HW 3 peer review

Loading data firstly.

```
rm(list=ls())
data = read.csv('sleep.csv', header=TRUE, sep = ',')
colnames(data)[1] = "Species"
dim(data)
```

```
## [1] 42 11
```

```
head(data)
```

```
##                  Species   BodyWt BrainWt NonDreaming Dreaming TotalSleep
## 1 Africangiantpouchedrat    1.000     6.6         6.3      2.0        8.3
## 2          Asianelephant 2547.000  4603.0         2.1      1.8        3.9
## 3                 Baboon   10.550   179.5         9.1      0.7        9.8
## 4            Bigbrownbat    0.023     0.3        15.8      3.9       19.7
## 5         Braziliantapir  160.000   169.0         5.2      1.0        6.2
## 6                    Cat    3.300    25.6        10.9      3.6       14.5
##   LifeSpan Gestation Predation Exposure Danger
## 1      4.5        42         3        1      3
## 2     69.0       624         3        5      4
## 3     27.0       180         4        4      4
## 4     19.0        35         1        1      1
## 5     30.4       392         4        5      4
## 6     28.0        63         1        2      1
```

Total Variables

```
# Response Variables
NonDreaming <- data$NonDreaming
Dreaming <- data$Dreaming
TotalSleep <- data$TotalSleep #(for this assignment, we will not use TotalSleep)

#Continuous Variables
BodyWt <- data$BodyWt
BrainWt <- data$BrainWt
LifeSpan <- data$LifeSpan
Gestation <- data$Gestation

#Categorical Variables
Predation <- data$Predation
Exposure <- data$Exposure
Danger <- data$Danger
```

Using linear regression, you will investigate the association of the explanatory

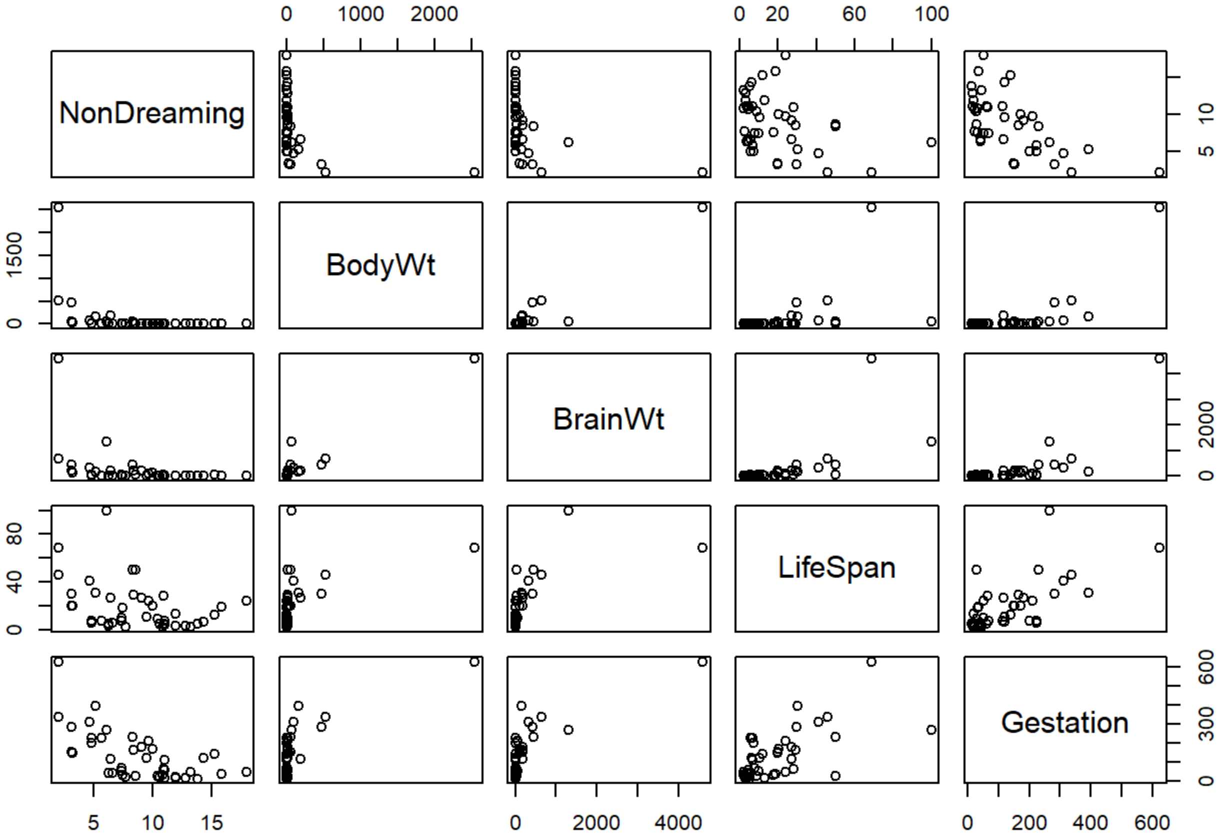
variables to the response variable NonDreaming first, and then repeat the homework questions using the second response variable, Dreaming.

Question 1: Exploratory Data Analysis.

1a. Using scatterplots, describe the relationship between NonDreaming and the continuous independent variables: BodyWt, BrainWt, LifeSpan and Gestation. Describe the general trend (direction and form).

Answer: Scatter plots show that there probably have strong linear relationship between NonDreaming and Gestation, which is negative relationship. There may have a weak relationship between NonDreaming and LifeSpan, the clusters of data points seems to be negative trend. In addition, it appears that BodyWt or BrainWt have no effects on NonDreaming, which suggestion there are no obvious linear relationships between them.

```
# For NonDreaming and continuous independent variables
plot(data[,c("NonDreaming", "BodyWt", "BrainWt", "LifeSpan", "Gestation")])
```



1b. Calculate and interpret the correlation coefficients for continuous variables

Answer: From the result of correlation coefficients, we found that there have strong linear correlations between NonDreaming and Gestation (-0.6061048), which is negative trend. In addition, there seems to have very weak linear correlations between NonDreaming and other three variables including BodyWt, Brainwt and LifeSpan based on the correlation coefficients, which are negative trend. All those outcome match the results from above scatter plots.

```
NonDreamingcor <- cor(data[c(2,3,7,8)],data[4])
NonDreamingcor
```

```
##            NonDreaming
## BodyWt      -0.3936373
## BrainWt     -0.3867947
## LifeSpan    -0.3722345
## Gestation   -0.6061048
```

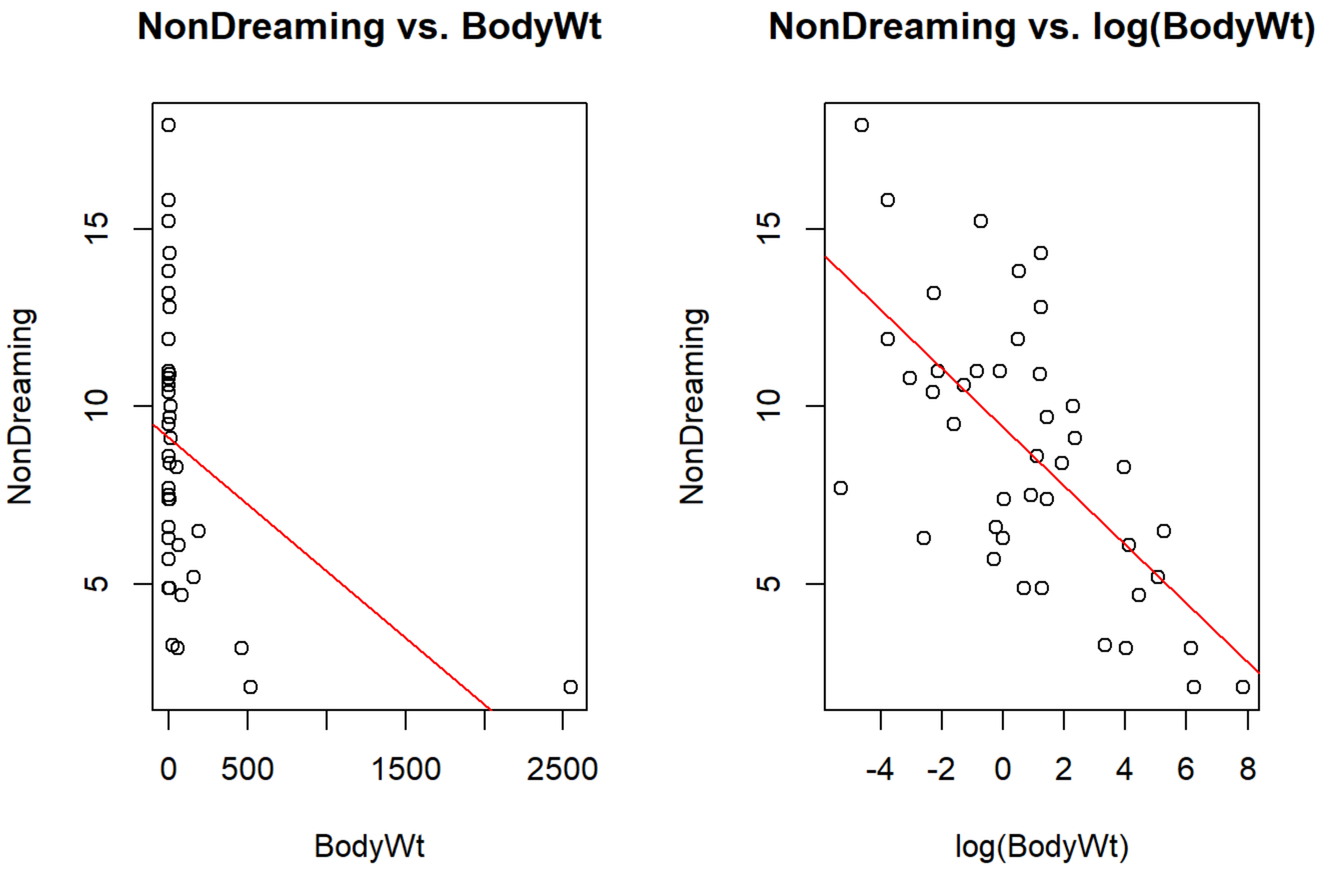
1c. Improving linearity: Using the initial scatterplots, are you able to visually validate the direction and strength of the correlation coefficients? If you see clusters of data points, try adding a directional line (abline) to the scatterplot by individually inspecting each predicting variable. You may need to transform the predicting continuous variable(s) to improve the linearity of the data. You can also transform the predicting variable NonDreaming, to improve linearity, although not required.

Answer: It is hard to draw clear conclusions based on initial scatterplots before transformation. For better linearity, we will improve them by transformation, such as using log for responose variable or predictor variable, and analyze them with individual plot and regression line as followings. After transformation, the following scatterplots of linearity have been greatly improved.

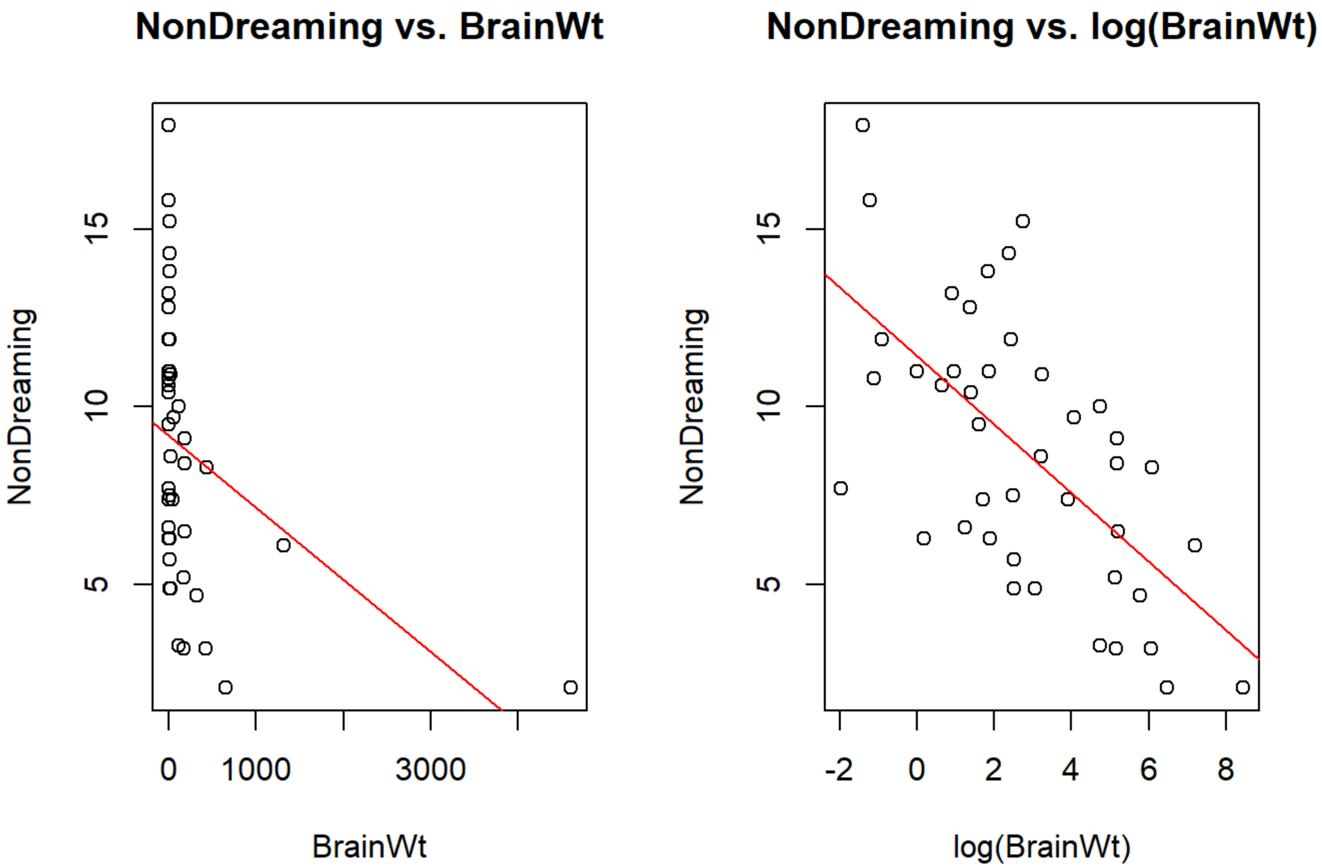
Visually inspect each continuous predicting variable. Include final plots for each variable in your report. Here is starter code for the first variable, BodyWt.

