# Diagnostic Plots and Transformations of Data

Rafael Romero

2025-02-20

## Contents

**Question 1** **1**
- **1. A.)** Plot the residual against the fitted values. What departures from the linear regression model can be studied from this plot? What are your findings? . . . . . 1
- **1. B.)** Prepare a QQ plot of the residuals. What departures from the linear regression model can be studied from this plot? What are your findings? . . . . . 2
- **1. C.)** Plot the residuals against the ACT scores. What departures from the linear regression model can be studied from this plot? What are your findings? . . . . . 3
- **1. D.)** Discuss possible remedial measures when there are departures in (a), (b) and (c). . . . . 4

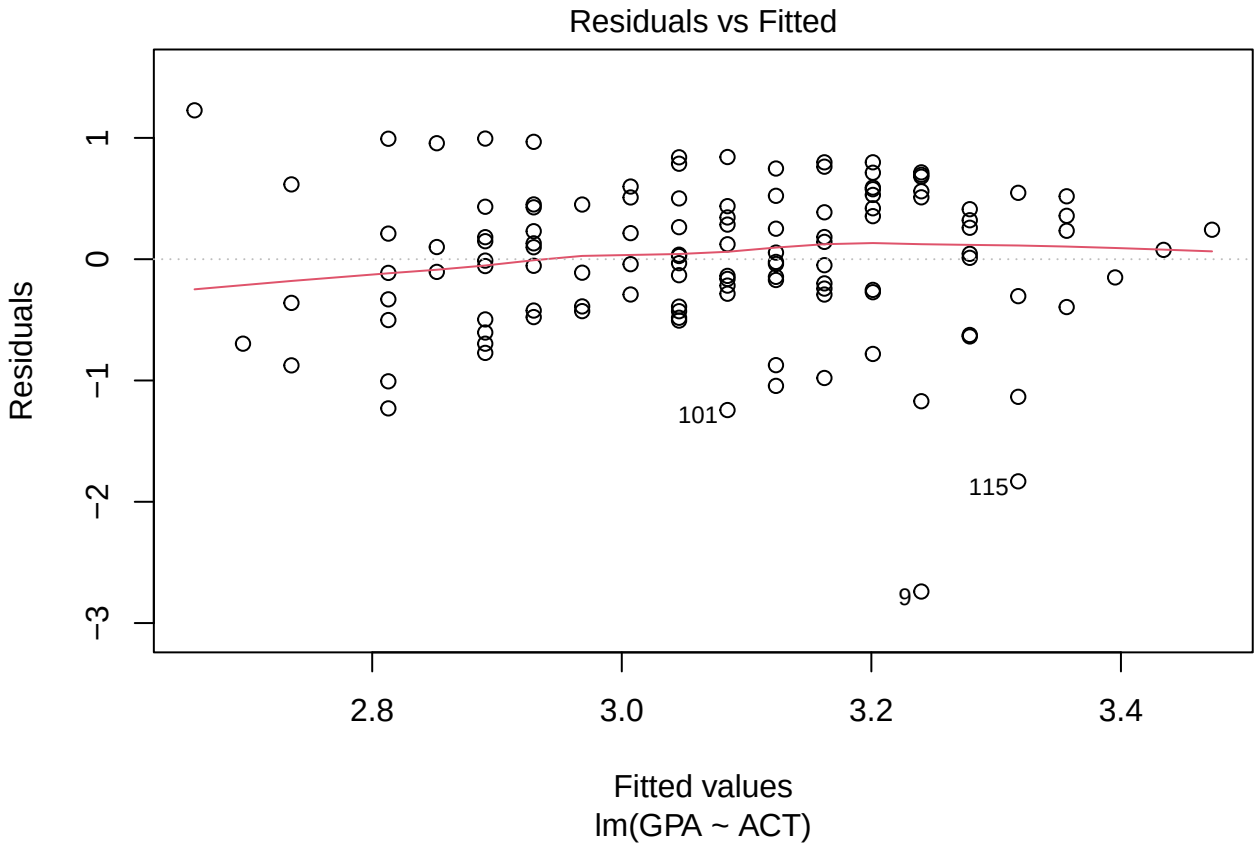
**Question 2 Refer to the Airfreight breakage problem in the first two assignments. Construct diagnostic plots and comment on your findings.** **4**

**Question 3** **6**
- **3. A.)** Plot the data and fit a linear regression model. . . . . 6
- **3. B.)** Construct diagnostic plots and comment on your findings. . . . . 7
- **3. C.)** Use the Box-Cox procedure to find an appropriate power transformation. . . . . 9
- **3. D.)** Use the transformation Y = log10Y and obtain the estimated linear regression function for the transformed data. . . . . 10
- **3. E.)** Plot transfermed data and the estimated regression line in (d). Does the regression . . . . . 11
- **1. F.)** Construct diagnostic plots for the fit in (d) and comments on your findings. . . . . 11
- **3. G.)** Express the estimated regression function in (d) in the original units. . . . . 13

## Question 1

### 1. A.) Plot the residual against the fitted values. What departures from the linear regression model can be studied from this plot? What are your findings?

