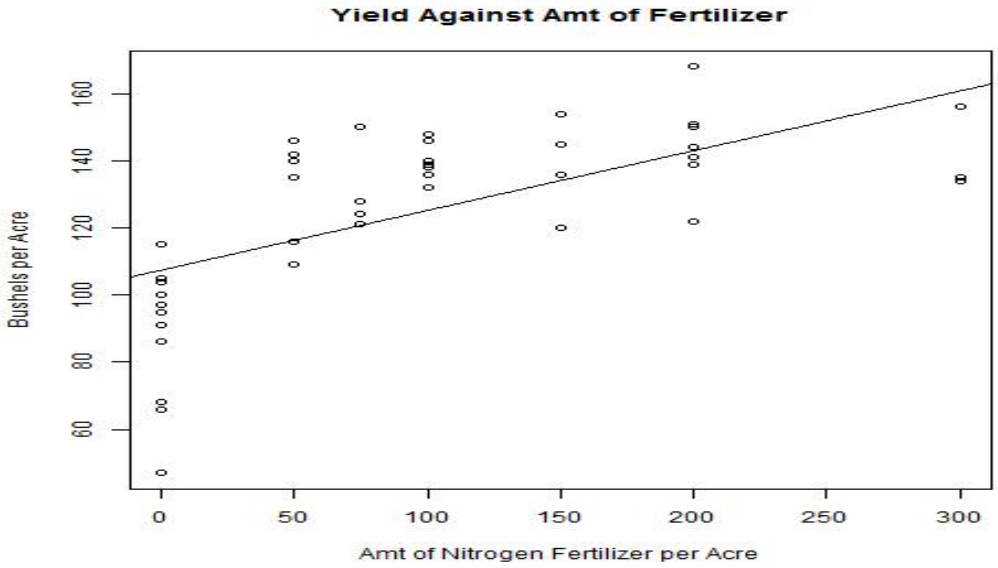
Stat 6021: Homework Set 5 Solutions

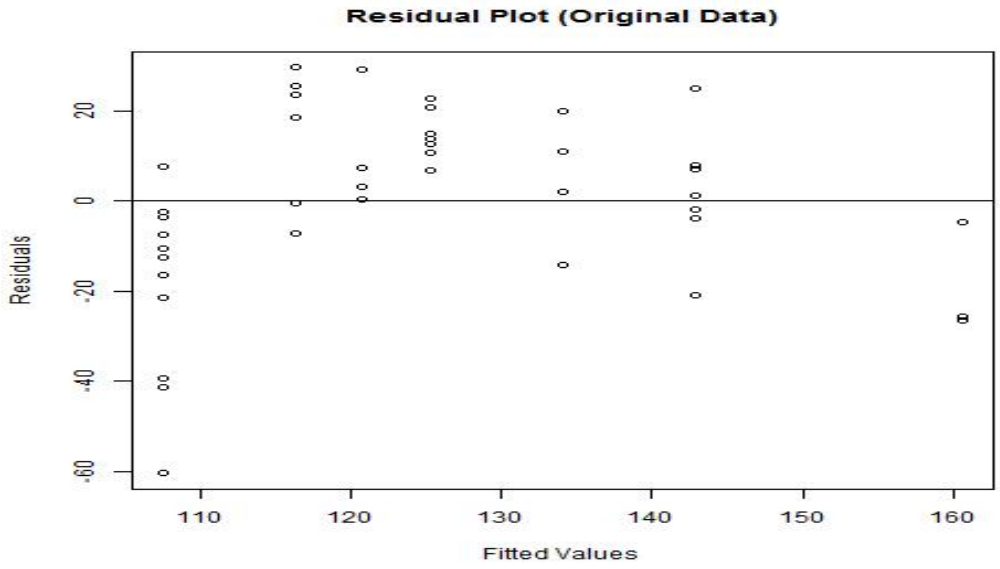
1. (a) The response is the yield of corn. The predictor is the amount of nitrogen fertilizer applied. The scatterplot is displayed below. There appears to be a curved relationship between the amout of fertilizer applied and the yield of corn (not linear).