For NonDreaming variable:

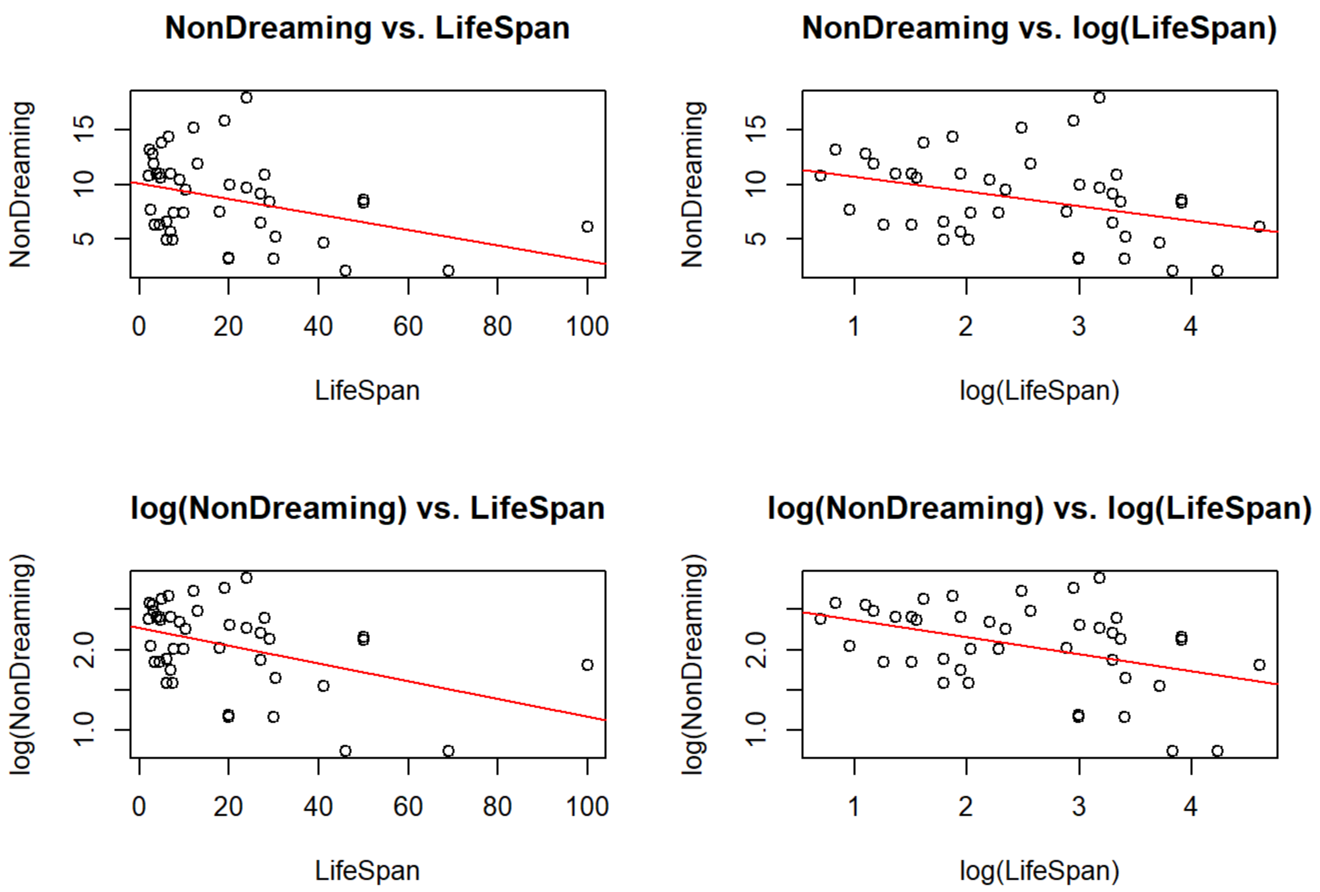
```
#### INspect BodyWt
par(mfrow = c(1,2))
{plot(BodyWt, NonDreaming, main = 'NonDreaming vs. BodyWt')
abline(lm(NonDreaming ~ BodyWt, data = data), col = 'red')}

#### Transform BodyWt
{plot(log(BodyWt), NonDreaming, main = 'NonDreaming vs. log(BodyWt)')
abline(lm(NonDreaming ~ log(BodyWt), data = data), col = 'red')}
```



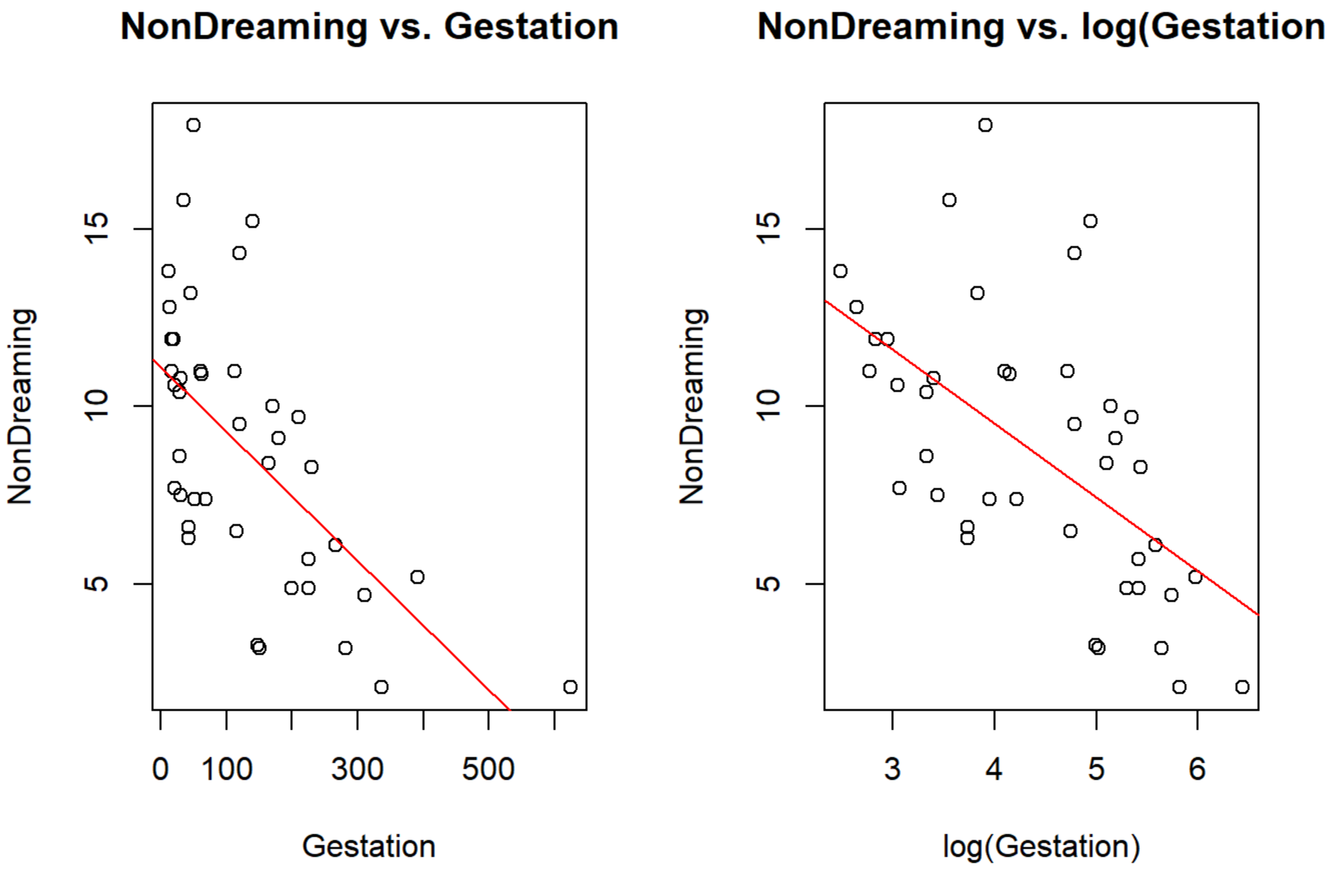
```
#### INspect BrainWt
par(mfrow = c(1,2))
{plot(BrainWt, NonDreaming, main = 'NonDreaming vs. BrainWt')
abline(lm(NonDreaming ~ BrainWt, data = data), col = 'red')}

#### Transform BrainWt
{plot(log(BrainWt), NonDreaming, main = 'NonDreaming vs. log(BrainWt)')
abline(lm(NonDreaming ~ log(BrainWt), data = data), col = 'red')}
```



```
#### Inspect LifeSpan
par(mfrow = c(2,2))
{plot(LifeSpan, NonDreaming, main = 'NonDreaming vs. LifeSpan')
abline(lm(NonDreaming ~ LifeSpan, data = data), col = 'red')}

#### Transform LifeSpan
{plot(log(LifeSpan), NonDreaming, main = 'NonDreaming vs. log(LifeSpan)')
abline(lm(NonDreaming ~ log(LifeSpan), data = data), col = 'red')}

{plot(LifeSpan, log(NonDreaming), main = 'log(NonDreaming) vs. LifeSpan')
abline(lm(log(NonDreaming) ~ LifeSpan, data = data), col = 'red')}

{plot(log(LifeSpan), log(NonDreaming), main = 'log(NonDreaming) vs. log(LifeSpan)')
abline(lm(log(NonDreaming) ~ log(LifeSpan), data = data), col = 'red')}
```



```
#### Transform Gestation
par(mfrow = c(1,2))
{plot(Gestation, NonDreaming, main = 'NonDreaming vs. Gestation')
abline(lm(NonDreaming ~ Gestation, data = data), col = 'red')}

#### Transform Gestation
{plot(log(Gestation), NonDreaming, main = 'NonDreaming vs. log(Gestation)')
abline(lm(NonDreaming ~ log(Gestation), data = data), col = 'red')}
```



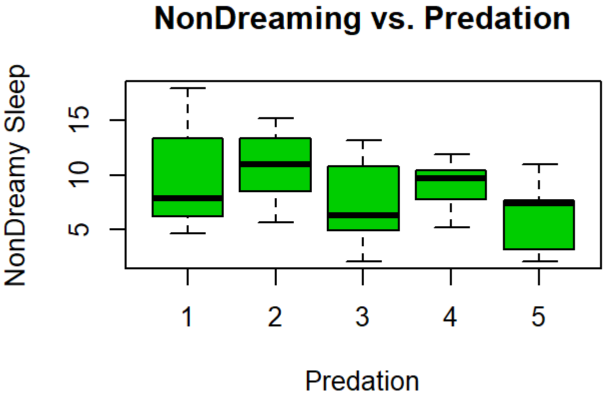
For Categorical variables:

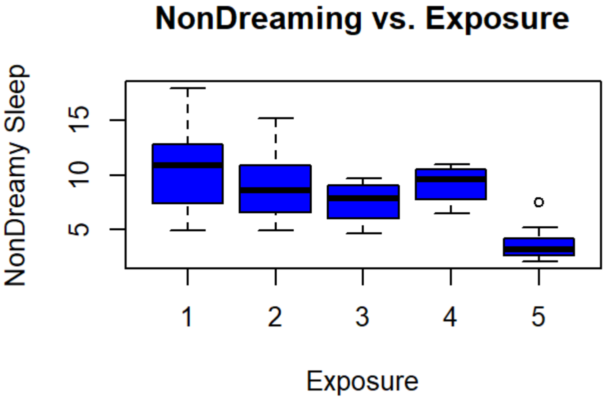
1d. Using boxplots, describe the relationship between NonDreaming and the categorical independent variables Predation, Exposure, and Danger. Does NonDreaming vary with the categorical variables?

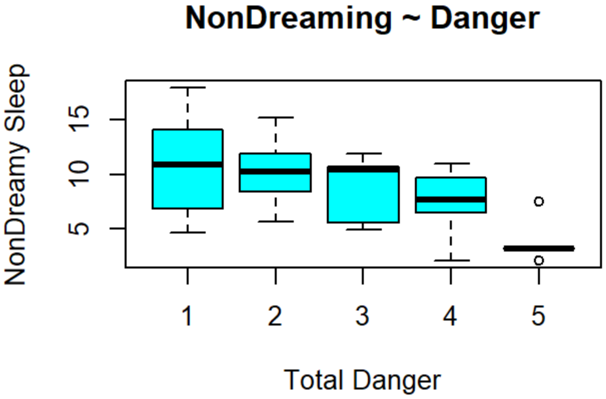
Answer: The criterion for judgement is that if the median lie of box A lies outside of box B entirely, then there is likely to be a difference between the two groups.So we can use this method for following analysis.

In overview, it is less between group variability from the results of each boxplot below, or the changes of NonDreaming are not significant with individual variables in the following boxplots except the type No.5. (No.5 is maximum for categorical variable Predation, No.5 is most exposed for Exposure, No.5 if most danger for Danger)

```
par(mfrow = c(2,2))
boxplot(NonDreaming ~ Predation, xlab = "Predation", ylab = "NonDreamy Sleep", main =
'NonDreaming vs. Predation', col = 3)

boxplot(NonDreaming ~ Exposure, xlab = "Exposure", ylab = "NonDreamy Sleep", main = '
NonDreaming vs. Exposure', col = 4)

boxplot(NonDreaming ~ Danger, xlab = "Total Danger", ylab = "NonDreamy Sleep", main =
'NonDreaming ~ Danger', col = 5)
```







1e. Based on this section for exploratory analysis, is it reasonable to assume a linear regression model? Would you suggest that NonDreaming varies with all or only some of the independent variables? Would you recommend using the categorical variables Predation, Exposure, and Danger in the model? Why?