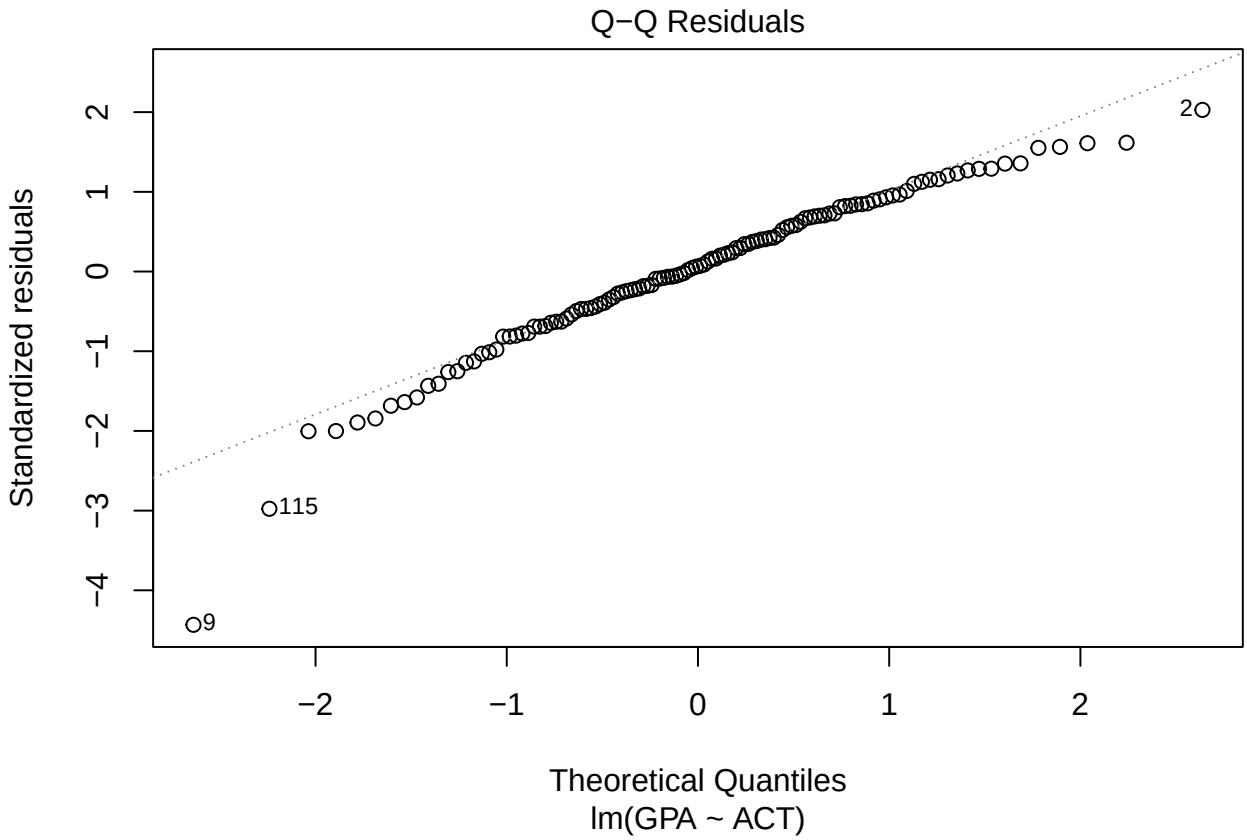
```
data <- read.table("data/GPA.txt", header = TRUE)

GPA <- data$GPA #response variable, dependent
ACT <- data$ACT #predictor variable, independent

model <- lm(GPA~ACT)
plot(model, which=1)
```




**CONCLUSION:** We can analyze the assumptions of the linear regression model and whether or not they are met such as linearity. We notice a relatively straight red line which indicates a linear relationship between ACT and GPA. We can also notice the spread of the variance still relatively stays similar throughout the line. We can also analyze the outliers which are the points furthest from the red line. In conclusion, the randomly scattered residuals and relatively straight line indicates that the linear model is appropriate and a linear relationship between GPA and ACT scores exists.

### 1. B.) Prepare a QQ plot of the residuals. What departures from the linear regression model can be studied from this plot? What are your findings?

