HW Module 5

Matt Scheffel

2022-09-30

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(faraway)

Data<-cornnit
```

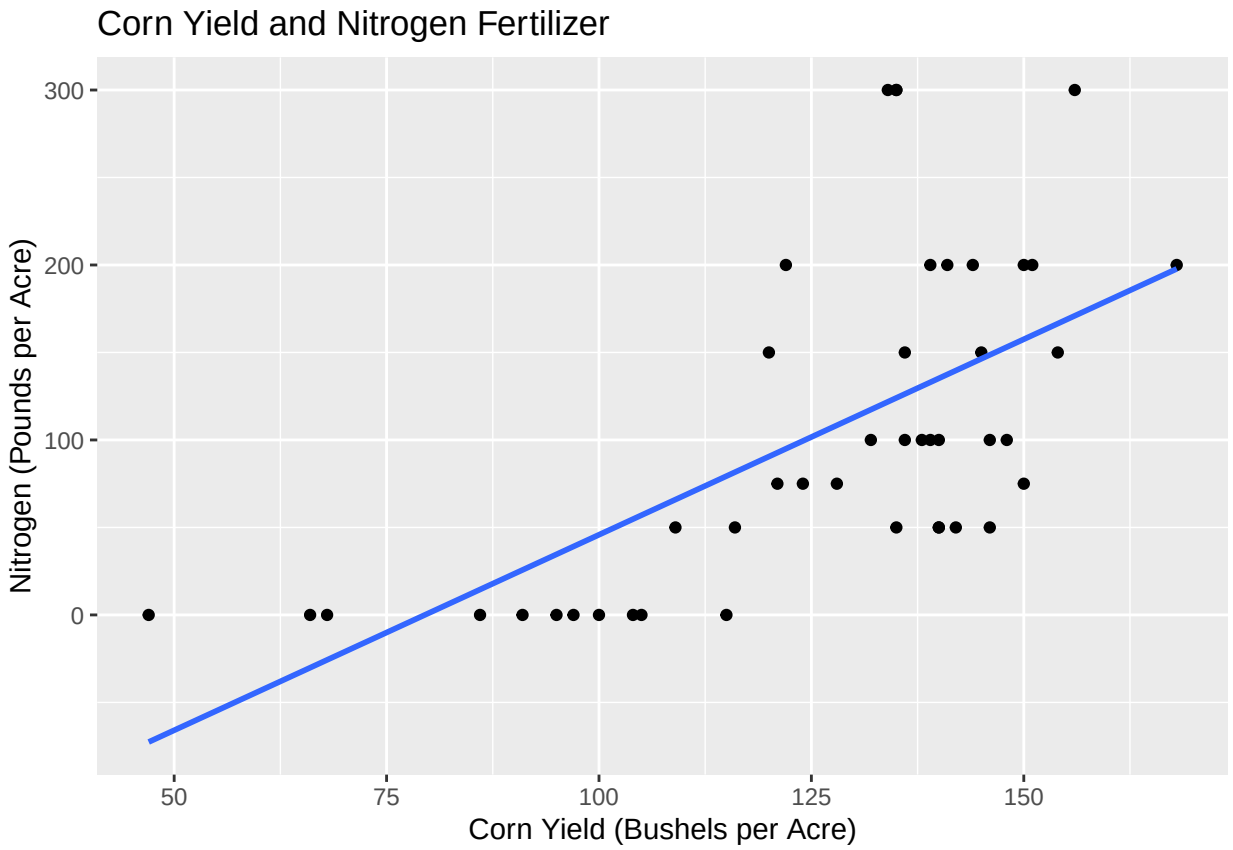
1A

The response variable is the Corn Yield.

The predictor is the Nitrogen Fertilizer.