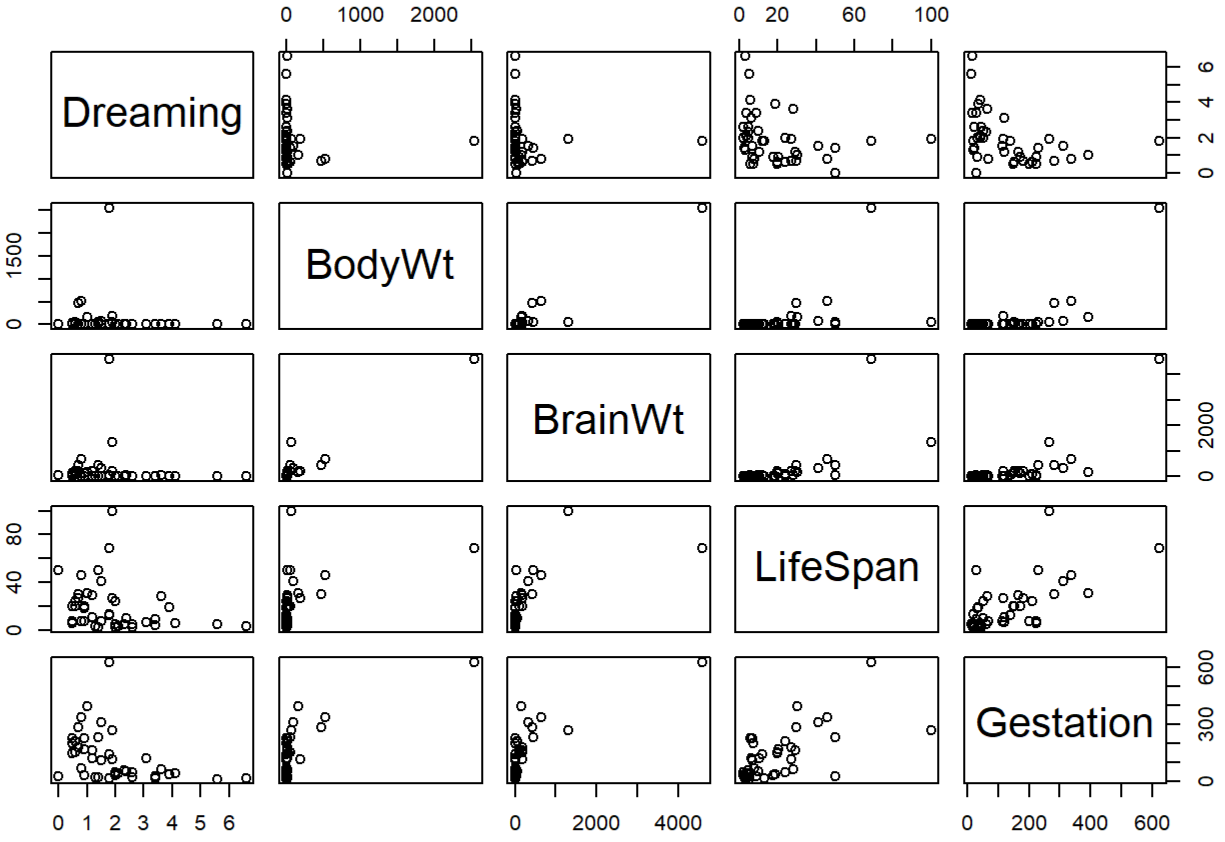
Answer: No. we can not consider it is reasonable to assume a linear regression model for the original analysis except relationship of NonDreaming and Gestation. However, if we have transfomed the variables, such as using log, there will have a very strong linear relationship between them. Therefore, without transformation, NonDreaming varies with only some of the independent variables such as Gestation. In addition, I don’t recommend using the categorical variables directly because the is less variability within them based on boxplots.

Question 2: Repeat Questions 1a, 1b, 1c, 1d, and 1e exercises for the second response variable Dreaming. Label your answers 2b, 2c, 2d, 2e.

2a. Using scatterplots, describe the relationship between NonDreaming and the continuous independent variables: BodyWt, BrainWt, LifeSpan and Gestation. Describe the general trend (direction and form).

Answer: Scatterplots indicate that there probably have strong linear relationship between Dreaming and Gestation, which is negative relationship. There may have a weak relationship between Dreaming and LifeSpan, the clusters of data points seems to be negative trend, there also have some outliers. In addition, it appears that BodyWt or BrainWt have no effects on Dreaming, which suggestion there are no obvious linear relationships between them. Scatterplots show as below.

```
# For Dreaming and continuous independent variables
plot(data[, c("Dreaming", "BodyWt", "BrainWt", "LifeSpan", "Gestation")])
```



2b. Calculate and interpret the correlation coefficients for continuous variables

Answer: From the results of correlation coefficients below, we found there probably have some linear correlations between Dreaming and Gestation, which is negative trend. It is weak linear correlation between Dreaming and LifeSpan, which is negative trend. In addition, there have no obvious linear correlation between Dreaming and other three variables including BodyWt, Brainwt and LifeSpan. Those results match the analysis from above scatterplot.

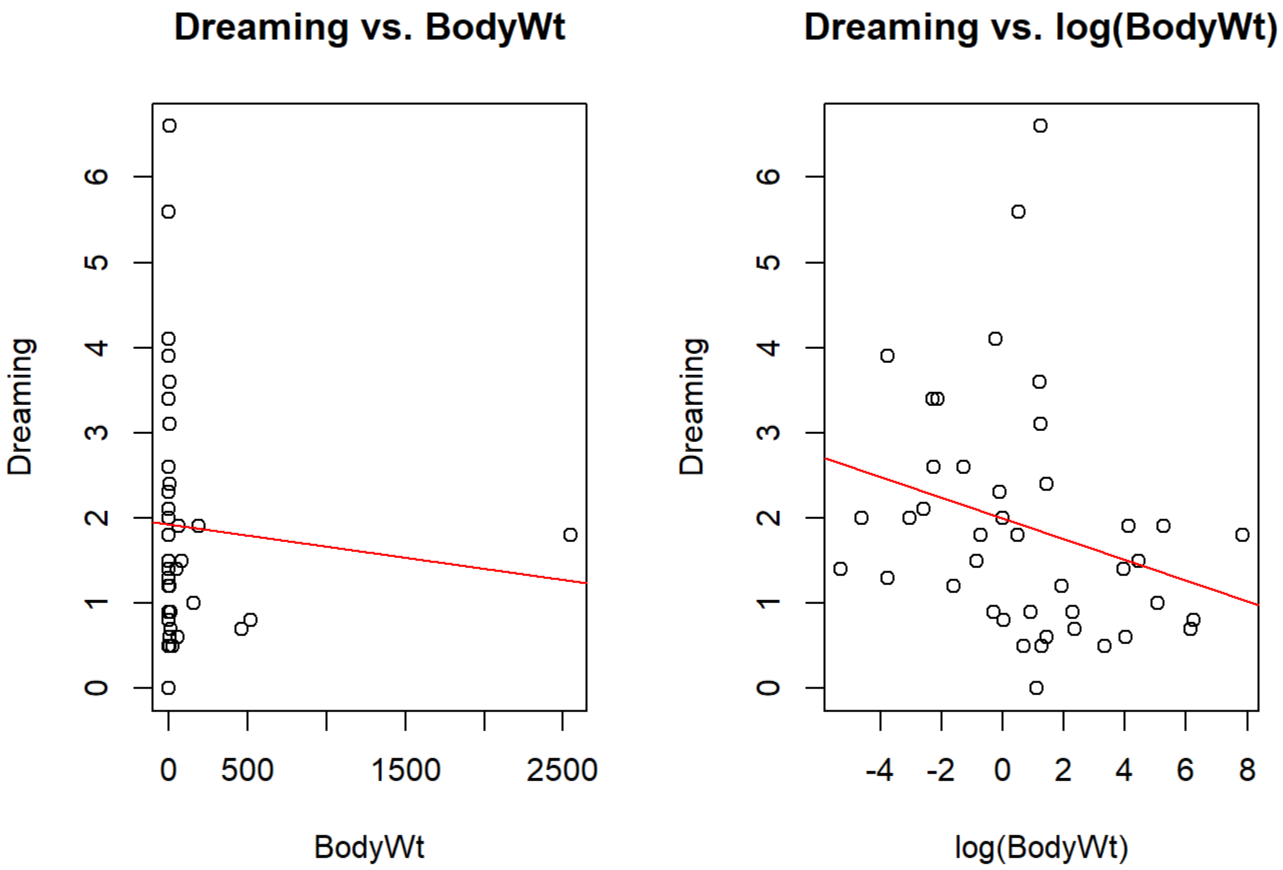
```
Dreamingcor <- cor(data[c(2,3,7,8)],data[5])
Dreamingcor
```

```
##              Dreaming
## BodyWt    -0.07488845
## BrainWt   -0.07427740
## LifeSpan  -0.26834006
## Gestation -0.40893177
```

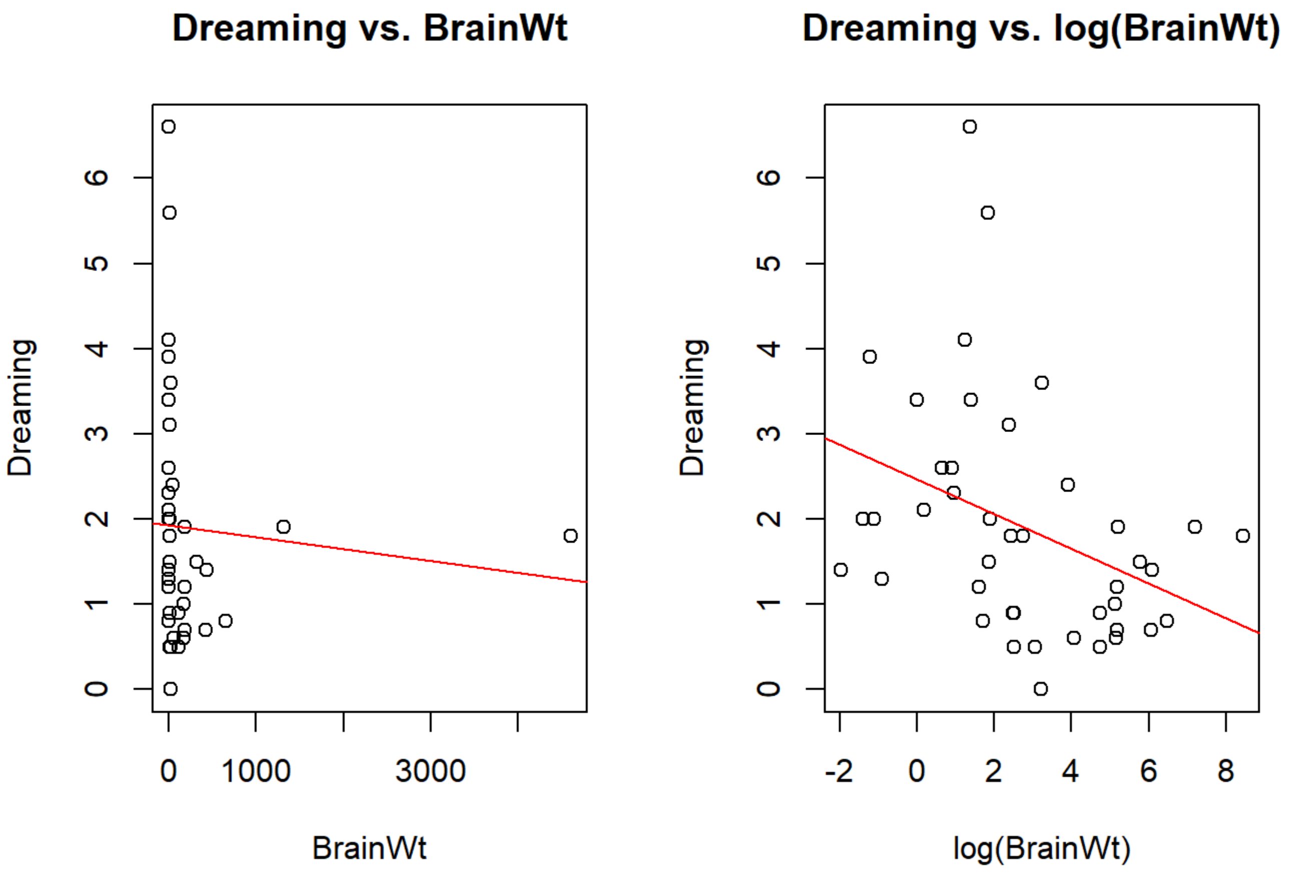
2c. Improving linearity: Using the initial scatterplots, are you able to visually validate the direction and strength of the correlation coefficients? If you see clusters of data points, try adding a directional line (abline) to the scatterplot by individually inspecting each predicting variable. You may need to transform the predicting continuous variable(s) to improve the linearity of the data. You can also transform the predicting variable Dreaming, to improve linearity, although not required.

Answer: If no transformation, it will be hard to draw clear conclusions based on initial scatterplots. For better linearity, we will improve them by transformation, such as using log for responose variable or predictor variables, and analyze them with individual plot and regression line as followings. After transformation, the following scatterplots of linearity have been greatly improved.