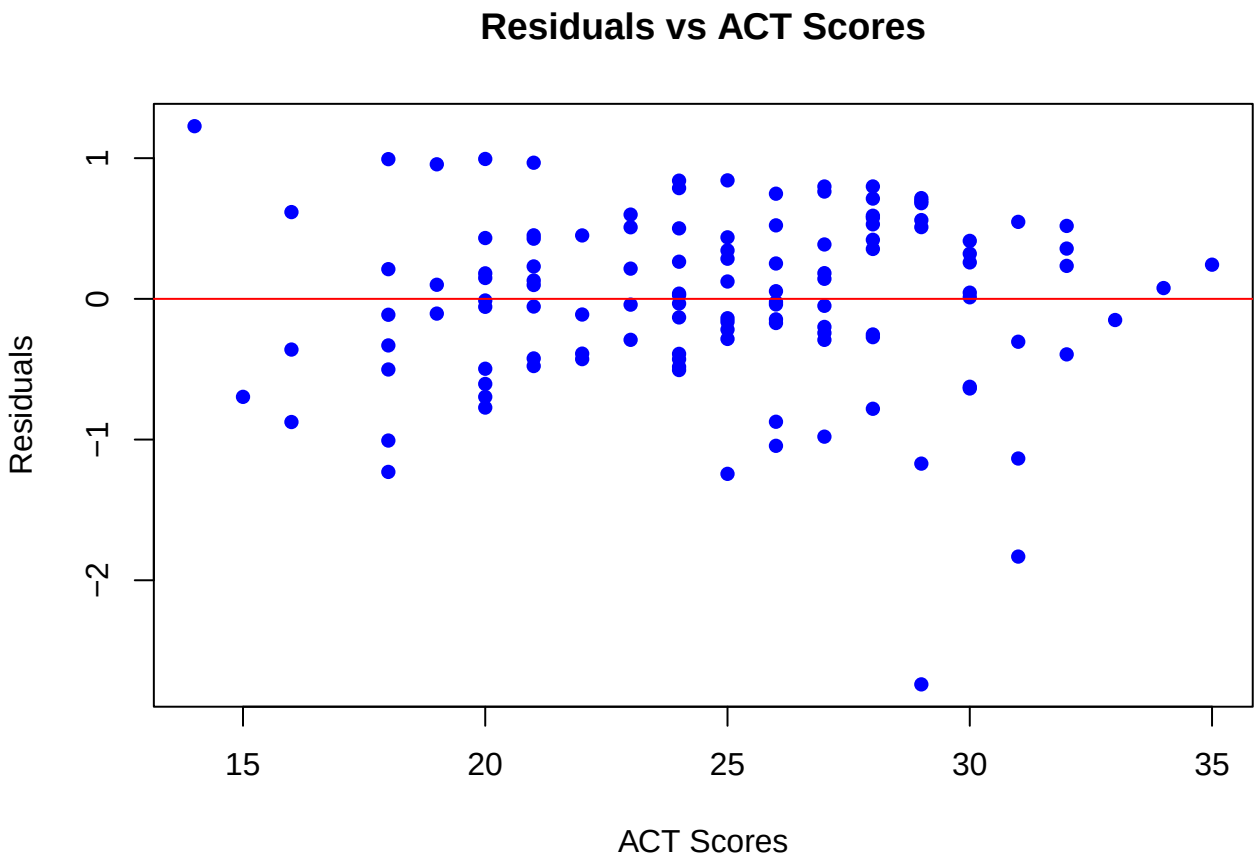
```
plot(model, which =2)
```




**CONCLUSION:** With the Q-Q Residual plot we can analyze the normality of the residuals which is an important assumption for our linear model. The points must closely follow the fitted line in order to indicate a normal distribution. We can see that it is relatively straight in the model but tends to fall under the line at the beginning and at the end. This tells us that there may be outliers or residuals can be non-normally distributed.

### 1. C.) Plot the residuals against the ACT scores. What departures from the linear regression model can be studied from this plot? What are your findings?

```
plot(ACT, model$residuals,
     main = "Residuals vs ACT Scores",
     xlab = "ACT Scores",
     ylab = "Residuals",
     pch = 16, col = "blue")
abline(h = 0, col = "red")
```



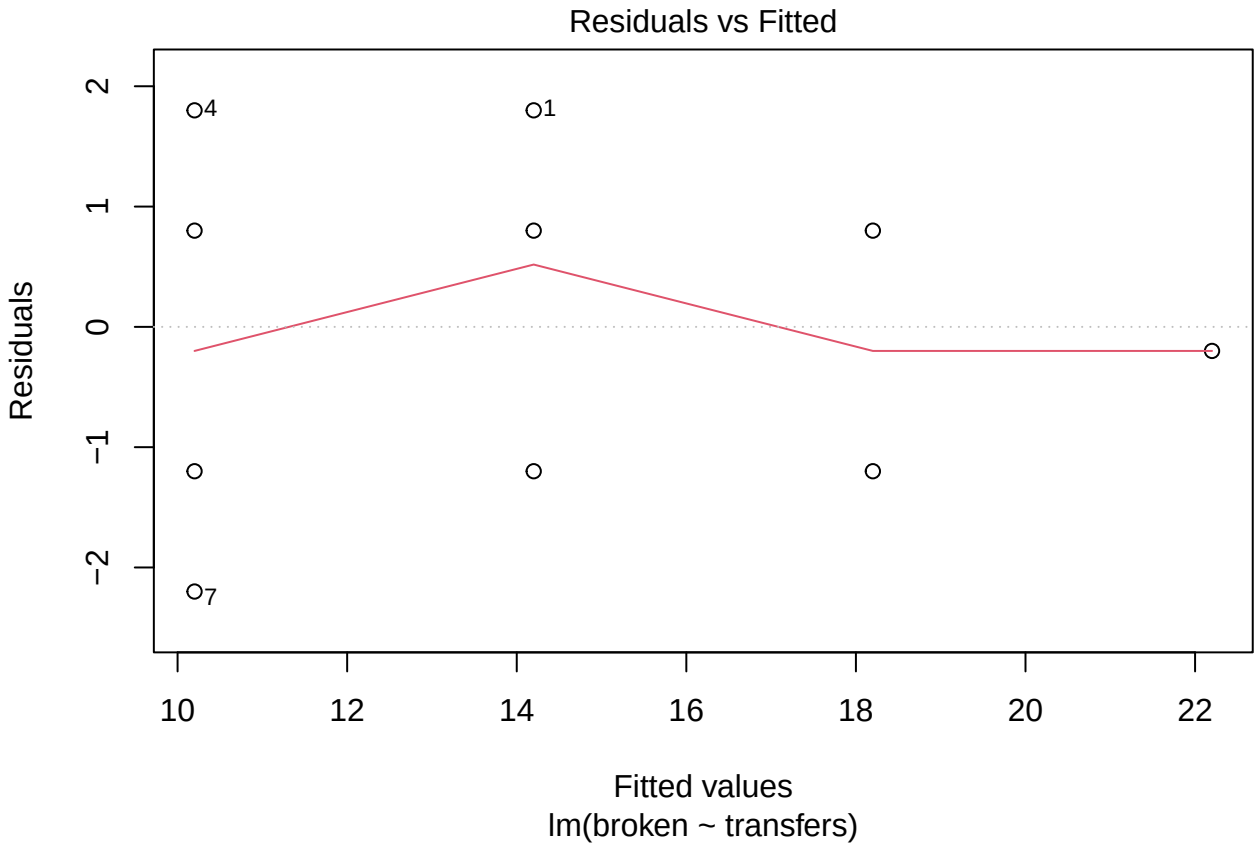

**Conclusion:** We can analyze the linear relationship betweeen GPA and ACT scores. The residuals are fairly well randomly spread out as act scores increase which tells us that the linear transformation is appropriate and a transformation may not be needed. Thus, a linear relationship between act scores and GPA is appropriate.

### 1. D.) Discuss possible remedial measures when there are departures in (a), (b) and (c).

Possible remedial measures include transformation of the data or developing a more appropriate model, but transformations are usually simpler as we can transform X and/or Y and use the transformed variables in order to have an adequate linear regression model. If linearity is not met we can use a polynomial function. For nonconstancy of error variance we can use weighted least squares to ensure normality. For outliers we can do sensitivity analysis and/or robust methods.