```
ggplot(Data, aes(x=yield,y=nitrogen))+
geom_point()+
geom_smooth(method = "lm", se=FALSE)+
labs(x="Corn Yield (Bushels per Acre)", y="Nitrogen (Pounds per Acre)",
title="Corn Yield and Nitrogen Fertilizer")
```

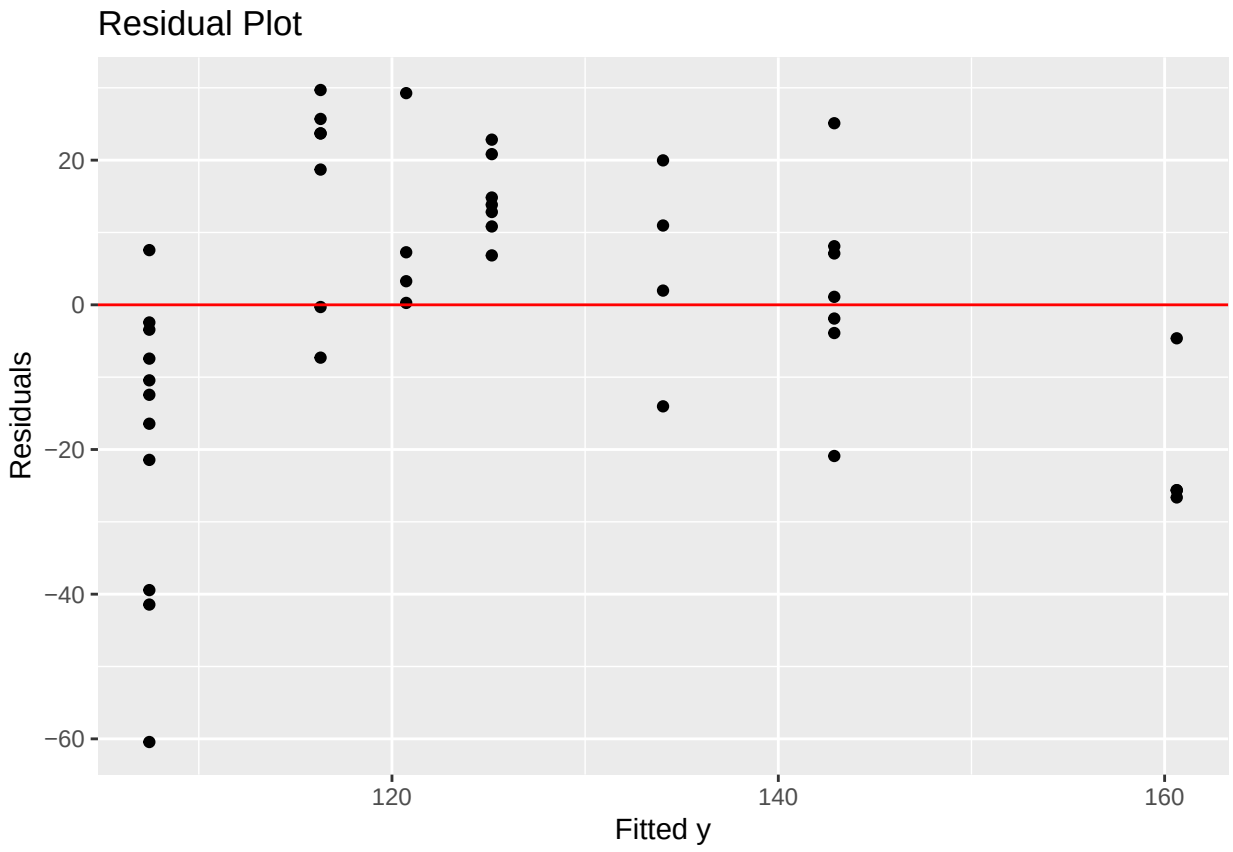
```
## `geom_smooth()` using formula 'y ~ x'
```



The scatterplot seems to indicate that there may be a linear relationship, but there are still some significant outliers. Assumptins 1 and 2 may not be met.

1 B

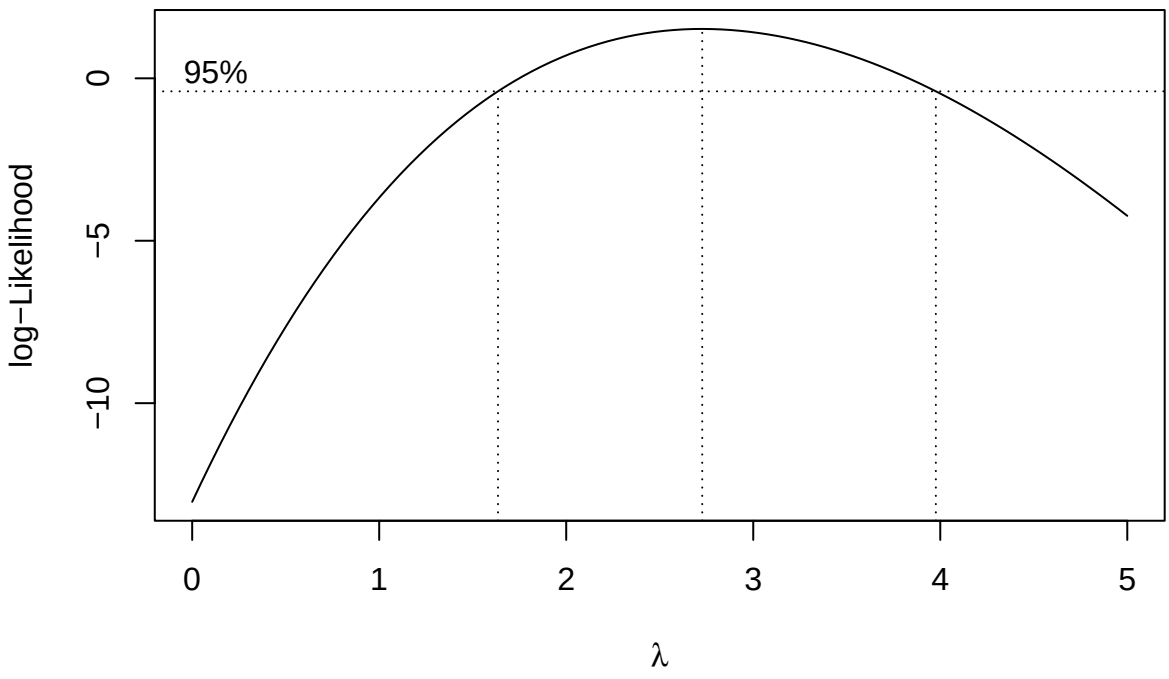
```
result<-lm(yield~nitrogen, data=Data)
Data$yhat<-result$fitted.values
Data$res<-result$residuals
ggplot(Data, aes(x=yhat,y=res))+
geom_point()+
geom_hline(yintercept=0, color="red")+
labs(x="Fitted y", y="Residuals", title="Residual Plot")
```



I will consider transforming the predictor first since I am unsure about the linear relationship of the variables.

1C