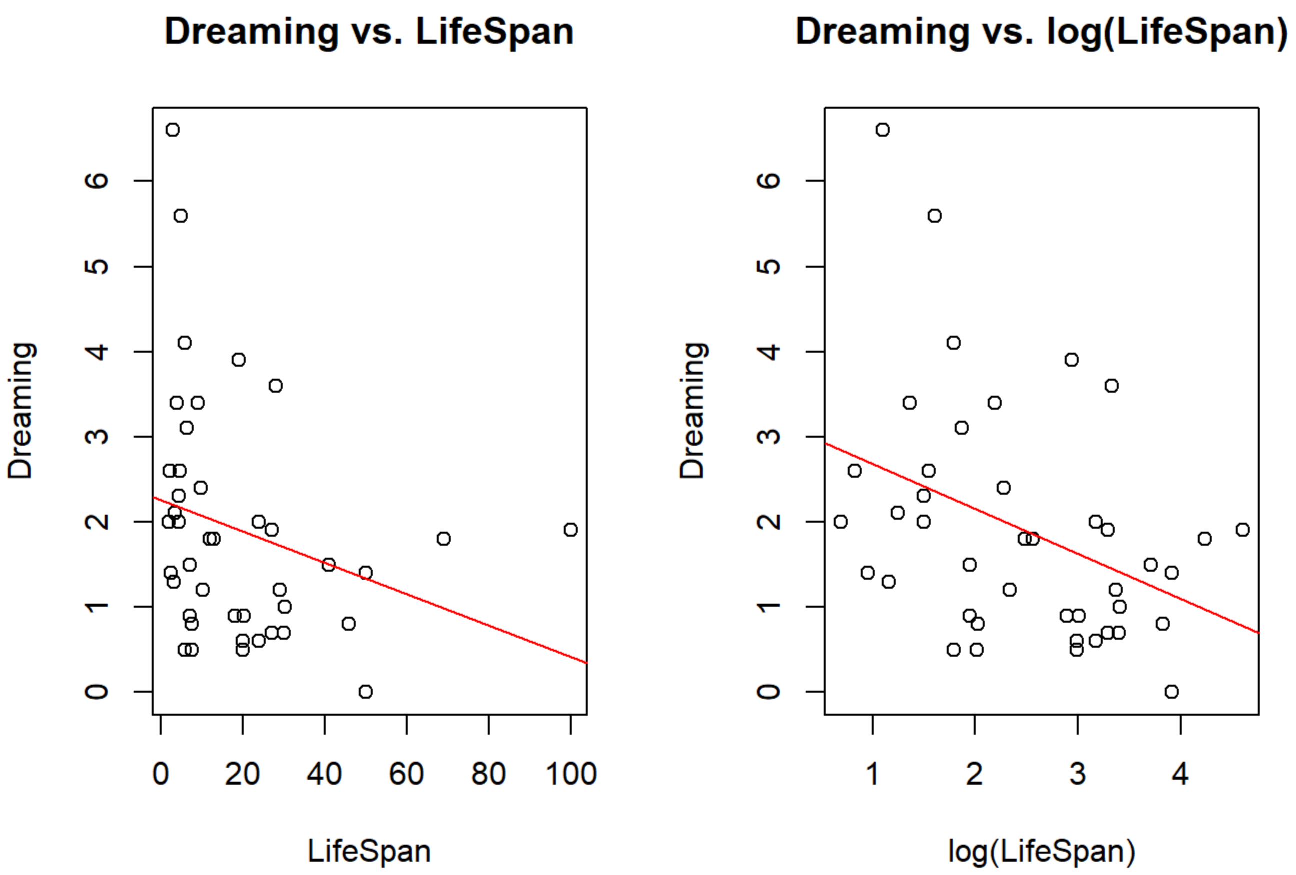
```
#### Inspect BodyWt
par(mfrow = c(1,2))
{plot(BodyWt, Dreaming, main = 'Dreaming vs. BodyWt')
abline(lm(Dreaming ~ BodyWt, data = data), col = 'red')}

#### Transform BodyWt
{plot(log(BodyWt), Dreaming, main = 'Dreaming vs. log(BodyWt)')
abline(lm(Dreaming ~ log(BodyWt), data = data), col = 'red')}
```



```
#### Inspect BrainWt
par(mfrow = c(1,2))
{plot(BrainWt, Dreaming, main = 'Dreaming vs. BrainWt')
abline(lm(Dreaming ~ BrainWt, data = data), col = 'red')}
#### Transform BrainWt
{plot(log(BrainWt), Dreaming, main = 'Dreaming vs. log(BrainWt)')
abline(lm(Dreaming ~ log(BrainWt), data = data), col = 'red')}
```

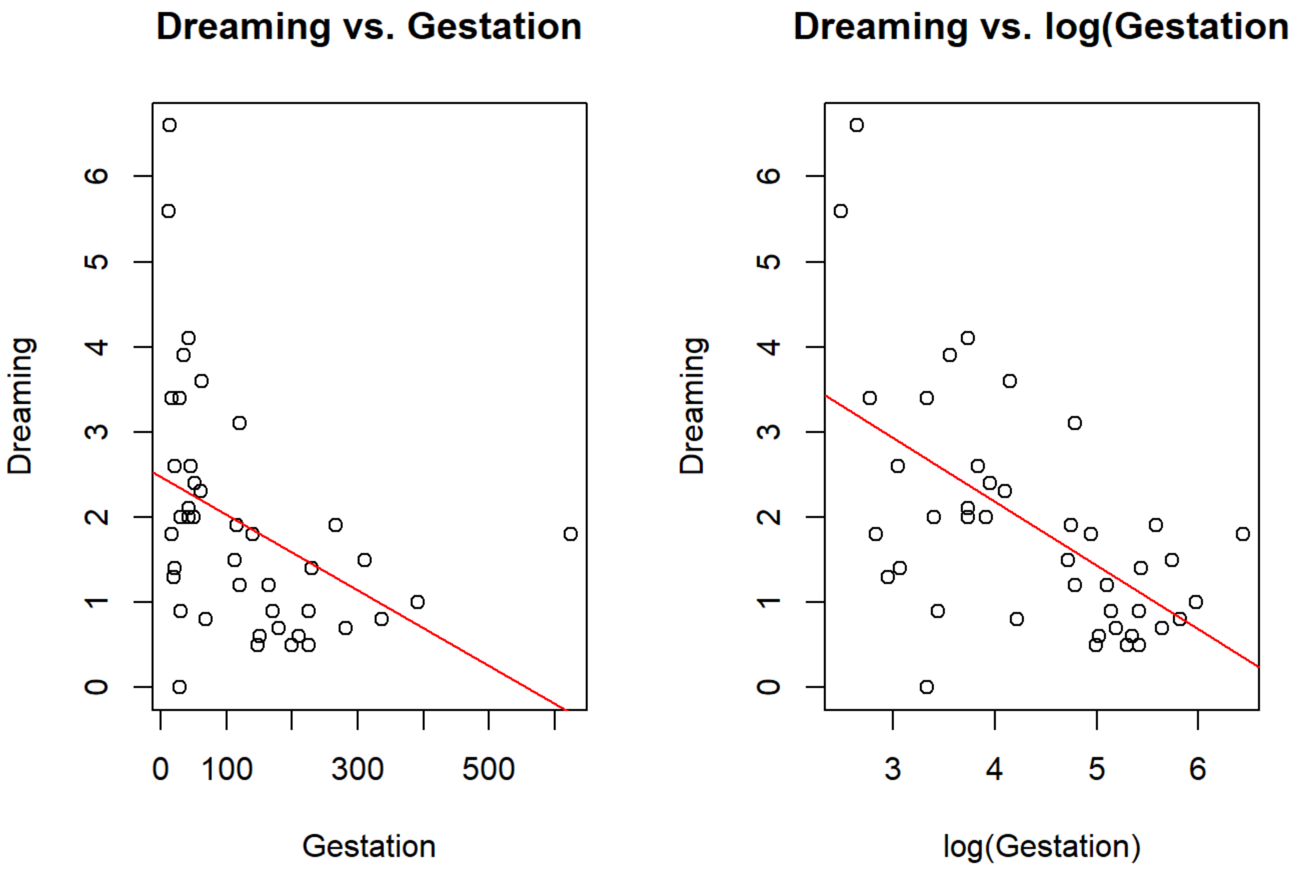


```
#### Inspect LifeSpan
par(mfrow = c(1,2))
{plot(LifeSpan, Dreaming, main = 'Dreaming vs. LifeSpan')
abline(lm(Dreaming ~ LifeSpan, data = data), col = 'red')}

#### Transform LifeSpan
{plot(log(LifeSpan), Dreaming, main = 'Dreaming vs. log(LifeSpan)')
abline(lm(Dreaming ~ log(LifeSpan), data = data), col = 'red')}
```


Dreaming vs. LifeSpan

Dreaming vs. log(LifeSpan)

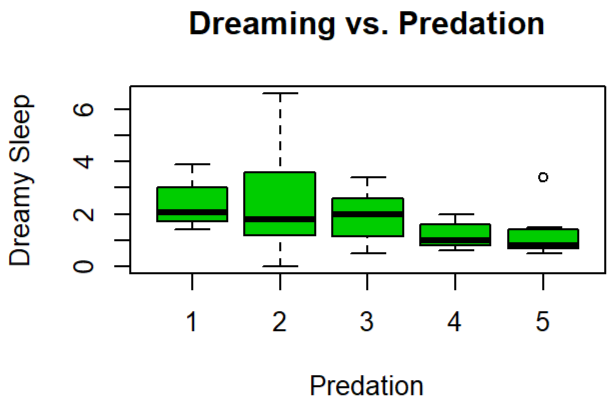
```
#### Transform Gestation
par(mfrow = c(1,2))
{plot(Gestation, Dreaming, main = 'Dreaming vs. Gestation')
abline(lm(Dreaming ~ Gestation, data = data), col = 'red')}

#### Transform Gestation
{plot(log(Gestation), Dreaming, main = 'Dreaming vs. log(Gestation')
abline(lm(Dreaming ~ log(Gestation), data = data), col = 'red')}
```

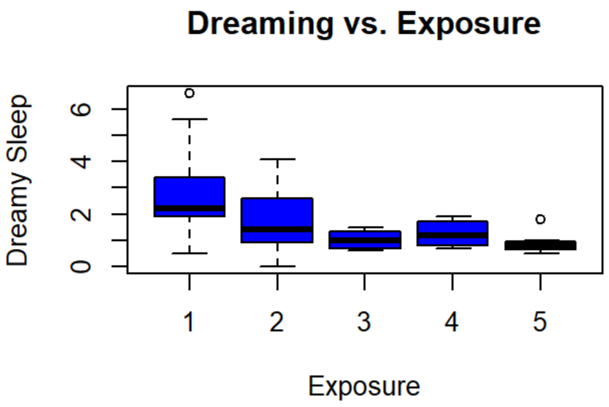


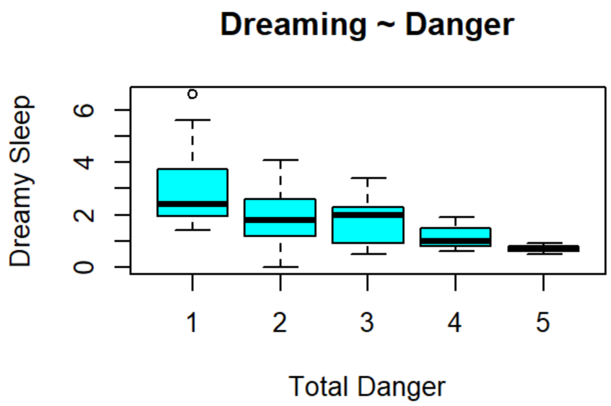
2d. Using boxplots, describe the relationship between Dreaming and the categorical independent variables Predation, Exposure, and Danger. Does Dreaming vary with the categorical variables?

Answer: In overview, the Dreaming decreases when the degrees of categorical variables increases such from minimum Predation to maximum, from least Exposure to most, from lease Danger to most. There appears to have many changes of Dreaming with the categorical variables, and many changes are significant, such as the variability between Dreaming and expsoure, Dreaming and Danger. The boxplots show below.

```
par(mfrow = c(2,2))
boxplot(Dreaming ~ Predation, xlab = "Predation", ylab = "Dreamy Sleep", main = 'Dreaming vs. Predation', col = 3)

boxplot(Dreaming ~ Exposure, xlab = "Exposure", ylab = "Dreamy Sleep", main = 'Dreaming vs. Exposure', col = 4)

boxplot(Dreaming ~ Danger, xlab = "Total Danger", ylab = "Dreamy Sleep", main = 'Dreaming ~ Danger', col = 5)
```







#### **2e.** Based on this section for exploratory analysis, is it reasonable to assume a linear regression model? Would you suggest that Dreaming varies with all or only some of the independent variables? Would you recommend using the categorical variables Predation, Exposure, and Danger in the model? Why? #### **Answer:** No. We can not assume a linear regression model because there almost no linear relationship between them…

Answer: No. we can not consider it is reasonable to assume a linear regression model for the original analysis except relationship of Dreaming and Gestation. However, if we have transfomed the variables, such as using log, there will have a very strong linear relationship between them. Therefore, without transformation, Dreaming varies with only some of the independent variables such as Gestation. In addition, I may recommend using the categorical variables because the trendency is very obvious, and many changes are significant such as variability between Dreaming and Exposure, Dreaming and Danger based on above boxplots.

Question 3: Fitting the Linear Regression Model.