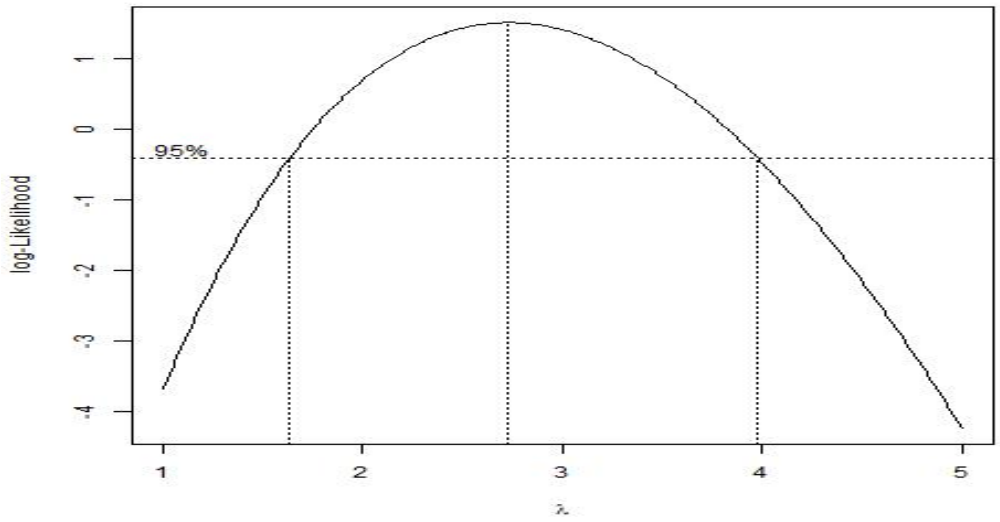
   (b) The residual plot without any data transformations is displayed below. Two things to notice:

   - 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.**

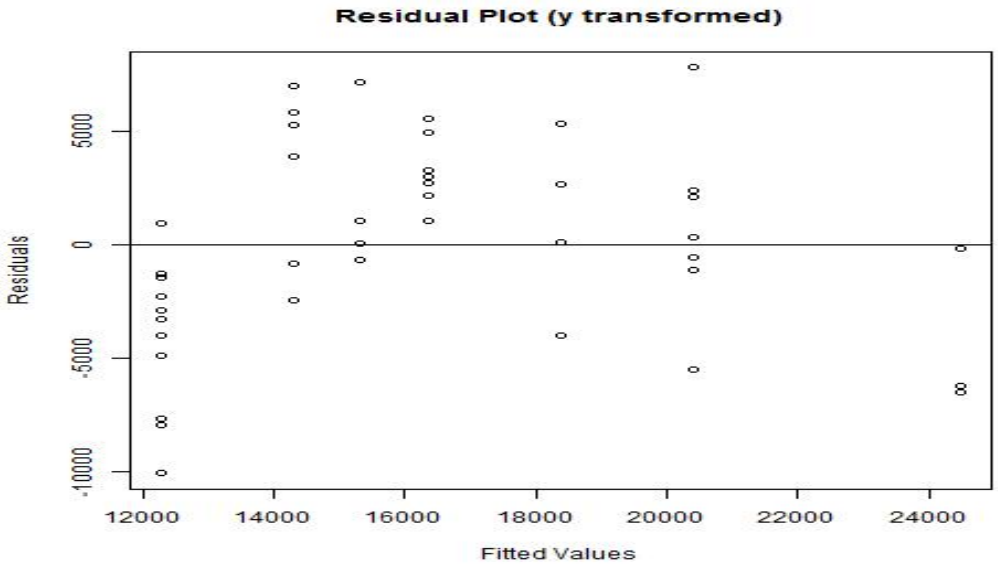


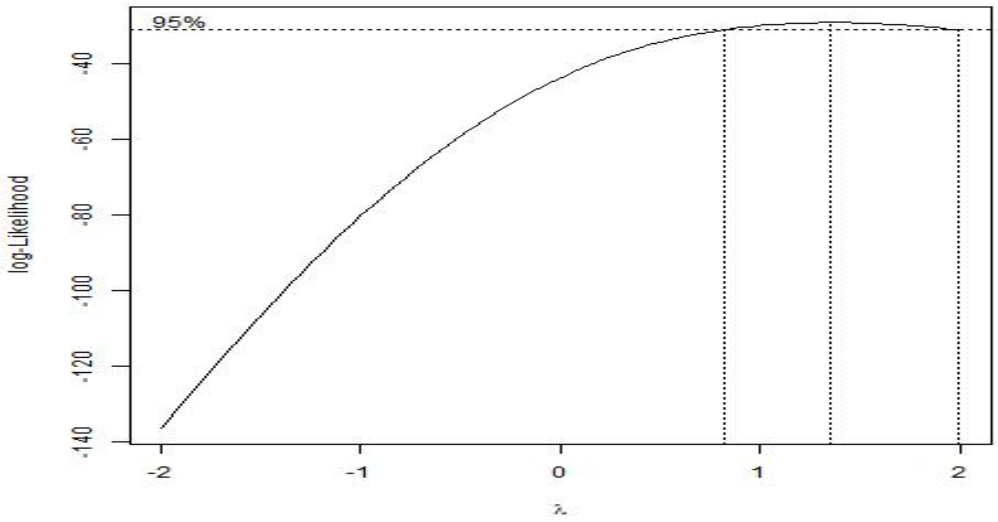
(c) Based on the Box Cox plot below, raising the response variable to power of a value between slightly less than 2 and 4 should work. To keep things simple, I will raise the response variable to the power of 2.



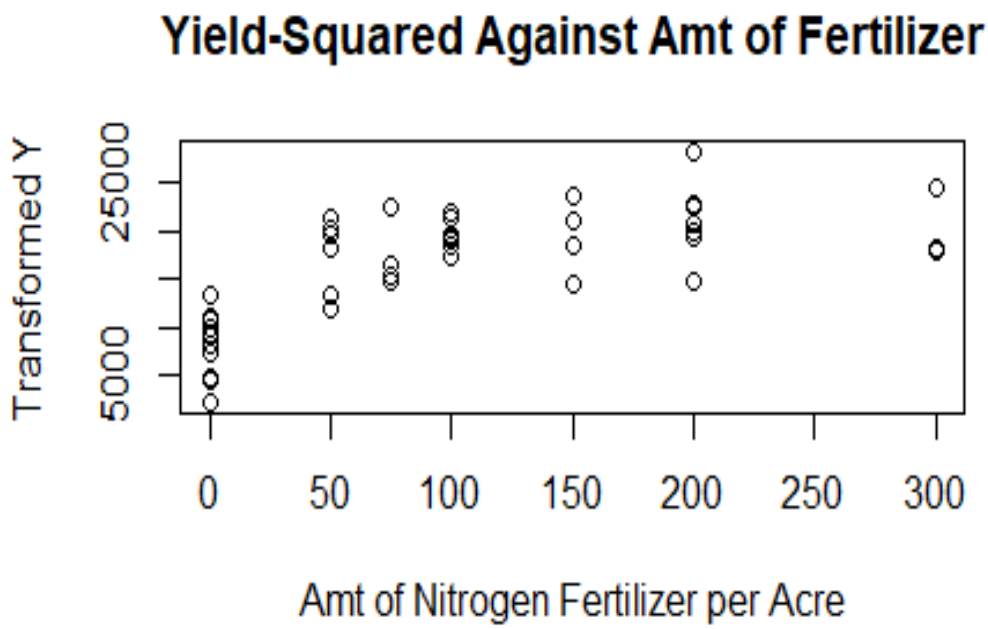
(d) The plots below are the residual plot and Box Cox plot after raising using the transformation $y^* = y^2$. Two things to notice:

- 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. We create a scatterplot of y^* against the predictor to decide how to transform the predictor.