```
bc = boxcox(result, lambda=seq(0,5))
```



```
best.lam = bc$x[which(bc$y==max(bc$y))]
```

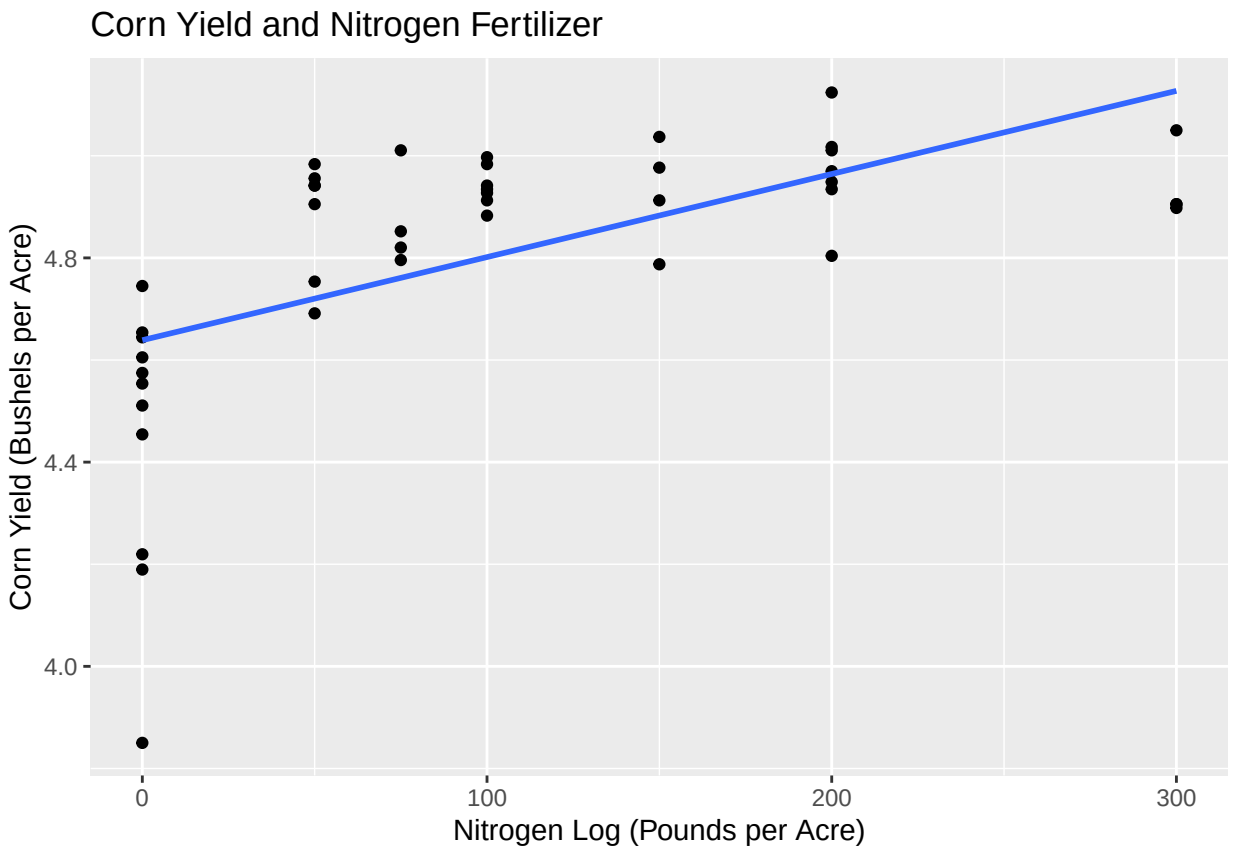
The BoxCox plot helps me determine that I will need to transform the response variable. Variance is likely not constant.

1D

```
##log transform y
Data$log.yield<-log(Data$yield)

##scatter plot of y* against x