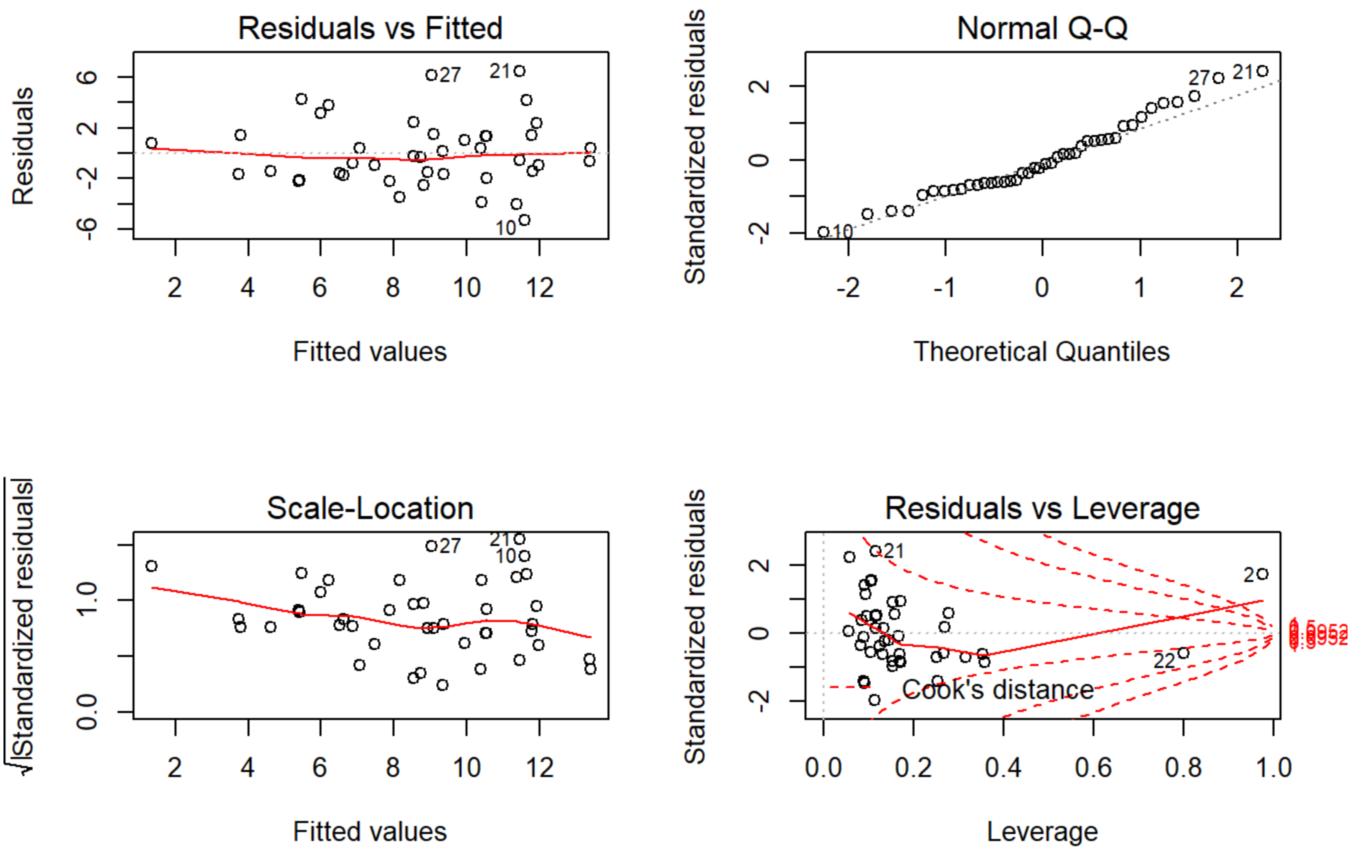
Plot the full model for NonDreaming without transforming the response variable or predicting variables. Remember to exclude the two response variables for sleep and the Species column.

```
model3 <- lm(NonDreaming ~. -Species -Dreaming -TotalSleep, data = data)
summary(model3)
```

```
##
## Call:
## lm(formula = NonDreaming ~ . - Species - Dreaming - TotalSleep,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2864 -1.6503 -0.4501  1.4037  6.4473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.331488   1.256475  10.610  2.5e-12 ***
## BodyWt       0.003332   0.005568   0.598   0.5535
## BrainWt     -0.001294   0.003342  -0.387   0.7010
## LifeSpan    -0.001181   0.043509  -0.027   0.9785
## Gestation   -0.013804   0.006563  -2.103   0.0429 *
## Predation    1.414774   1.027350   1.377   0.1775
## Exposure     0.224418   0.643644   0.349   0.7295
## Danger      -2.799115   1.275630  -2.194   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.857 on 34 degrees of freedom
## Multiple R-squared:  0.5404, Adjusted R-squared:  0.4458
## F-statistic: 5.711 on 7 and 34 DF,  p-value: 0.0002014
```

```
par(mfrow = c(2,2))
plot(model3, cook.levels = c(4/42,0.5,1))
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



1. What are the model parameters and what are their estimates?

Answer: This is multiple linear model. Residual standard error is 2.857 on 34 degrees of freedom, multiple R-squared is 0.5404, which means 54.04% variability can be explained. F-statistic is 5.711 on 7 and 34 degrees of freedom (n-p-1 = 42-7-1).

The estimate b0 (intercept) is 13.331488, b1 coefficient for variable BodyWt is 0.003332, estimate of variable BrainWt is -0.001294, estimate of variable LifeSpan is -0.001181, estimate of variable Gestation is -0.013804, estimate of variable Predation is 1.414774, estimate of variable Exposure is 0.224418, estimate of variable Danger is -2.799115

2. What is the equation for the regression line?

Answer: Equation of regression line:

NonDreaming = 13.331488+0.003332*BodyWt-0.001294*BrainWt -0.001181*LifeSpan-0.013804*Gestation+1.414774*Predation+0.224418*Exposure-2.799115*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The variables of Gestation (p-value = 0.0429) and Danger (p-value = 0.0351)

4. Interpret the estimated value of the parameters, including the error term, corresponding to BodyWt and Predation in the context of the problem.

Answer: The estimated intercept is 13.331488, which is an estimate the expected value of response variable (NonDreaming) when all the predictor variables equal zero. In addition, when other variables are hold in the fixed model, if BodyWt increases by 1 unit, the NonDreaming will increase by 0.003332 units, the standard error of estimate is 0.005568, p-value is 0.5535. When other variables are hold in the fixed model, if Predation increased by 1 unit, the NonDreaming will increase by 1.414774 units, the standard error of estimate is 1.027350, p-value is 0.1775.

5. Check the assumptions of the model through plotting. Note potential outliers, if any.

Answer: (1). To evaluate constant variance and independence assumption, we can use scatter plot of Residuals vs Fitted, we found that most of residuals are scatted around the 0 line, which indicates that we do have constant variance. In addition, we found several possible outlier points, which are points of No.10, 21 and 27.

(2). To evaluate the normality we can use Normal QQ plot, we found that most of residuals distrubute along a straight line except two ends, the distrubtion of residuals have a little deviate from the straight line at two tails (a little right skewed), there appears to have outlier points of No.10, 21 and 27.

(3) For plots of Scale_Location, sqrt of standard resuduals and fitted values can be evaluate the costant variance or independence, we found severl potential outlier such as points of No.10, 21, 27. In addition, from plot of Residuals vs leverage, cook’s distance are useful to identify these outliers, we found severl potential outlier such as points of No.2, 21 and 22.

3a. Change model3 to log transform the response variable, NonDreaming.

```
model3a <- lm(log(NonDreaming) ~. -Species -Dreaming -TotalSleep, data = data)
summary(model3a)
```

```
##
## Call:
## lm(formula = log(NonDreaming) ~ . - Species - Dreaming - TotalSleep,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60553 -0.20880  0.03197  0.16087  0.60877
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6679033  0.1503207  17.748   <2e-16 ***
## BodyWt      -0.0002651  0.0006661  -0.398   0.6931
## BrainWt      0.0001303  0.0003998   0.326   0.7464
## LifeSpan    -0.0022691  0.0052053  -0.436   0.6657
## Gestation   -0.0018214  0.0007851  -2.320   0.0265 *
## Predation    0.1608420  0.1229089   1.309   0.1994
## Exposure     0.0224126  0.0770035   0.291   0.7728
## Danger      -0.3207756  0.1526124  -2.102   0.0430 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3418 on 34 degrees of freedom
## Multiple R-squared:  0.6418, Adjusted R-squared:  0.5681
## F-statistic: 8.703 on 7 and 34 DF,  p-value: 4.29e-06
```

1. What are the model parameters and what are their estimates?

Answer: This model is Log-Linear Model. Residual standard error is 0.3418 on 34 degrees of freedom (n-p-1), multiple R-squared is 0.6418, which means 64.18% variability can be explained. F-statistic is 8.703 on 7 and 34 degrees of freedom (n-p-1 = 42-7-1).

The estimate b0 (intercept) is 2.6679033, b1 coefficient for variable BodyWt is -0.0002651, estimate of variable BrainWt is 0.0001303, estimate of variable LifeSpan is -0.0022691, estimate of variable Gestation is -0.0018214, estimate of variable Predation is .1608420, estimate of variable Exposure is 0.0224126, estimate of variable Danger is -0.3207756.

2. What is the equation for the regression line?

Answer: log(NonDreaming) = 2.6679033-0.0002651*BodyWt+0.0001303*BrainWt -0.0022691*LifeSpan-0.0018214*Gestation+0.1608420*Predation+0.0224126*Exposure-0.3207756*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: predicting variables of Gestation (p-value=0.0265) and Danger (p-value = 0.0430).

4. Interpret the estimated value of the parameters, including the error term,

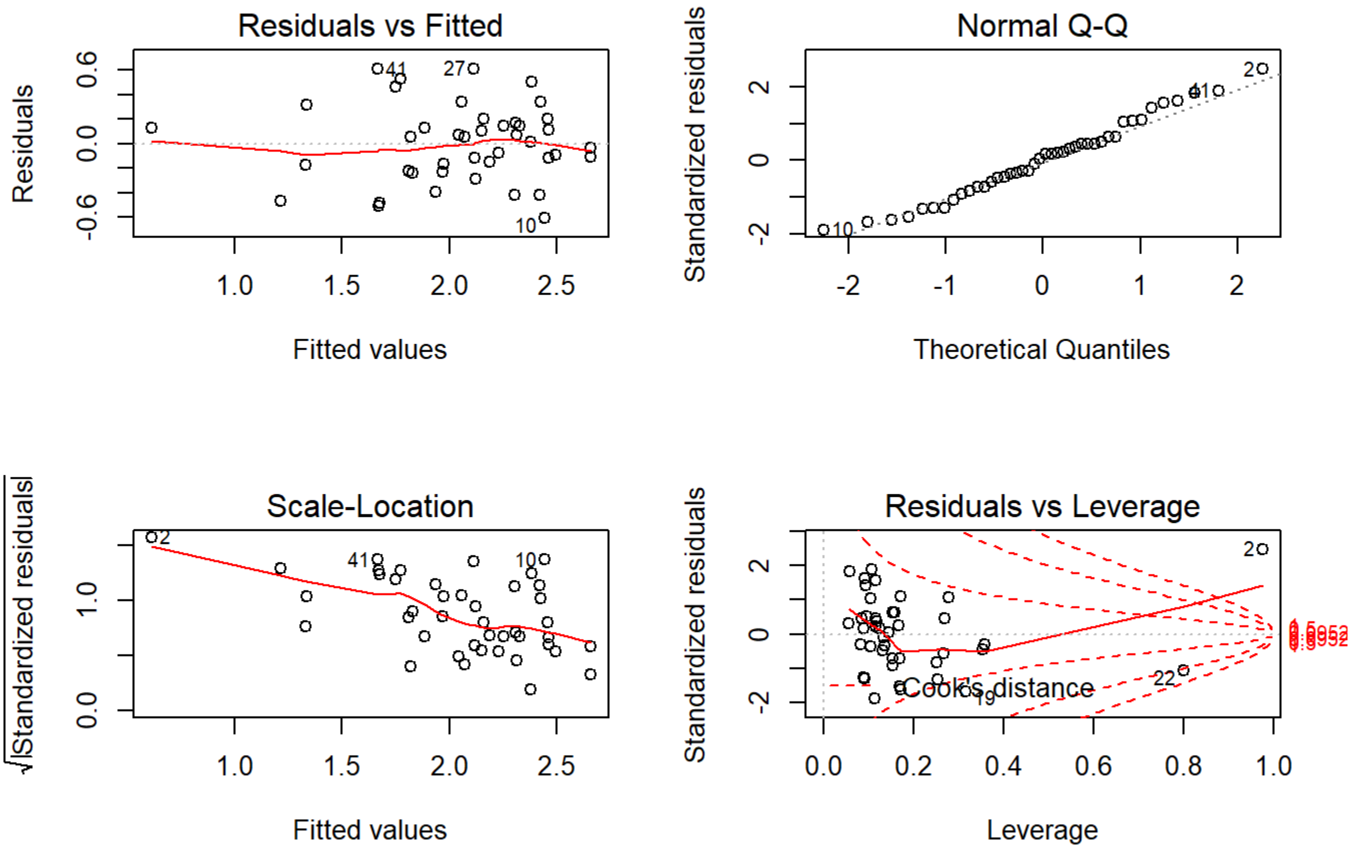
corresponding to BodyWt and Predation in the context of the problem.

Answer: In this Log-Linear model, when other variables are hold in the fixed model, if BodyWt increases by 1 unit, the NonDreaming will decrease by 0.02651%. the std. Error of estimate is 0.0006661 an p-value is 0.6931. In addition, when other variables are hold in the fixed model, if Predation increases by 1 unit, the NonDreaming will increase by 16.08420%. the std.Error of estimate is 0.1229089 an p-value is 0.1994.

5. Check the assumptions of the model through plotting. Note potential outliers, if any.

```
par(mfrow = c(2,2))
plot(model3a, cook.levels = c(4/42, 0.5, 1))
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Answer: (1). To evaluate constant variance and independence assumption, we can use scatter plot of Residuals vs Fitted, we found that most of residuals are scatted around the 0 line, which indicates that we do have constant variance. In addition, we found several possible outlier points, which are points of No.10, 27 and 41.

(2). To evaluate the normality we can use Normal QQ plot, we found that most of residuals distrubute along a straight line except two ends, the distrubtion of residuals have a little deviate from the straight line at two tails (a little right skewed), there appears to have outlier points of No.1, 2 and 10.

(3) For plots of Scale_Location, sqrt of standard resuduals and fitted values can be evaluate the costant variance or independence, we found severl potential outlier such as points of No.2, 10 and 41, which is deviate from the horizon line. In addition, cook’s distance are useful to identify these outliers. From plot of Residuals vs leverage, we found severl potential outlier such as points of No.1, 2 and 22.

3b. Change model3a to remove the log transform of NonDreaming, and add the log transformation of numeric response variables BrainWt, BodyWt, LifeSpan and Gestation.

```
model3b <- lm(NonDreaming ~ log(BodyWt) +log(BrainWt) +log(LifeSpan) +log(Gestation)
+Exposure +Predation +Danger, data = data)
summary(model3b)
```

```
## 
## Call:
## lm(formula = NonDreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + 
##     log(Gestation) + Exposure + Predation + Danger, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.0359 -1.4743 -0.1921  1.8385  5.8191 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     11.4819     2.8885   3.975 0.000348 ***
## log(BodyWt)     -0.3917     0.4787  -0.818 0.418974    
## log(BrainWt)    -0.5591     0.7019  -0.797 0.431228    
## log(LifeSpan)    1.2991     0.7492   1.734 0.091993 .  
## log(Gestation)  -0.5083     0.6613  -0.769 0.447419    
## Exposure         0.5310     0.6196   0.857 0.397516    
## Predation        1.6141     0.9812   1.645 0.109177    
## Danger          -2.9313     1.1628  -2.521 0.016562 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.58 on 34 degrees of freedom
## Multiple R-squared:  0.6253, Adjusted R-squared:  0.5481 
## F-statistic: 8.105 on 7 and 34 DF,  p-value: 8.706e-06
```

1. What are the model parameters and what are their estimates?

Answer: This model is Linear-Log model. In the model, Residual standard error is 2.58 on 34 degrees of freedom (n-p-1), multiple R-squared is 0.6253, which means 62.53% variability can be explained. F-statistic is 8.105 on 7 and 34 degrees of freedom (n-p-1 = 42-7-1).

The estimate b0 (intercept) is 11.4819, b1 coefficient for variable log(BodyWt) is -0.3917, estimate of variable log(BrainWt) is -0.5591, estimate of variable log(LifeSpan) is 1.2991, estimate of variable log(Gestation) is -0.5083, estimate of variable Exposure is 0.5310, estimate of variable Predation is 1.6141, estimate of variable Danger is -2.9313.

2. What is the equation for the regression line?

Answer: Equation of this model:

NonDreaming = 11.4819-0.3917*log(BodyWt)-0.5591*log(BrainWt)+1.2991*log(LifeSpan)-0.5083*log(Gestation)+0.5310*Exposure+1.6141*Predation-2.9313*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The predictor variable of Danger (p-value = 0.016562).

4. Interpret the estimated value of the parameters, including the error term, corresponding to log(BodyWt) and Predation in the context of the problem.