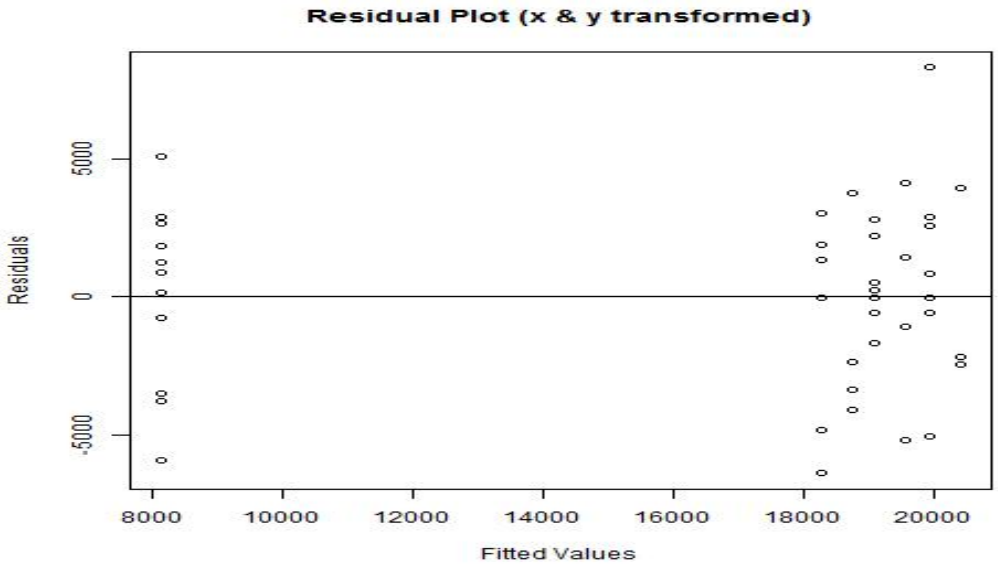


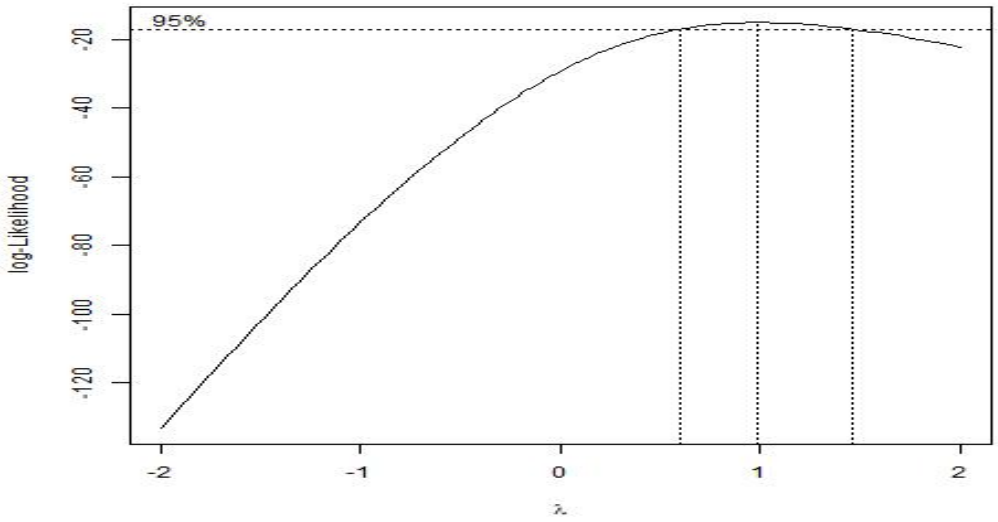
Based on the shape of the scatterplot of y^* against the predictor, we consider a log transformation to the predictor. However, we note that some observations have a value of 0 for the predictor, so we add a small constant, 0.01, first to the predictor and then apply a log transformation to the predictor.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

The plots below are the residual plot and Box Cox plot of profile log likelihoods regressing y^* on x^*. Two things to notice:

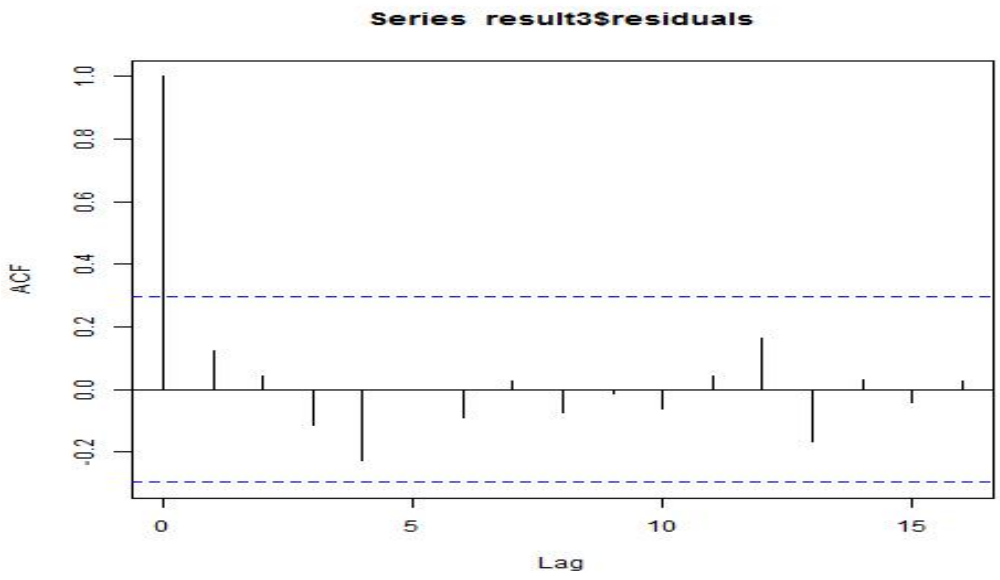
- The variance of the residuals is a lot more constant. The Box Cox plot indicates we no longer need to transform the response variable since 1 lies within the 95% CI.
- The residuals are evenly scattered across the horizontal axis.