ggplot(Data, aes(x=nitrogen,y=log.yield))+
geom_point()+
geom_smooth(method = "lm", se=FALSE)+
labs(x="Nitrogen Log (Pounds per Acre)", y="Corn Yield (Bushels per Acre)",
title="Corn Yield and Nitrogen Fertilizer")
```

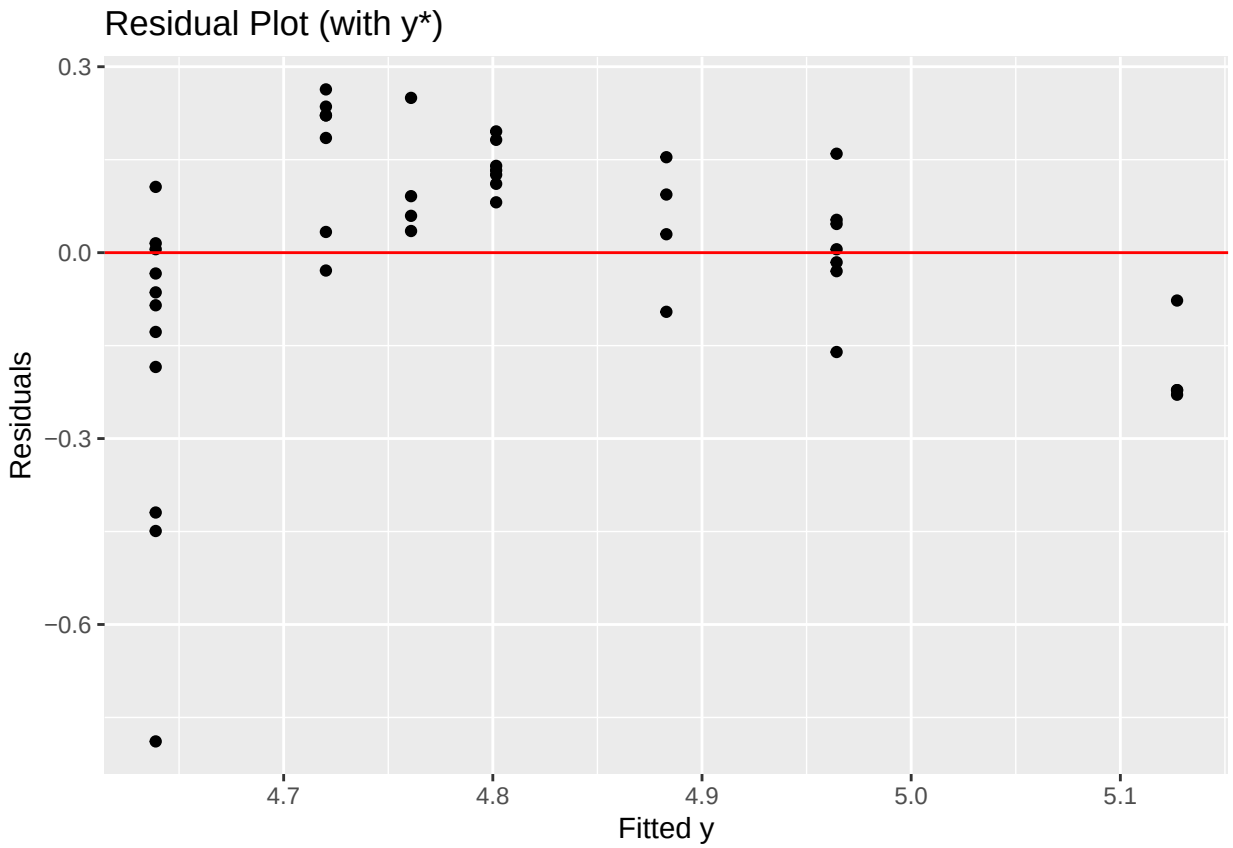
```
## `geom_smooth()` using formula 'y ~ x'
```



Relationship is more linear but there are still many outliers.