Answer: In this Linear-Log model, when other variables are hold in the fixed model, if BodyWt increases by 1%, the NonDreaming will decrease by 0.003917 unit. the std.Error of estimate is 0.4787 an p-value is 0.418974. In addition, when other variables are hold in the fixed model, if Predation increases by 1 unit, the NonDreaming will increase by 1.6141 units. the std.Error of estimate is 0.9812 and p-value is 0.109177.

5. Did model3b improve over model3a? Explain how you determined if the model improved or not.

Answer: We found that the differences are very small between model3a and model3b according to model parameters such as R-squared, F-statistic value, p-value and other factors. For example: R-square of model3a is 0.6418, R-square of model3b is 0.6253.

3c. Because the Danger variable is an interpolation of the Exposure and Predation variables, let's keep Danger and remove the other two from the model using model3b as your baseline.

```
model3c <- lm(NonDreaming ~ log(BodyWt) +log(BrainWt) +log(LifeSpan) +log(Gestation)
+Danger, data = data)
summary(model3c)
```

```
##
## Call:
## lm(formula = NonDreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) +
##     log(Gestation) + Danger, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6447 -1.7321  0.0363  1.3016  5.5696
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.0446     2.5383   5.139 9.82e-06 ***
## log(BodyWt)     -0.4294     0.4663  -0.921   0.3633
## log(BrainWt)    -0.4437     0.7008  -0.633   0.5307
## log(LifeSpan)    1.0811     0.6867   1.574   0.1241
## log(Gestation)  -0.6992     0.6362  -1.099   0.2790
## Danger          -0.8723     0.3249  -2.685   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.629 on 36 degrees of freedom
## Multiple R-squared:  0.5878, Adjusted R-squared:  0.5305
## F-statistic: 10.27 on 5 and 36 DF,  p-value: 3.573e-06
```

1. What are the model parameters and what are their estimates?

Answer: This model is Linear-Log Model. Residual standard error is 2.629 on 36 degrees of freedom (n-p-1), multiple R-squared is 0.5878, which means 58.78% variability can be explained.F-statistic is 10.27 on 5 and 36 degrees of freedom (n-p-1 = 42-5-1).

The estimate b0 (intercept) is 13.0446, b1 coefficient for variable log(BodyWt) is -0.4294, estimate of variable log(BrainWt) is -0.4437, estimate of variable log(LifeSpan) is 1.0811, estimate of variable log(Gestation) is -0.6992, estimate of variable Danger is -0.8723.

2. What is the equation for the regression line?

Answer: Equation of this regression line is

NonDreaming = 13.0446-0.4294_log(BodyWt)-0.4437_log(BrainWt)+1.0811_log(LifeSpan)-0.6992_log(Gestation)-0.8723*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The predicting variable is Danger (p-value = 0.0109).

4. Interpret the estimated value of the parameters, including the error term, corresponding to log(BodyWt) and Danger in the context of the problem

Answer: In this Linear-Log model, when other variables are hold in the fixed model, if BodyWt increases by 1%, the NonDreaming will decrease by 0.004294 unit. the std.Error of estimate is 0.4663 and p-value is 0.3633. In addition, when other variables are hold in the fixed model, if Danger increases by 1 unit, the NonDreaming will decrease by 0.8723 units. the std.Error of estimate is 0.3249 and p-value is 0.0109.

5. Did model3c improve over model3b? Explain how you determined if the model improved or not.

Answer: No, model3c seems to have a little weaker power comparing to model 3b according to model parameters such as R-squared, F-statistic value, p-value and other factors. For example, R-squared of model3b 0.6253, R-squared of model3c is 0.5878.

3d. For our final model, let’s attempt to improve the data assumptions and model predictability by adding back the transformation of the response variable, NonDreaming, using model3c as your baseline.

```
finalmodel <- lm(log(NonDreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log
(Gestation) + Danger, data = data)
summary(finalmodel)
```

```
##
## Call:
## lm(formula = log(NonDreaming) ~ log(BodyWt) + log(BrainWt) +
##     log(LifeSpan) + log(Gestation) + Danger, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69689 -0.25990  0.02811  0.20292  0.57709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.78927    0.32012   8.713 2.15e-10 ***
## log(BodyWt)    -0.11203    0.05881  -1.905  0.06478 .
## log(BrainWt)    0.03583    0.08838   0.405  0.68756
## log(LifeSpan)   0.07153    0.08660   0.826  0.41425
## log(Gestation) -0.13977    0.08023  -1.742  0.09003 .
## Danger         -0.11576    0.04097  -2.825  0.00766 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3316 on 36 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.5934
## F-statistic: 12.97 on 5 and 36 DF,  p-value: 3.036e-07
```

1. What are the model parameters and what are their estimates?

Answer: This model is Log-Log Model. Residual standard error is 0.3316 on 36 degrees of freedom (n-p-1), multiple R-squared is 0.643, which means 64.3% variability can be explained. F-statistic is 12.97 on 5 and 36 degrees of freedom (n-p-1 = 42-5-1), p-value of F is 3.036e-07.

The estimate b0 (intercept) is 2.78927, b1 coefficient for variable log(BodyWt) is -0.11203, estimate of variable log(BrainWt) is 0.03583, estimate of variable log(LifeSpan) is 0.07153, estimate of variable log(Gestation) is -0.13977, estimate of variable Danger is -0.11576.

2. What is the equation for the regression line?

Answer: The equation for the regression line is

log(NonDreaming) = 2.78927-0.11203*log(BodyWt)+0.03583*log(BrainWt)+0.07153*log(LifeSpan)-0.13977*log(Gestation)-0.11576*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The predicing variable of Danger (p-value = 0.00766).

4. Interpret the estimated value of the parameters, including the error term, corresponding to log(BodyWt) and Danger in the context of the problem.

Answer: In this Log-Log model, when other variables are hold in the fixed model, if BodyWt increases by 1%, the NonDreaming will decrease by 0.11203 %. the std.Error of estimate is 0.05881 and p-value is 0.06478. In addition, when other variables are hold in the fixed model, if Danger increases by 1 unit, the NonDreaming will decrease by 11.576%. the std.Error of estimate is 0.04097 and p-value is 0.00766.

5. Did finalmodel improve over model3c? Explain how you determined if the model improved or not.

Answer: Yes, finalmodel has a better improvement power comparing to model3c according to model parameters such as R-squared, F-statistic value, p-value and other factors, for example, R-squared of finalmodel is 0.643, R-squared of model3c is 0.5878.

Original Question 4 - Repeat Questions 3, 3a, 3b, 3c and 3d for the response variable Dreaming. Label your answers 4, 4a, 4b, 4c, 4d.

2-13-2019 Updated Question 4 - Repeat Questions 3, 3a, 3b, 3c and 3d with the response variable Dreaming. Label your answers 4, 4a, 4b, 4c, 4d. Important! Because row 11 (for species Echidna) has a zero value for Dreaming, lets remove that row of data prior to running Question 4, and rename our data variable data2.

```
data2 = data[-11, ]
head(data2)
```

```
##                   Species   BodyWt BrainWt NonDreaming Dreaming TotalSleep
## 1 Africangiantpouchedrat    1.000     6.6         6.3      2.0        8.3
## 2          Asianelephant 2547.000  4603.0         2.1      1.8        3.9
## 3                 Baboon   10.550   179.5         9.1      0.7        9.8
## 4            Bigbrownbat    0.023     0.3        15.8      3.9       19.7
## 5         Braziliantapir  160.000   169.0         5.2      1.0        6.2
## 6                    Cat    3.300    25.6        10.9      3.6       14.5
##   LifeSpan Gestation Predation Exposure Danger
## 1      4.5        42         3        1      3
## 2     69.0       624         3        5      4
## 3     27.0       180         4        4      4
## 4     19.0        35         1        1      1
## 5     30.4       392         4        5      4
## 6     28.0        63         1        2      1
```

Question 4: Fitting the Linear Regression Model.

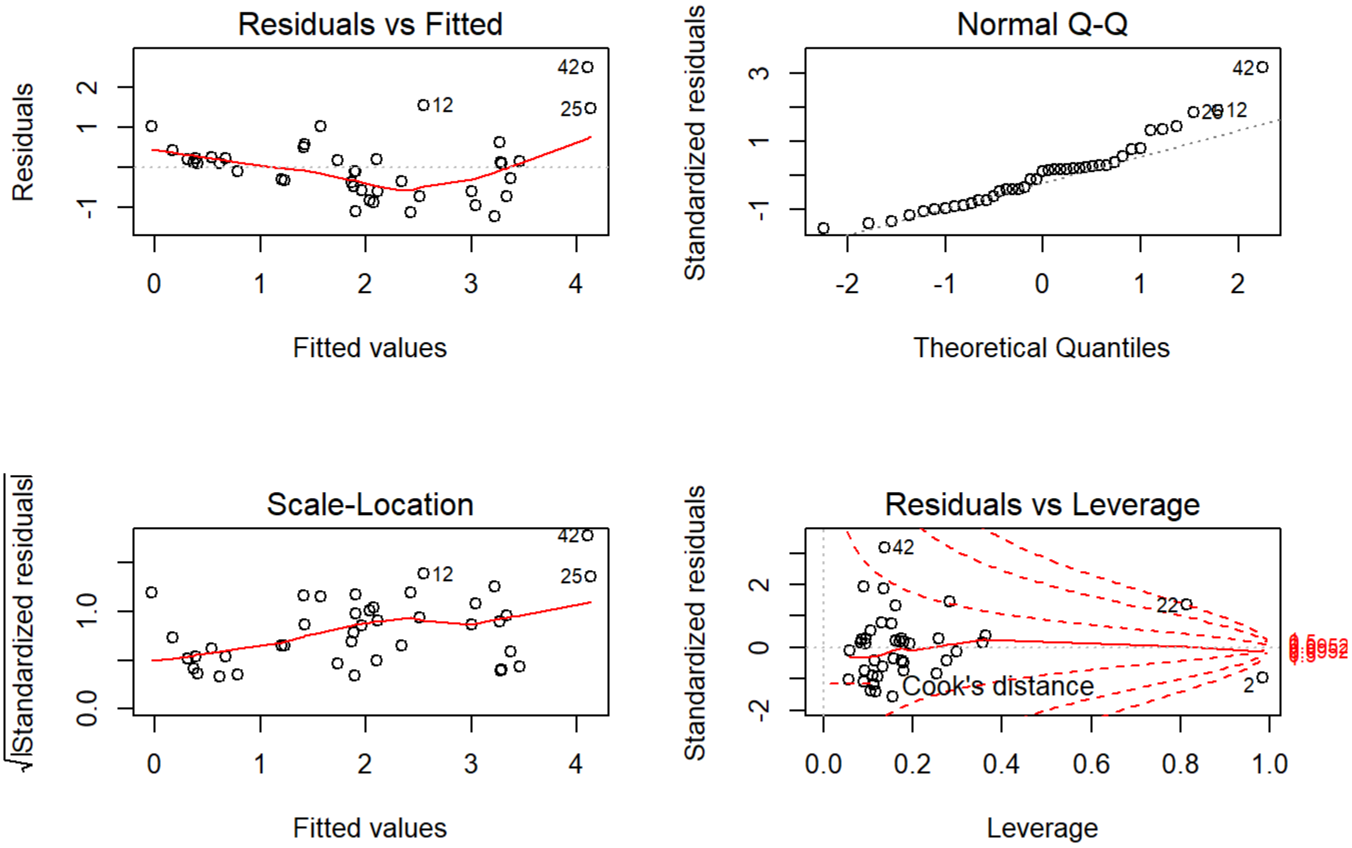
Plot the full model for Dreaming without transforming the response variable or predicting variables. Remember to exclude the two response variables for sleep and the Species column.

```
model4 <- lm(Dreaming ~. -Species -NonDreaming -TotalSleep, data = data2)
summary(model4)
```

```
##
## Call:
## lm(formula = Dreaming ~ . - Species - NonDreaming - TotalSleep,
##     data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22331 -0.56344  0.08977  0.23318  2.49739
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.820083   0.372492  10.255 8.57e-12 ***
## BodyWt       0.003615   0.001813   1.994  0.05444 .
## BrainWt     -0.001041   0.001086  -0.958  0.34505
## LifeSpan     0.011875   0.015420   0.770  0.44674
## Gestation   -0.007219   0.002081  -3.470  0.00147 **
## Predation    0.859964   0.304644   2.823  0.00800 **
## Exposure     0.295094   0.191910   1.538  0.13367
## Danger      -1.675645   0.378480  -4.427 9.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8463 on 33 degrees of freedom
## Multiple R-squared:  0.6865, Adjusted R-squared:  0.6199
## F-statistic: 10.32 on 7 and 33 DF,  p-value: 8.706e-07
```

```
par(mfrow = c(2,2))
plot(model4, cook.levels = c(4/42,0.5,1))
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



#### 1. What are the model parameters and what are their estimates?

Answer: This is multiple linear model. In the model, Residual standard error is 0.8463 on 33 degrees of freedom, multiple R-squared is 0.6865, which means 68.65% variability can be explained.F-statistic is 10.32 on 7 and 33 degrees of freedom (n-p-1 = 41-7-1).

The estimate b0(intercept) is 3.820083, b1 coefficient for variable BodyWt is 0.003615, estimate of variable BrainWt is -0.001041, estimate of variable LifeSpan is 0.011875, estimate of variable Gestation is -0.007219, estimate of variable Predation is 0.859964, estimate of variable Exposure is 0.295094, estimate of variable Danger is -1.675645.

2. What is the equation for the regression line?

Answer: Equation of regression line is

Dreaming = 3.820083 + 0.003615*BodyWt -0.001041*BrainWt +0.011875*LifeSpan-0.007219*Gestation+0.859964*Predation+0.295094*Exposure-1.675645*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The variables of Gestation (p-value = 0.00147), Predation (p-value = 0.00147) and Danger (p-value = 9.86e-05).

4. Interpret the estimated value of the parameters, including the error term, corresponding to BodyWt and Predation in the context of the problem.

Answer: The estimated intercept is 3.820083, which is an estimate the expected value of response variable (Dreaming) when all the predictor variables equal zero. In addition, when other variables are hold in the fixed model, if BodyWt increases by 1 unit, the Dreaming will increase by 0.003615 units, the standard error of estimate is 0.001813, p-value is 0.05444. When other variables are hold in the fixed model, if Predation increased by 1 unit, the Dreaming will increase by 0.859964 units, the standard error of estimate is 0.304644, p-value is 0.00800.

5. Check the assumptions of the model through plotting. Note potential outliers, if any.

Answer: (1). To evaluate constant variance and independence assumption, we can use scatter plot of Residuals vs Fitted, we found that many points of residuals are randomly scatted around the 0 line, but some residuals are still non-randomly scattered, which indicates that we may have constant variance to some degree. However, we found several possible outlier points, they are points of No.12, 25 and 42.

(2). To evaluate the normality we can use Normal QQ plot, we found that most of residuals distrubute along a straight line except two ends, the distrubtion of residuals have some deviation from the straight line at two tails especially the top tail (right skewed), there appears to have outlier points of No.12, 25 and 42.

(3) For plots of Scale_Location, sqrt of standard residuals vs fitted values can be evaluate the costant variance or independence, we found severl potential outlier such as points of No.12, 25, 42. In addition, cook's distance are useful to identify these outliers. From plot of Residuals vs leverage, we found severl potential outlier such as points of No.2, 22 and 42.

4a. Change model4 to log transform the response variable, Dreaming.

