogen,data=Data)
par(mfrow=c(2,2))
plot(ystar_result)
```



*My first transformation I completed is y^3, in an effort to stabilize the variance. I chose this value for lambda, because it is the closest whole number to our optimal value of lambda given in the confidence interval. In the Residuals vs Fitted graph you can see that from left to right the variance of our errors is relatively constant, so this transformation was successful in meeting assumption two.
We can see from the Residuals vs Fitted graph that assumption one is still not met as the errors do not have a mean of zero for each value of the predictor. This indicates that we need to perform a transformation on the x variable. In order to determine what type of transformation to perform we must create a new scatter plot of ystar against the predictor variable, nitrogen. We will evaluate the shape of the data points.

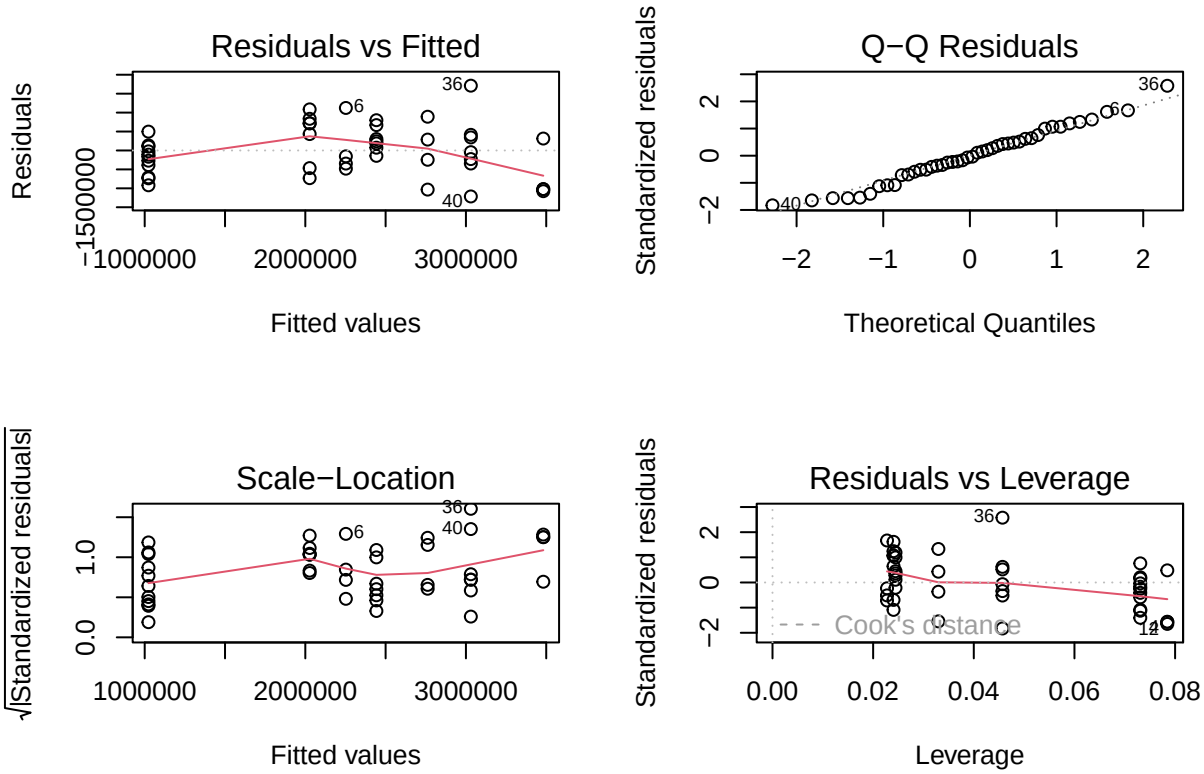
```
ggplot(Data, aes(x=nitrogen,y=ystar))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)+
  labs(x="Nitrogen (lbs/acre)", y="YStar", title="YStar against Nitrogen")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The shape of our data points starts at (0,0) and increases quite a bit then tapers off and approaches an upper limit. *This shape matches the shape of a square root function, so we will perform a square root transformation on the x variable.

```
xstar<-(Data$nitrogen) ** (1/2)
xstar_result<-lm(ystar~xstar)
par(mfrow=c(2,2))
plot(xstar_result)
```



In the residuals vs fitted graph we can see that the vertical variance of the data is rather consistent from left to right, as we would expect, because the transformation of the x variable does not affect this. So assumption 2 is met.
In the residuals vs fitted graph we can see that the error of the means is not perfectly zero, but is substantially improved by this transformation of the x variable compared to the eariler residual vs fitted graphs. I would say assumption 1 is met.
In the Q-Q Residuals graph we can see that the errors are normally distributed, so assumption 4 is met.

Problem 2

(a)
Based on the provided scatterplot, I would transform the y variable first, as the variance is decreasing from left to right and the data points are not evenly dispersed on either side of the regression line. Assumption 1 and 2 are not met. Transforming the response variable has the potential to resolve both of these.
(b)
The Log-Likelihoods confidence interval shows 95% confidence for lambda between about -0.05 and 0.1. Zero is within this range and is an appropriate choice, this would be the log transformation of the y variable. I agree with my classmate's decision.
(c)
The estimated regression equation for this model is:

$$log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$y^* = 1.51 - 0.45x$$

1.50792 is Beta0
-0.44993 is Beta1
Beta1 is negative, so the predicted response variable decreases by (1 - (eˆ-0.44993)) x 100 = 63.77 percent for a one unit increase in the predictor.
Beta0 is the log of the concentration at time equal to 0.