```
result2<-lm(Data$log.yield~Data$nitrogen, data=Data)
##create residual plot
Data$yhat2<-result2$fitted.values
Data$res2<-result2$residuals
ggplot(Data, aes(x=yhat2,y=res2))+
geom_point()+
```

```
geom_hline(yintercept=0, color="red")+
labs(x="Fitted y", y="Residuals", title="Residual Plot (with y*)")
```

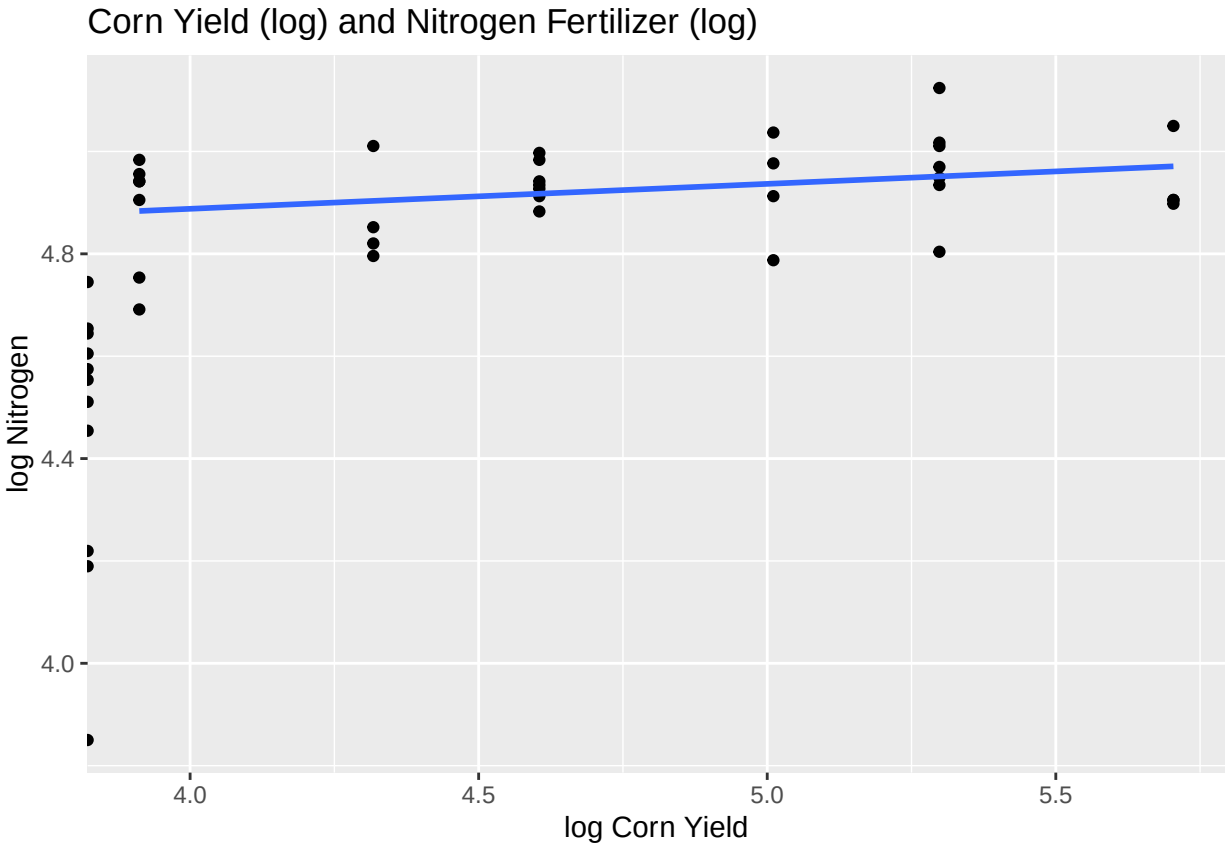


Residual plot is still fairly unchanged, suggesting the need for another transformation of the predictor variable.