```
model4a <- lm(log(Dreaming) ~. -Species -NonDreaming -TotalSleep, data = data2)
summary(model4a)
```

```
##
## Call:
## lm(formula = log(Dreaming) ~ . - Species - NonDreaming - TotalSleep,
##     data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56473 -0.23549 -0.03553  0.17015  0.72288
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4684458  0.1442539  10.180 1.03e-11 ***
## BodyWt       0.0020701  0.0007020   2.949  0.00582 **
## BrainWt     -0.0005974  0.0004208  -1.420  0.16503
## LifeSpan     0.0104864  0.0059717   1.756  0.08837 .
## Gestation   -0.0042079  0.0008057  -5.223 9.57e-06 ***
## Predation    0.3747033  0.1179789   3.176  0.00323 **
## Exposure     0.1015576  0.0743207   1.366  0.18103
## Danger      -0.7743707  0.1465730  -5.283 8.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3277 on 33 degrees of freedom
## Multiple R-squared:  0.8066, Adjusted R-squared:  0.7656
## F-statistic: 19.67 on 7 and 33 DF,  p-value: 4.342e-10
```

1. What are the model parameters and what are their estimates?

Answer: This model is Log-Linear Model. Residual standard error is 0.3277 on 33 degrees of freedom (n-p-1), multiple R-squared is 0.8066, which means 80.66% variability can be explained.F-statistic is 19.67 on 7 and 33 degrees of freedom (n-p-1 = 41-7-1).

The estimate b0(intercept) is 1.4684458, b1 coefficient for variable BodyWt is 0.0020701, estimate of variable BrainWt is -0.0005974, estimate of variable LifeSpan is 0.0104864, estimate of variable Gestation is -0.0042079, estimate of variable Predation is 0.3747033, estimate of variable Exposure is 0.1015576, estimate of variable Danger is -0.7743707.

2. What is the equation for the regression line?

Answer: log(Dreaming) = 1.4684458+0.0020701*BodyWt-0.0005974*BrainWt +0.0104864*LifeSpan-0.0042079*Gestation+0.3747033*Predation+0.1015576*Exposure-0.7743707*Danger

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: predicting variables of BodyWt (p-value = 0.00582), Gestation (p-value=9.57e-06), Predation (p-value = 0.00323) and Danger (p-value = 8.00e-06).

4. Interpret the estimated value of the parameters, including the error term,

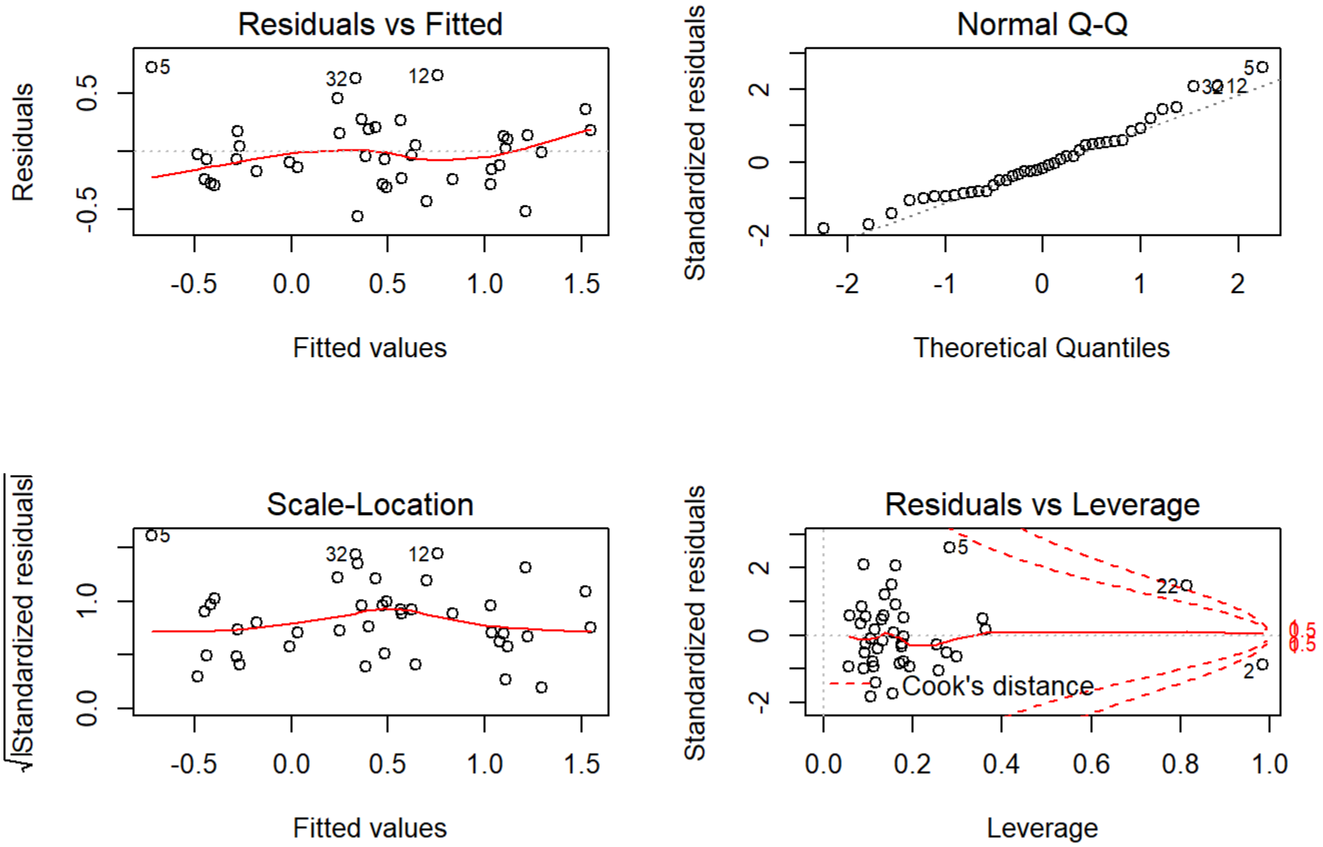
corresponding to BodyWt and Predation in the context of the problem.

Answer: In this Log-Linear model, when other variables are hold in the fixed model, if BodyWt increases by 1 unit, the Dreaming will increase by 0.20701 %. the std.Error of estimate is 0.0007020 an p-value is 0.00582. In addition, when other variables are hold in the fixed model, if Predation increases by 1 unit, the Dreaming will increase by 37.47033 %. the std.Error of estimate is 0.1179789 an p-value is 0.00323.

5. Check the assumptions of the model through plotting. Note potential outliers, if any.

```
par(mfrow = c(2,2))
plot(model4a)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Answer: (1). To evaluate constant variance and independence assumption, we can use scatter plot of Residuals vs Fitted, we found that most residuals are randomly scatted around the 0 line, which indicates that we probably have many constant variance. In addition, we found several possible outlier points, which are points of No.5, 12 and 32.

(2). To evaluate the normality we can use Normal QQ plot, we found that many residuals distrubute along a straight line except two ends, the distrubtion of residuals have some deviation from the straight line at two tails (right skewed), there appears to have outlier points of No.5, 12 and 32.

(3) For plots of Scale_Location, sqrt of standard resuduals and fitted values can be evaluate the costant variance or independence, we found severl potential outlier such as points of No.5, 12 and 32, which is deviate from the horizon line. In addition, cook’s distance are useful to identify these outliers. From plot of Residuals vs leverage, we found severl potential outlier such as points of No.2, 5 and 22.

4b. Change model4a to remove the log transform of Dreaming, and add the log transformation of numeric response variables BrainWt, BodyWt, LifeSpan and Gestation.

```
model4b <- lm(Dreaming ~ log(BodyWt) +log(BrainWt) +log(LifeSpan) +log(Gestation) +Ex
posure +Predation +Danger, data = data2)
summary(model4b)
```

```
##
## Call:
## lm(formula = Dreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) +
##     log(Gestation) + Exposure + Predation + Danger, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44161 -0.26871 -0.09673  0.34675  1.45088
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.43127    0.80648   9.214 1.21e-10 ***
## log(BodyWt)     0.44017    0.13251   3.322 0.002195 **
## log(BrainWt)   -0.35662    0.19429  -1.836 0.075449 .
## log(LifeSpan)   0.02462    0.21895   0.112 0.911153
## log(Gestation) -0.82406    0.19193  -4.294 0.000145 ***
## Exposure        0.26488    0.17167   1.543 0.132367
## Predation       0.59634    0.27182   2.194 0.035392 *
## Danger         -1.36005    0.32232  -4.220 0.000180 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.714 on 33 degrees of freedom
## Multiple R-squared:  0.7768, Adjusted R-squared:  0.7295
## F-statistic: 16.41 on 7 and 33 DF,  p-value: 4.24e-09
```

1. What are the model parameters and what are their estimates?

Answer: This model is Linear-Log model. In this model, Residual standard error of model is 0.714 on 33 degrees of freedom (n-p-1), multiple R-squared is 0.7768, which means 77.68% variability can be explained.F-statistic is 16.41 on 7 and 33 degrees of freedom (n-p-1 = 41-7-1).

The estimate b0(intercept) is 7.43127, b1 coefficient for variable log(BodyWt) is 0.44017, estimate of variable log(BrainWt) is -0.35662, estimate of variable log(LifeSpan) is 0.02462, estimate of variable log(Gestation) is -0.82406, estimate of variable Exposure is 0.26488, estimate of variable Predation is 0.59634, estimate of variable Danger is 0.59634.

2. What is the equation for the regression line?

Answer: Equation of this model:

Dreaming = 7.43127+0.44017*log(BodyWt)-0.35662*log(BrainWt)+0.02462*log(LifeSpan)-0.82406*log(Gestation)+0.26488*Exposure+0.59634*Predation-1.36005*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The predictor variable of BodyWt (p-value = 0.002195), Gestation (p-value = 0.000145), Predation (p-value = 0.035392) and Danger (p-value = 0.000180).

4. Interpret the estimated value of the parameters, including the error term, corresponding to log(BodyWt) and Predation in the context of the problem.

Answer: In this Linear-Log model, when other variables are hold in the fixed model, if BodyWt increases by 1%, the Dreaming will increase by 0.0044017 units. the std.Error of estimate is 0.13251 an p-value is 0.002195. In addition, when other variables are hold in the fixed model, if Predation increases by 1 unit, the Dreaming will increase by 0.59634 units. the std.Error of estimate is 0.27182 and p-value is 0.035392.

5. Did model4b improve over model4a? Explain how you determined if the model improved or not.

Answer: No. Both of model4a and model4b are very good model, and model4a is better than model4b according to model parameters such as R-squared, F-statistic value, p-value and other factors, for example, R-squared of model4a is 0.8066, R-squared of model4b ia 0.7768.

4c. Because the Danger variable is an interpolation of the Exposure and Predation variables, let’s keep Danger and remove the other two from the model using model4b as your baseline.

```
model4c <- lm(Dreaming ~ log(BodyWt) +log(BrainWt) +log(LifeSpan) +log(Gestation) +Da
nger, data = data2)
summary(model4c)
```

```
##
## Call:
## lm(formula = Dreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) +
##     log(Gestation) + Danger, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8985 -0.5195 -0.1136  0.3598  1.5668
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.94819    0.74532  10.664 1.53e-12 ***
## log(BodyWt)     0.43897    0.13553   3.239  0.00263 **
## log(BrainWt)   -0.32370    0.20370  -1.589  0.12103
## log(LifeSpan)  -0.02089    0.21328  -0.098  0.92252
## log(Gestation) -0.88941    0.19472  -4.568 5.88e-05 ***
## Danger         -0.55009    0.09443  -5.825 1.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7642 on 35 degrees of freedom
## Multiple R-squared:  0.7288, Adjusted R-squared:  0.6901
## F-statistic: 18.82 on 5 and 35 DF,  p-value: 4.721e-09
```

1. What are the model parameters and what are their estimates?

Answer: This model is Linear-Log Model. In this model, Residual standard error is 0.7642 on 35 degrees of freedom (n-p-1), multiple R-squared is 0.7288, which means 72.88% variability can be explained.F-statistic is 18.82 on 5 and 35 degrees of freedom (n-p-1 = 41-5-1).

The estimate b0(intercept) is 7.94819, b1 coefficient for variable log(BodyWt) is 0.43897, estimate of variable log(BrainWt) is -0.32370, estimate of variable log(LifeSpan) is -0.02089, estimate of variable log(Gestation) is -0.88941, estimate of variable Danger is -0.55009.

2. What is the equation for the regression line?

Answer: Equation of this regression line is

Dreaming = 7.94819+0.43897*log(BodyWt)-0.32370*log(BrainWt)-0.02089*log(LifeSpan)-0.88941*log(Gestation)-0.55009*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The predicting variable is BodyWt (p-value = 0.00263), Gestation (p-vlaue = 5.88e-05) and Danger (p-value = 1.31e-06).

4. Interpret the estimated value of the parameters, including the error term, corresponding to log(BodyWt) and Danger in the context of the problem

Answer: In this Linear-Log model, when other variables are hold in the fixed model, if

BodyWt increases by 1%, the Dreaming will increase by 0.0043897 units. the std.Error of estimate is 0.13553 and p-value is 0.00263. In addition, when other variables are hold in the fixed model, if Danger increases by 1 unit, the Dreaming will decrease by 0.55009 units. the std.Error of estimate is 0.09443 and p-value is 1.31e-06.

