ISYE 6414 Homework 3 – Solutions to Peer Assessment

Spring 2019

Section A: No assignment due for this section.

Section B: Multiple Linear Regression - R Data Analysis

In this problem, the sleep patterns of mammals were collected and studied by Allison, T., and Cicchetti, D. V. in 1976. The study attempted to understand the interrelationships between sleep and survival from natural predators. The study's premise is that by correlating sleep habits and other ecological characteristics of species, it may be possible to clarify the importance of sleep in mammals.

The data set sleep.csv  Includes brain and body weight, life span, gestation time, time sleeping, and predation and danger indices for 62 species of mammals (data cleansing removed null values which reduced the number of species to 42). The data set variables are explained in the chart below. Also in the file is a column labeled "Species" which is the name of the species of mammal studied. Of interest in this assignment are the sleep time response variables NonDreaming, Dreaming and TotalSleep.

Variable	Description
BodyWt	body weight (kg)
BrainWt	brain weight (g)
NonDreaming	slow wave ("nondreaming") sleep (hrs/day)
Dreaming	paradoxical ("dreaming") sleep (hrs/day)
TotalSleep	total sleep, sum of slow wave and paradoxical sleep (hrs/day)
LifeSpan	maximum life span (years)
Gestation	gestation time (days)
Predation	predation index (1-5) 1 = minimum (least likely to be preyed upon); 5 = maximum (most likely to be preyed upon)
Exposure	sleep exposure index (1-5) 1 = least exposed (e.g. animal sleeps in a well-protected den); 5 = most exposed
Danger	overall danger index (1-5) (based on the above two indices and other information) 1 = least danger (from other animals); 5 = most danger (from other animals)

Source

Allison, T., and Cicchetti, D. V. (1976). Sleep in mammals: ecological and constitutional correlates. *Science* **194** (November 12), 732-734.

The electronic data file was obtained from the Statlib database. The data file has been modified to remove null values.

This assignment provides repetition and practice for data assumption analysis for continuous and categorical variables, multiple linear regression modeling, transformation, and interpretation.

A reference guide from a different class will help with your interpretations, and is attached here. Interpretations for this assignment fall into four categories.

Level-Level: Linear, no log transformations.

Linear-Log: predicting variable(s) are log transformed.

Log-Linear: response variable is log transformed.

Log-Log: predicting variable(s) plus the response variable are transformed.

Please refer to the document for a reference in responding to the questions regarding the interpretation of the model parameters.

Instructions for reading the data. The data file name is sleep.csv  . To read the data in R, save the file in your working directory (make sure you have changed the directory if different from the R working directory) and read the data using the R function below and run the variable set up code as follows:

R-Code

```
data = read.csv("sleep.csv", header=TRUE, sep = ', ')
head(data)

#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).

R-Code

pairs(data) #or plot(data)

1a. Solution

Because the observations seemingly form a line, it is difficult to assess the linear relationship (direction and form) between NonDreaming-BodyWt and NonDreaming-BrainWt without transforming these predictors. Recommendation as well to analyze potential outliers.

The relationship between NonDreaming-LifeSpan and NonDreaming-Gestation appear linear in form and negative in direction, and its recommended to analyze potential outliers.

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

R-Code

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

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

1b. Solution

Each continuous predicting variable is negatively correlated with the response, NonDreaming. Gestation has the strongest negative correlation at -0.61, while the remaining predictors range from -0.39 and -0.37 in strength.

- 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 response variable NonDreaming, to improve linearity, although not required.

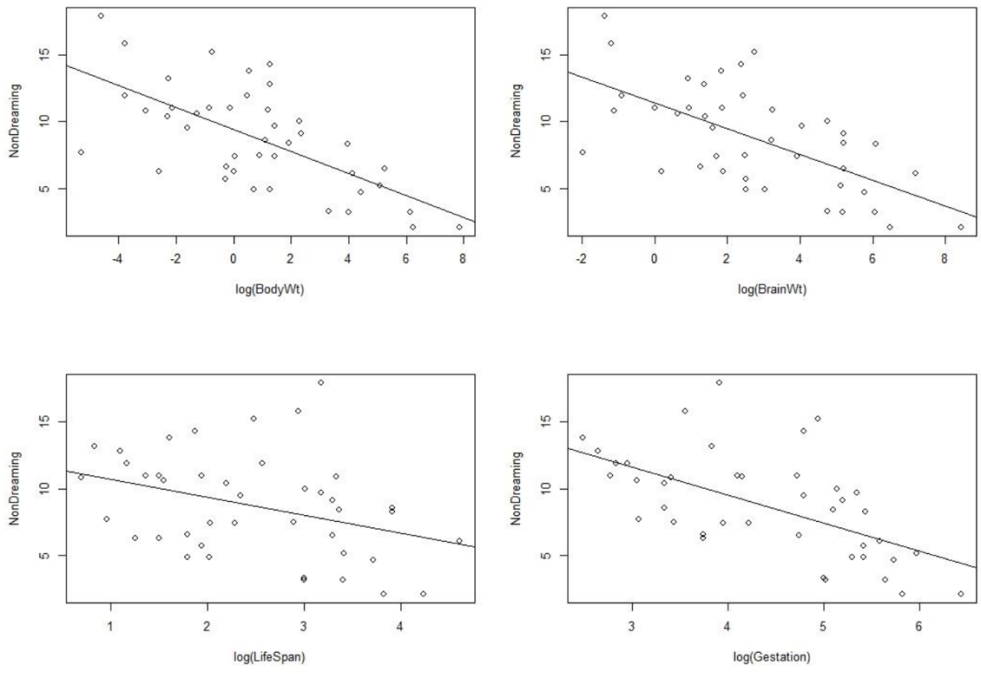
- Visually inspect each continuous predicting variable. Include final plots for each variable in your report. Here is starter code for the first variable, BodyWt.

R-Code –