```
##transform x
Data$log.nitrogen<-log(Data$nitrogen)
##scatterplot of y* against x*
ggplot(Data, aes(x=log.nitrogen,y=log.yield))+
geom_point()+
geom_smooth(method = "lm", se=FALSE)+
labs(x="log Corn Yield", y="log Nitrogen",
title="Corn Yield (log) and Nitrogen Fertilizer (log)")
```

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

```
## Warning: Removed 11 rows containing non-finite values (stat_smooth).
```



This latest transformation displays a strong linear relationship and indicates we liely need no more transformations.

```
summary(result2)
```

```
##
## Call:
## lm(formula = Data$log.yield ~ Data$nitrogen, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78867 -0.07918  0.03414  0.13468  0.26340
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6388213  0.0471707  98.341  < 2e-16 ***
## Data$nitrogen 0.0016276  0.0003414   4.768 2.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2075 on 42 degrees of freedom
## Multiple R-squared:  0.3512, Adjusted R-squared:  0.3357
## F-statistic: 22.73 on 1 and 42 DF,  p-value: 2.256e-05
```

Regression equation:

$y = 4.6388 + 0.0016x$

```
#trying to remove NA values

#Data[is.na(Data) | Data=="Inf"] = NA

#result3<-lm(Data$log.yield~Data$log.nitrogen, data=Data)
##create residual plot
#yhat3<-result3$fitted.values
#res3<-result3$residuals
#ggplot(Data, aes(x=yhat3,y=res3))+
#geom_point()+
#geom_hline(yintercept=0, color="red")+
#labs(x="Fitted y", y="Residuals", title="Residual Plot (with y* and x*)")
```

10/2/22

Module 05 Homework

Question 2

2a) Based on Figure 1, I would recommend transforming the predictor, x. This is because the linearity is questionable (as seen by the presence of outliers in the scatter plot.) Residuals suggest a non-linear mean.

2b) I do agree with my classmate that the response variable should be transformed first. Based on the BoxCox plot, the distribution looks fairly normal. However, the variance seems likely to not be constant based on the residual plot, so it is a good idea to transform the response variable.

2c) Estimated regression equation:

$$y = 1.50792 - 0.44993x$$ (w/ y = concentration & x = time)

Regression coefficients $\hat{\beta}_1$ & $\hat{\beta}_0$:

In context, the regression coefficients $\hat{\beta}_1$ & $\hat{\beta}_0$ seem to fit the model. The coefficients seem to match the linear relationship shown in the figures and will likely fit better after the log transformation.