5. Did model4c improve over model4b? Explain how you determined if the model improved or not.

Answer: No. Both model 4c and model3b model are very good, the power of model4b is a little better than mdoel4c according to model parameters such as R-squared, F-statistic value, p-value and other factors, for example R-squared of model4b is 0.7768, R-squared of model4c is 0.7288.

4d. For our final model, let’s attempt to improve the data assumptions and model predictability by adding back the transformation of the response variable, Dreaming, using model3c as your baseline.

```
finalmodel_4d <- lm(log(Dreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log
(Gestation) + Danger, data = data2)
summary(finalmodel_4d)
```

```
##
## Call:
## lm(formula = log(Dreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) +
##     log(Gestation) + Danger, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70627 -0.20824  0.03513  0.17167  0.87600
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.10346    0.37019   8.383 6.90e-10 ***
## log(BodyWt)     0.14340    0.06732   2.130 0.040252 *
## log(BrainWt)   -0.11064    0.10117  -1.094 0.281612
## log(LifeSpan)   0.05317    0.10593   0.502 0.618862
## log(Gestation) -0.41163    0.09671  -4.256 0.000148 ***
## Danger         -0.28973    0.04690  -6.177 4.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3796 on 35 degrees of freedom
## Multiple R-squared:  0.7249, Adjusted R-squared:  0.6856
## F-statistic: 18.45 on 5 and 35 DF,  p-value: 6.022e-09
```

1. What are the model parameters and what are their estimates?

Answer: This model is Log-Log Model. In this model, Residual standard error is 0.3796 on 35 degrees of freedom (n-p-1), multiple R-squared is 0.7249, which means 72.49% variability can be explained. F-statistic is 18.45 on 5 and 35 degrees of freedom (n-p-1 = 41-5-1), p-value of F is 6.022e-09.

The estimate b0(intercept) is 3.10346, b1 coefficient for variable log(BodyWt) is 0.14340, estimate of variable log(BrainWt) is -0.11064, estimate of variable log(LifeSpan) is 0.05317, estimate of variable log(Gestation) is -0.41163, estimate of variable Danger is -0.28973.

2. What is the equation for the regression line?

Answer: The equation for the regression line is

log(Dreaming) = 3.10346+0.14340*log(BodyWt)-0.11064*log(BrainWt)+0.05317*log(LifeSpan)-0.41163*log(Gestation)-0.28973*Danger.

3. Which predicting variable(s) are significant at alpha = 0.05? What are their p-values?

Answer: The predicing variable of BodyWt (p-value = 0.040252), Gestation (p-value = 0.000148) and Danger (p-value = 4.51e-07).

4. Interpret the estimated value of the parameters, including the error term, corresponding to log(BodyWt) and Danger in the context of the problem.

Answer: In this Log-Log model, when other variables are hold in the fixed model, if BodyWt increases by 1%, the Dreaming will increase by 0.14340%. the std.Error of estimate is 0.06732 and p-value is 0.040252. In addition, when other variables are hold in the fixed model, if Danger increases by 1 unit, the Dreaming will decrease by 28.973%. the std.Error of estimate is 0.04690 and p-value is 4.51e-07.

5. Did finalmodel improve over model4c? Explain how you determined if the model improved or not.

Answer: Both model4d and model4c are very good model, the difference is very small between them according to model parameters such as R-squared, F-statistic value, p-value and other factors, for example, the R-squared of model4c is 0.7288, R-squared of finalmodel_4d is 0.7249.

Question 5 - Checking the Assumptions of the Model

5. Plot the relevant residual plots to check the final model assumptions. (Linearity does not need to be plotted or analyzed as this was addressed in the first section of the assignment). You should have 3 plots. Enumerate the assumptions and describe what graphical techniques you used. Interpret the displays with respect to the assumptions of the linear regression model. Be sure to include the analysis of outliers. Comment on any apparent departures from the assumptions of the linear regression model.

```
finalmodel <- lm(log(NonDreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log
(Gestation) + Danger, data = data)
#summary(finalmodel)
#plot(finalmodel, cook.levels = c(4/42,0.5,1))
library(car)
```

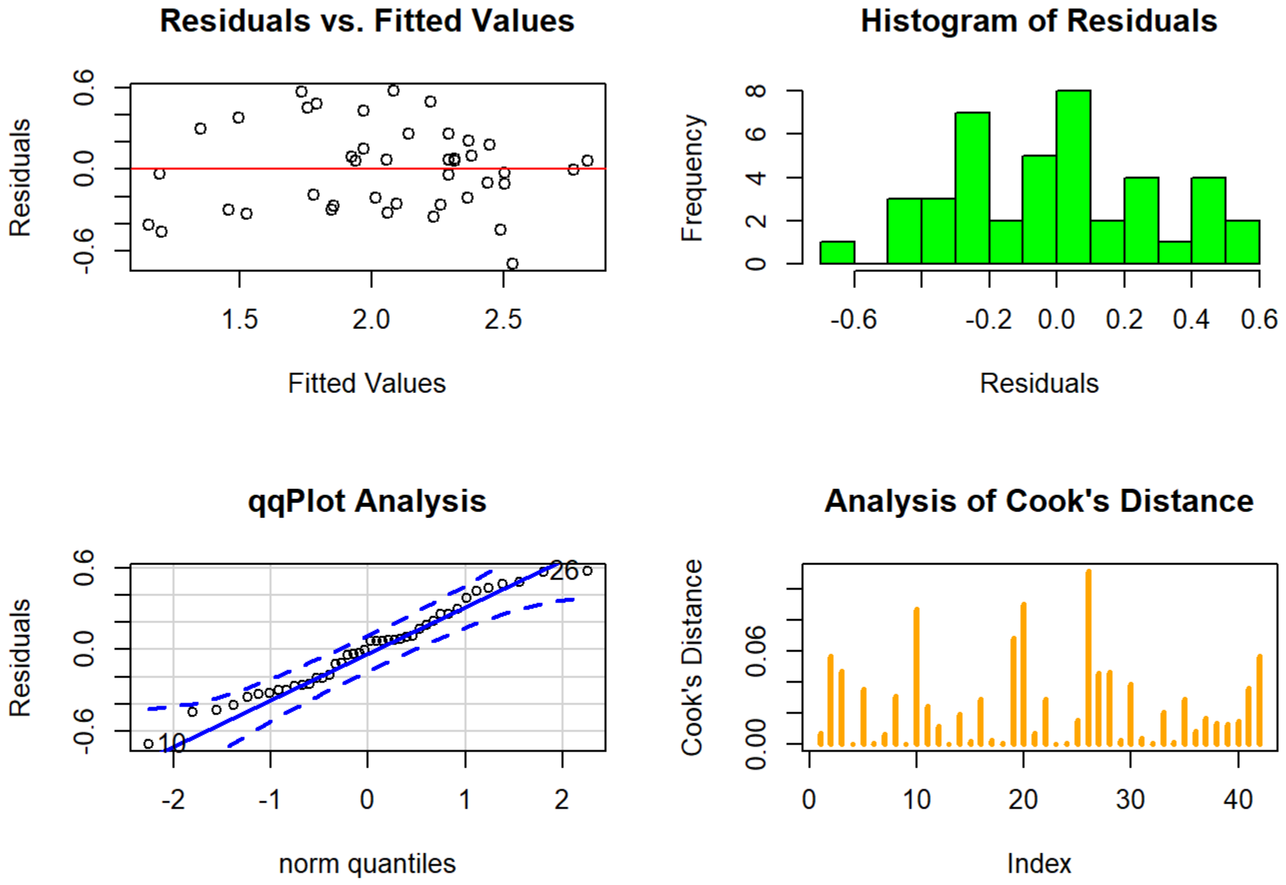
```
## Warning: package 'car' was built under R version 3.5.2
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.5.2
```

```
par(mfrow=c(2,2))
{plot(finalmodel$fitted, finalmodel$residuals, xlab = 'Fitted Values', ylab = 'Residu
als',
      main = 'Residuals vs. Fitted Values')
abline(0,0, col='red')}

hist(finalmodel$residuals, xlab = 'Residuals', main = 'Histogram of Residuals', nclas
s = 10, col = 'green')

qqPlot(finalmodel$residuals, ylab = 'Residuals', main = "qqPlot Analysis")
```

```
## [1] 10 26
```

```
cook = cooks.distance(finalmodel)
plot(cook, type = 'h', lwd = 3, col = 'orange', ylab = "Cook's Distance", main = "Ana
lysis of Cook's Distance")
```



Answer: In overview, there are four assumptions for fitting model: Linearity, Constant

Variance, Independence and Normality.

(1). Firstly, Linearity has been addressed in the first section of the assignment as above. For Constant Variance, we evaluate it with plot of Residuals vs Fitted Values, we do not see any obvious clustering of the residuals in the plot, and the points of residuals are randomly scatted around zero line, which indicates constant variance and independence.

(2). We also evaluate distribution of residuals with Histogram of Residuals which is almost symmetric except there are two or three gaps in the plot.

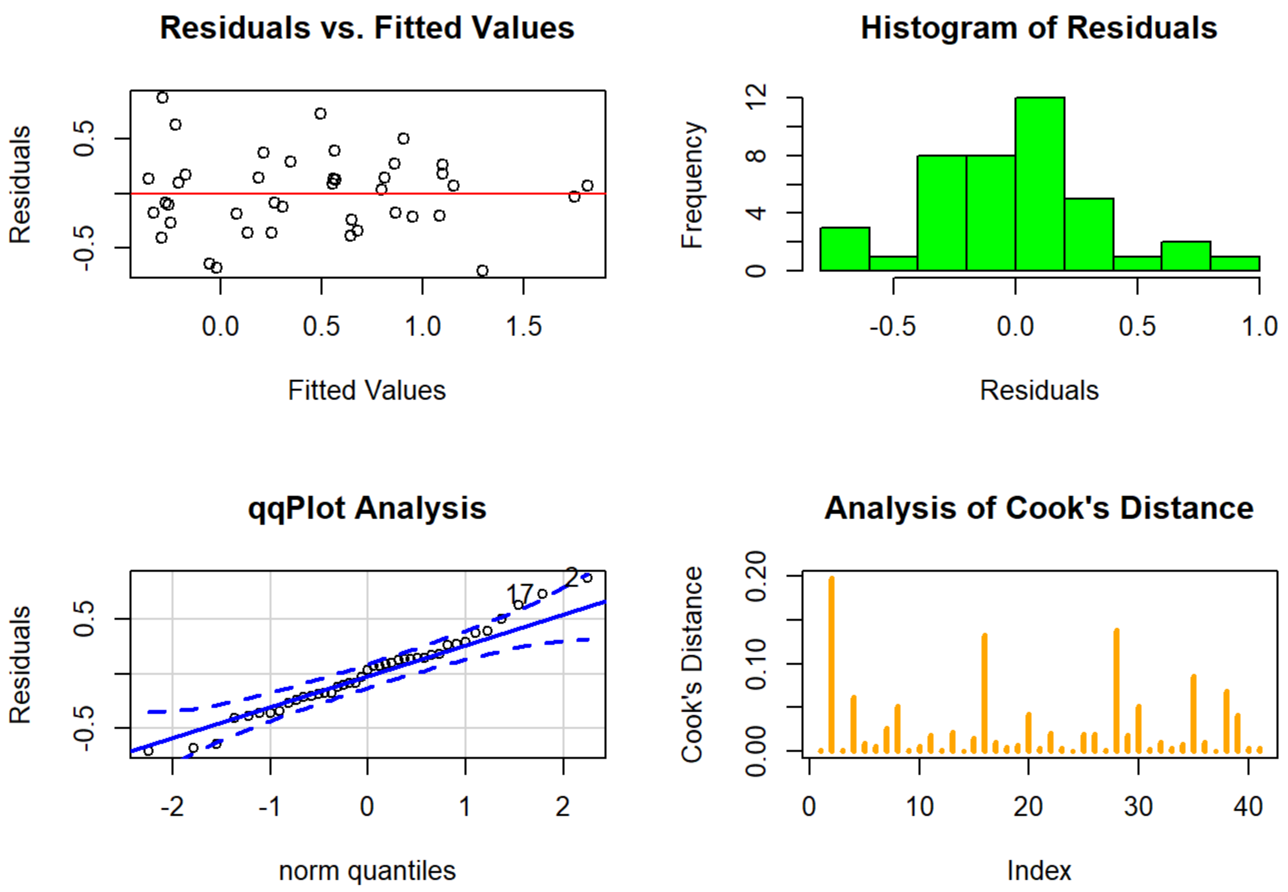
(3). In addition, we use qqPlot to evaluate normality of model, we found that most of residuals distrubute along a straight line, there are two potential outlier of points (No. 10 and 26).

(4). The last one is the plot of Cook’s distance, it is useful to identify these outliers or influential points. We see that a few vlaues are somewhat larger than the other values.

```
finalmodel_4d <- lm(log(Dreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log
(Gestation) + Danger, data = data2)
#summary(finalmodel_4d)
par(mfrow=c(2,2))
{plot(finalmodel_4d$fitted, finalmodel_4d$residuals, xlab = 'Fitted Values', ylab = '
Residuals',
      main = 'Residuals vs. Fitted Values')
abline(0,0, col='red')}

hist(finalmodel_4d$residuals, xlab = 'Residuals', main = 'Histogram of Residuals', nc
lass = 10, col = 'green')

qqPlot(finalmodel_4d$residuals, ylab = 'Residuals', main = "qqPlot Analysis")
```

```
##  2 17
##  2 16
```

```
cook = cooks.distance(finalmodel_4d)
plot(cook, type = 'h', lwd = 3, col = 'orange', ylab = "Cook's Distance", main = "Ana
lysis of Cook's Distance")
```



Answer: There are four assumptions for fitting model: Linearity, Constant Variance, Independence and Normality. Firstly, Linearity has been addressed in the first section of the assignment as above.

(1). The first plot is about Constant Variance, we evaluate it with plot of Residuals vs Fitted Values, we do not see any clustering of the residuals in the plot, and the points of residuals are randomly spread around zero line, which indicates both constant variance and the uncorrelated errors assumptions will hold.

(2). We also evaluate distribution of residuals with Histogram of Residuals, the residuals are almost symmetric distribution except a little right skewed, it also has no obvious gaps.

(3). In addition, we use qqPlot to evaluate normality of model, we found that many residuals distrubute along a straight line except two ends (a little right skewed on the right tail), there are two potential points of outlier: No. 2 and 17.

(4). The last plot is about Cook's distance, it is useful to identify these outlier or influential points, we see that three values is somewhat larger than the other values.