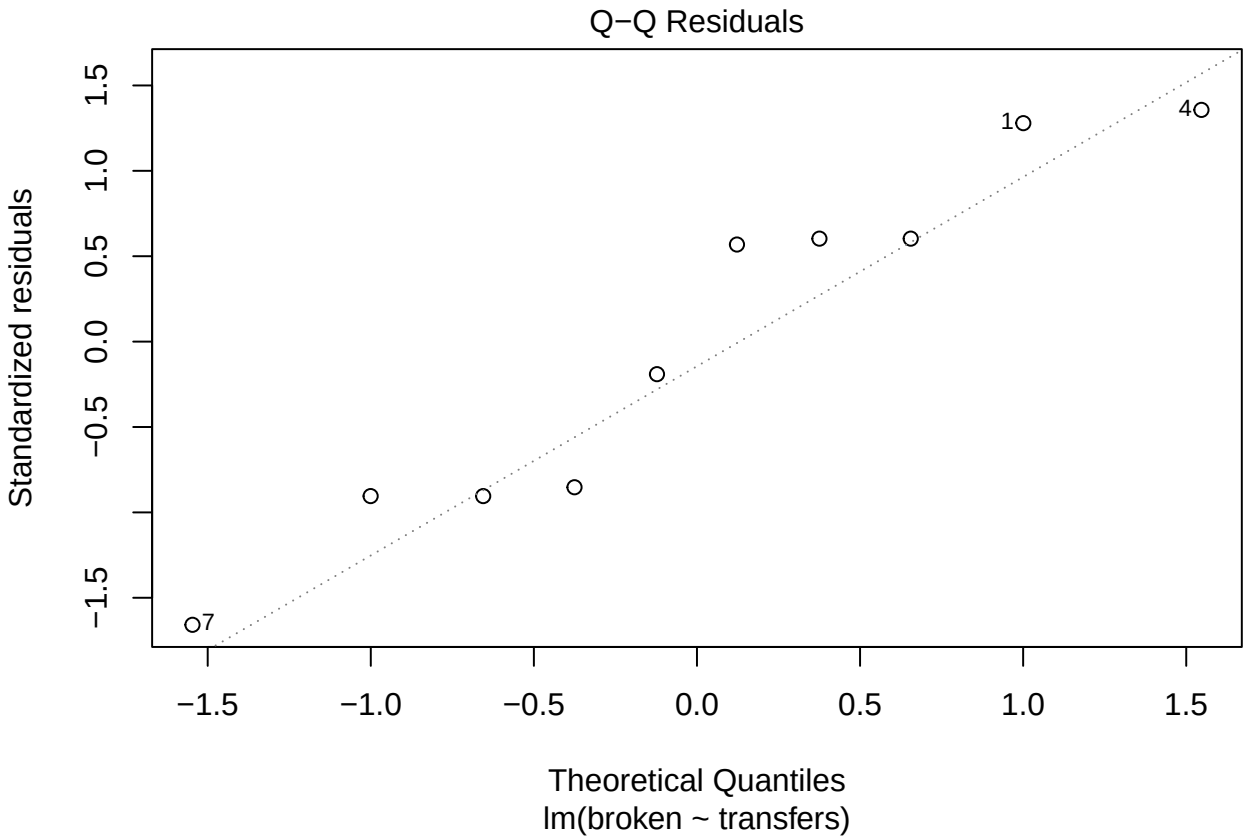
## Question 2 Refer to the Airfreight breakage problem in the first two assignments. Construct diagnostic plots and comment on your findings.

```
airfreight <- read.table("data/airfreight_breakage.txt", header = FALSE)

broken <- airfreight$V1 #response variable, dependent
transfers <- airfreight$V2 #predictor variable, independent
linmodel <- lm(broken~transfers)
plot(linmodel, which = 1)
```




**Residuals vs Fitted** Based on the residuals vs fitted plot we notice a funnel shap which means the variance increases as the fitted values are less which indicates unequal variance among the data. The residuals are also not evenly spread out which indicates that the model may not be a good fit as the data is not linear.

```
plot(linmodel, which=2)
```