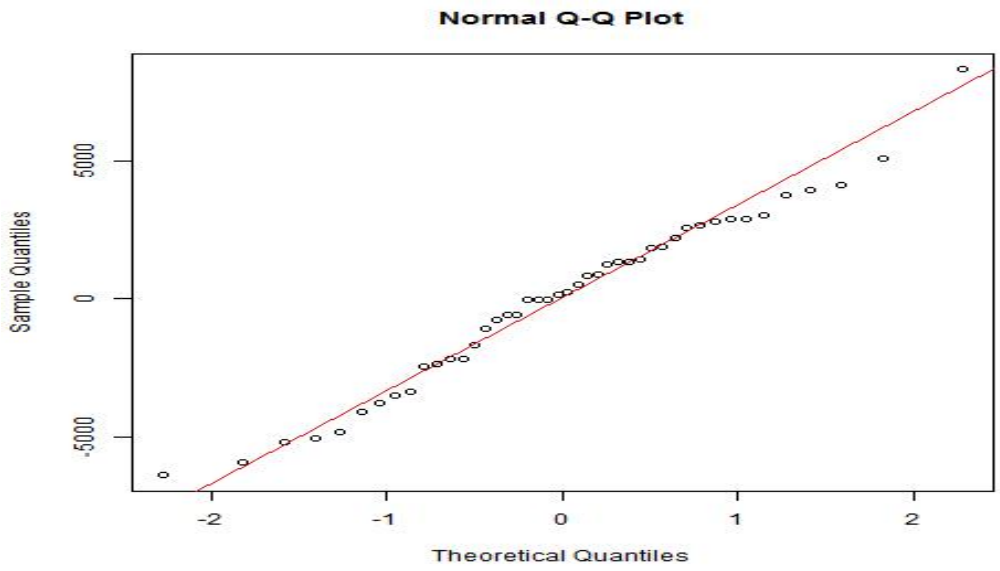
We no longer need to transform the variables.





Last, we check the other assumptions using an ACF plot and QQ plot of the residuals. Both plots appear fine: the ACF plot indicates no correlation between the residuals since all ACFs are insignificant for all lags, and the QQ plot indicates normality of the residuals as the residuals match their values under normality.





The regression equation is now

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

where $y^* = y^2$ and $x^* = \log(x + 0.01)$. Alternatively, this equation can be expressed as

$$y^2 = 13617.4 + 1188.3 \log(x + 0.01)$$

2. (a) The scatterplot of y against x suggests the relationship is not linear. Looking at the residual plot, the constant variance of error term assumption is not met, as the vertical variation of the residuals increases. I would transform the response variable first. Doing so will enable us to stabilize the variance. Once that is

achieved, we may also achieve linearity. If linearity is still not achieved, then I would transform the predictor. Transforming the predictor does not affect the variance of the error terms.

(b) I agree. Based on the output, we should choose $\lambda = 0$, and apply a log transformation on the response variable.

(c) $\hat{y^*} = 1.5079 - 0.4499x$, where $y^* = \ln(y)$. The predicted concentration of a solution is multiplied by $\exp(-0.44993) = 0.6376728$ when time increases by one unit. The predicted concentration of a solution is $\exp(1.5079) = 4.5172$ when time is 0.