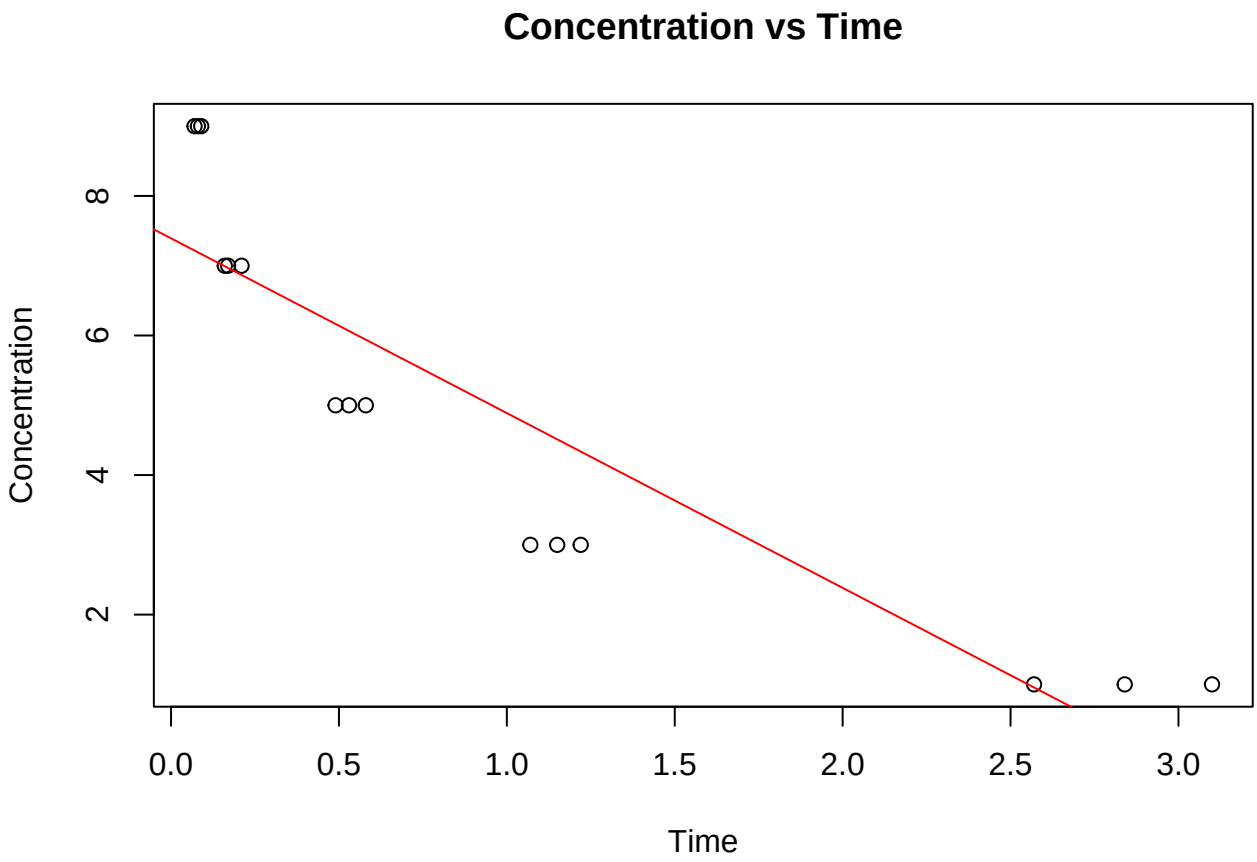

**QQ Residuals** Based on the Q-Q Residuals we notice that the residuals do not closely follow the line which indicates that the residuals are not normally distributed which is an important assumption for our linear model.

## Question 3

```
solution <- read.table("data/solution.txt", header = TRUE)
time <- solution$time #independent variable
concentration <- solution$concentration #response variable
```

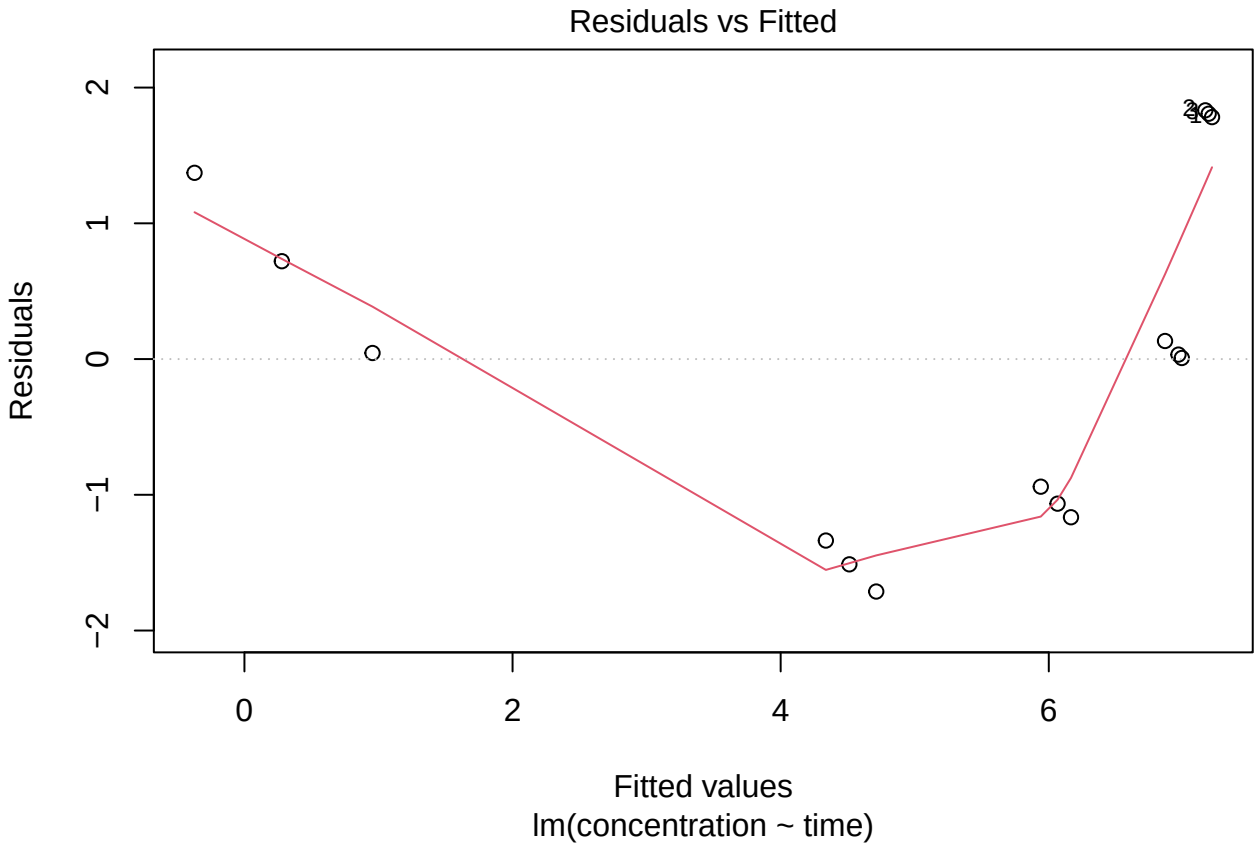
### 3. A.) Plot the data and fit a linear regression model.

```
fit1 <- lm(concentration ~ time)
plot(time, concentration,main = "Concentration vs Time", xlab = "Time", ylab = "Concentration")
abline(fit1, col="red")
```



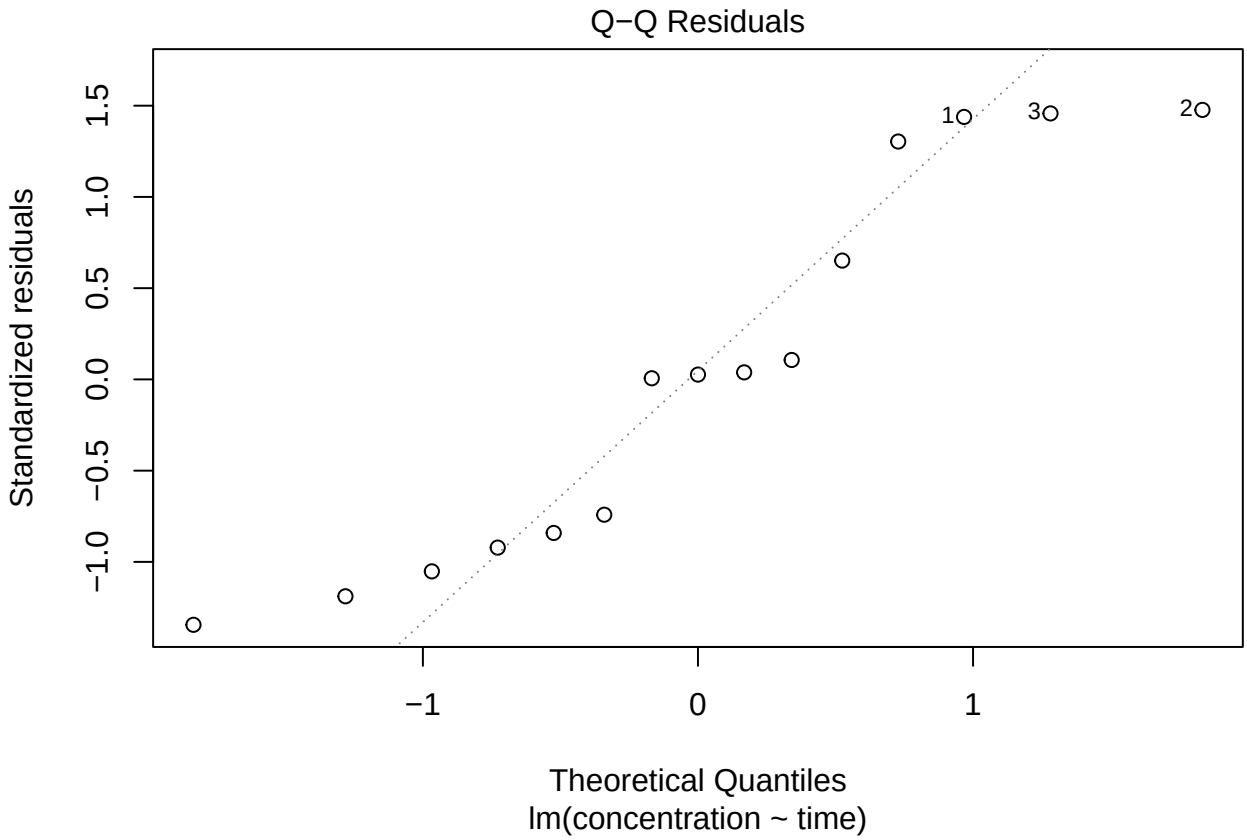

**3. B.) Construct diagnostic plots and comment on your findings.**

```
plot(fit1, which = 1)
```




**Conclusion** Based on the Residuals vs Fitted plot we can see that the residuals do not closely follow the model which indicates non-linearity and unequal variance. We must have the residuals randomly scattered at the horizontal line for the assumptions to be most likely met.

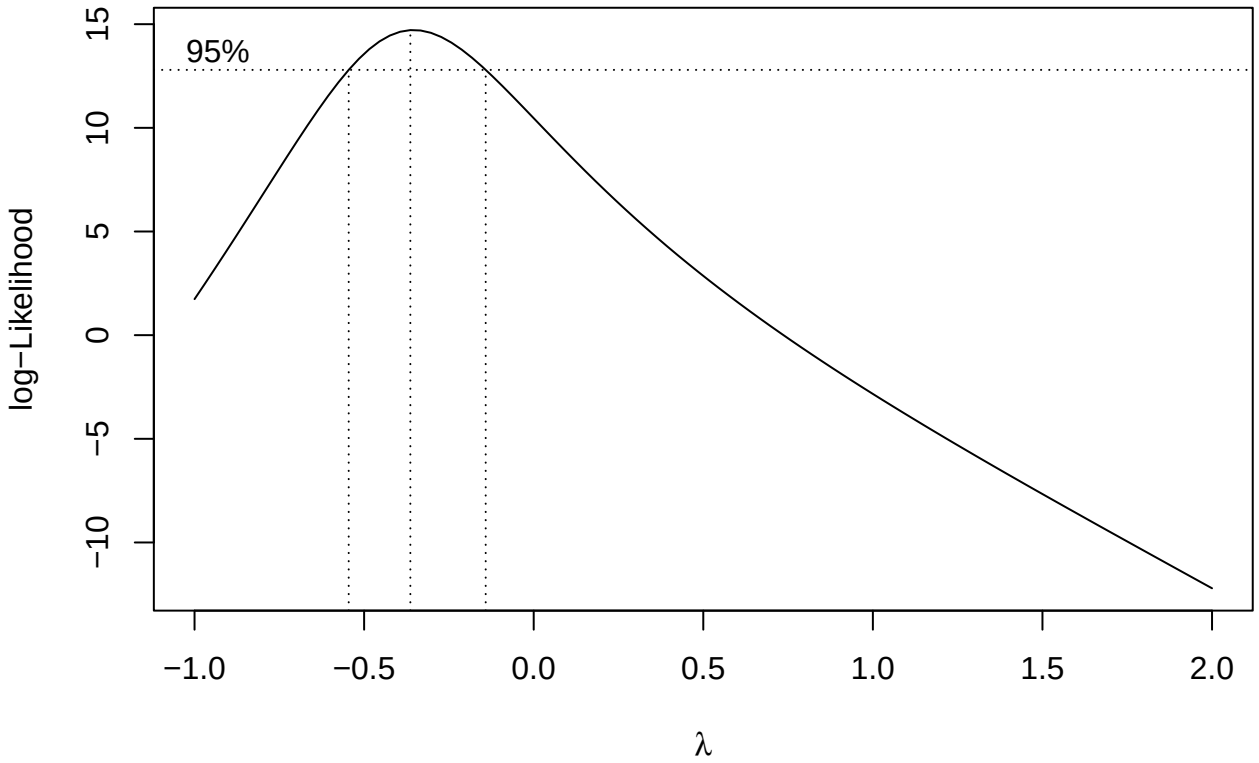
```
plot(fit1, which = 2)
```




**Conclusion** Based on the Q-Q Residuals plot the residuals do not closely follow the line which indicates that they may not be normally distributed which is an important assumption for our linear model.

### 3. C.) Use the Box-Cox procedure to find an appropriate power transformation.

```
boxcox(fit1, plotit=T, lambda = seq(-1,2,len=100))
```




**3. D.) Use the transformation Y = log10Y and obtain the estimated linear regression function for the transformed data.**

```
y = log10(concentration)
fit2 <- lm(y~time)
summary(fit2)
```

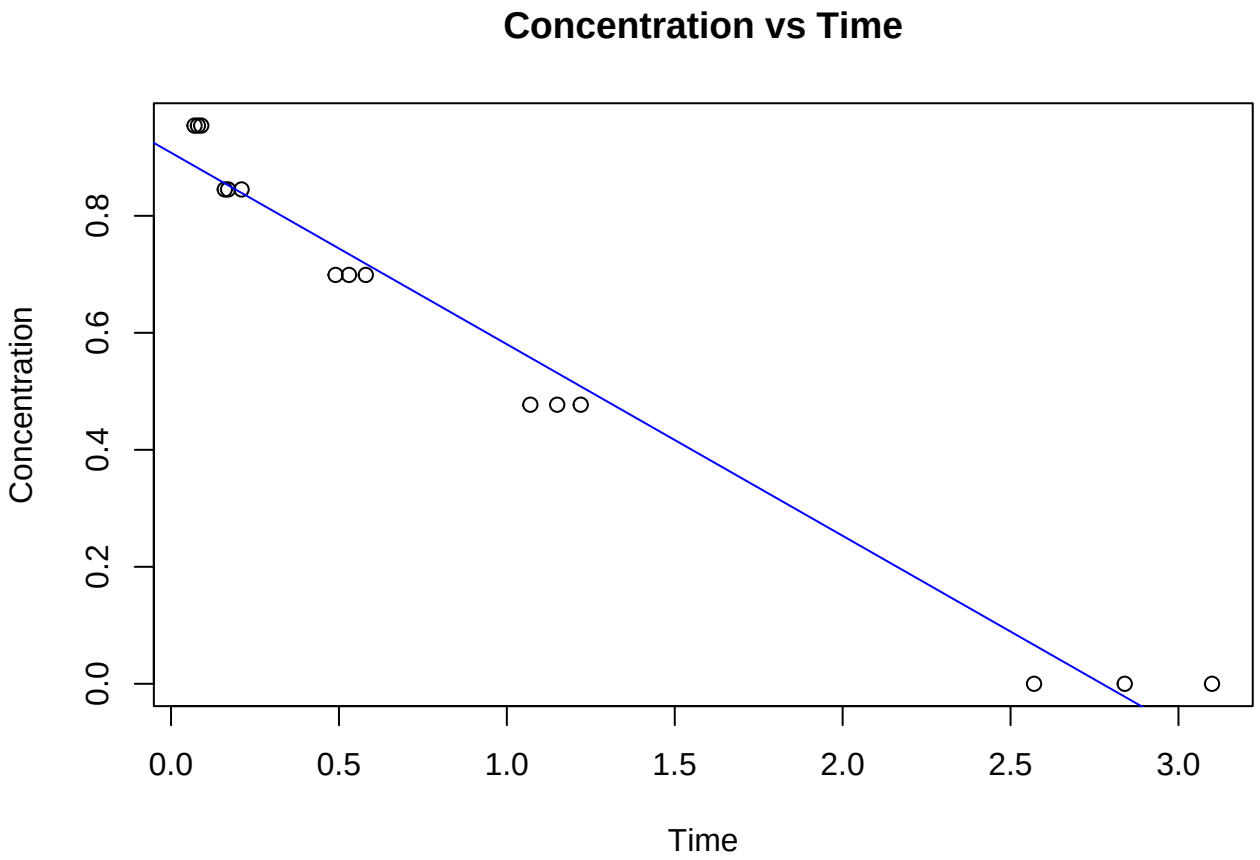
```
##
## Call:
## lm(formula = y ~ time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.0804 -0.0419 -0.0104  0.0457  0.1072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9079     0.0212    42.8  2.3e-15 ***
## time         -0.3274     0.0152   -21.5  1.5e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06 on 13 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.971
## F-statistic:  463 on 1 and 13 DF,  p-value: 1.51e-11
```

**Estimated Linear Regression Function:**

$$\hat{y} = 0.9079 - 0.3274(x)$$

### 3. E.) Plot transfermed data and the estimated regression line in (d). Does the regression

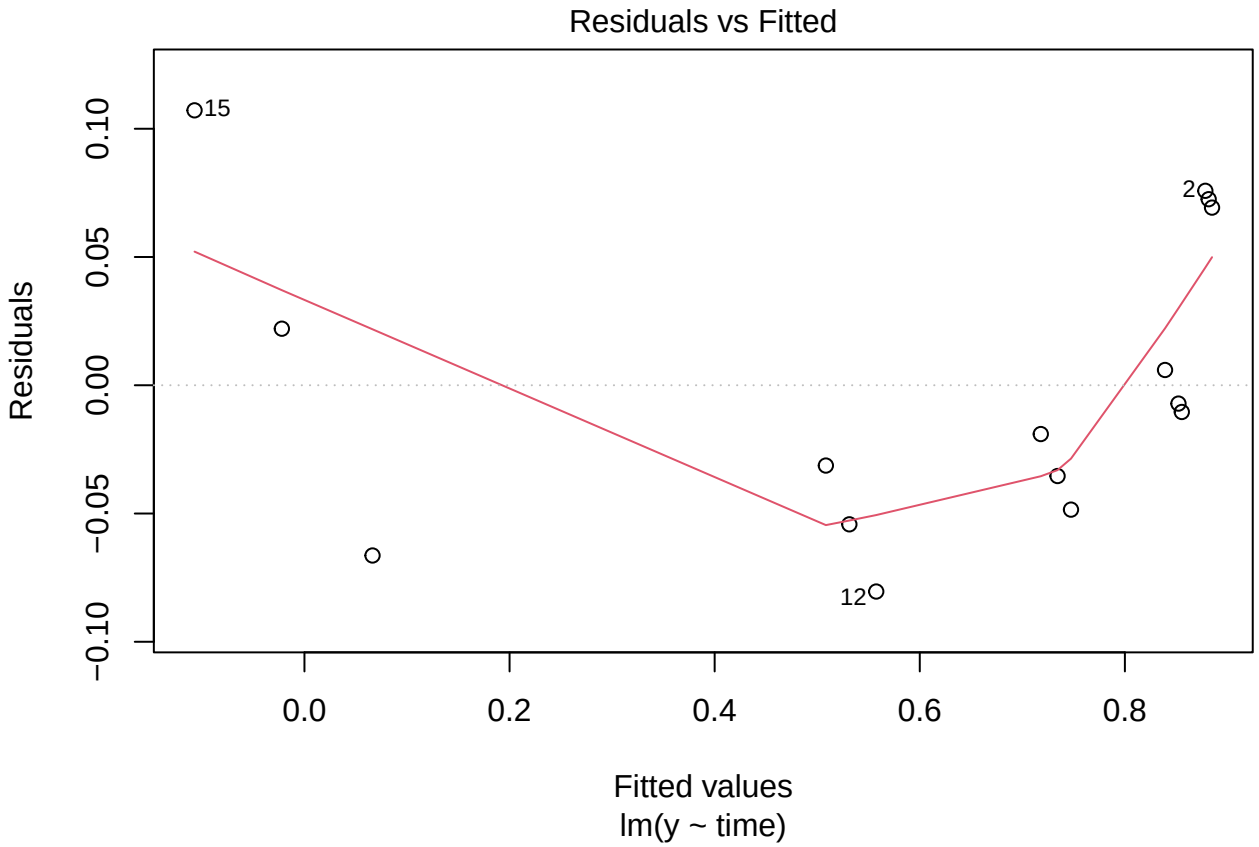
```
plot(time, y,main = "Concentration vs Time", xlab = "Time", ylab = "Concentration")
abline(fit2, col="blue")
```




**CONCLUSION:** The regression line appears to be a better fit to the transformed data than untransformed, but still misses a few points in the middle of the data.

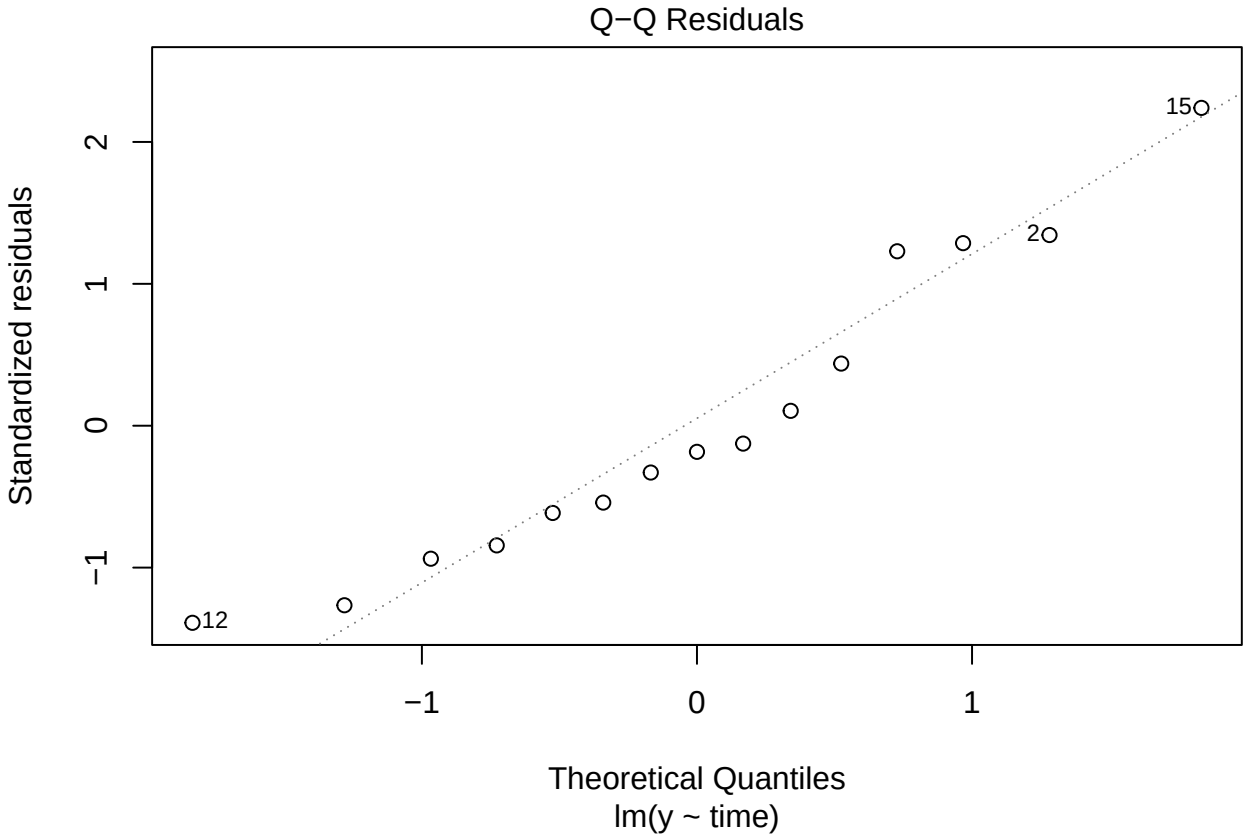
### 1. F.) Construct diagnostic plots for the fit in (d) and comments on your findings.

```
plot(fit2, which = 1)
```




**Residuals vs Fitted** Based on the plot the transformed data still expresses nonlinearity and non equal variances due to the curve in the plot

```
plot(fit2, which = 2)
```




**Q-Q Residuals** The residuals follow the line more closely than before which can indicate that it has a normal distribution, but it still does not follow the line too closely.

### 3. G.) Express the estimated regression function in (d) in the original units.

$$concentration = 0.9079 - 0.3274(time)$$