```
#Inspect BodyWt
plot(BodyWt, NonDreaming)
abline(lm(NonDreaming ~ BodyWt, data = data))
```

```
# Transform BodyWt
plot(log(BodyWt), NonDreaming)
abline(lm(NonDreaming ~log(BodyWt), data = data))

# Inspect and Transform BrainWt
plot(BrainWt, NonDreaming)
abline(lm(NonDreaming ~ BrainWt, data = data))
plot(log(BrainWt), NonDreaming)
abline(lm(NonDreaming ~ log(BrainWt), data = data))

# Inspect and Transform LifeSpan
plot(LifeSpan, NonDreaming)
abline(lm(NonDreaming ~ LifeSpan, data = data))
plot(log(LifeSpan), NonDreaming)
abline(lm(NonDreaming ~ log(LifeSpan), data = data))

# Inspect and Transform Gestation
plot(Gestation, NonDreaming)
abline(lm(NonDreaming ~ Gestation, data = data))
plot(log(Gestation), NonDreaming)
abline(lm(NonDreaming ~ log(Gestation), data = data))
```

1c. Solution



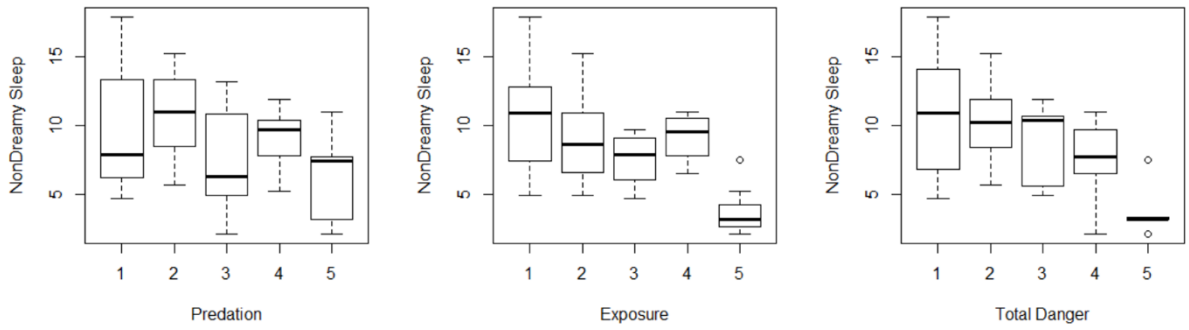
In each of the four plots, linear form and negative direction are clearly visible with the log transformation of all quantitative predicting variables.

Grading: Interpretation of visual plots can be subjective. Some students may conclude a linear relationship for LifeSpan and Gestation without transformation. Please give full credit to students who at a minimum, concluded that the log transform of both BodyWt and BrainWt improved the linearity of the relationship with the response variable NonDreaming.

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

R-Code
boxplot(NonDreaming ~ Predation, xlab = "Predation", ylab = "NonDreamy Sleep")
boxplot(NonDreaming ~ Exposure, xlab = "Exposure", ylab = "NonDreamy Sleep")
boxplot(NonDreaming ~ Danger, xlab = "Total Danger", ylab = "NonDreamy Sleep")

1d. Solution



NonDreaming appears to vary with each of the categorical predictors Predation, Exposure and Danger.

We see some degree of overlap in the box plots above. Generally, for each of the categorical variables, NonDreamy sleep is higher when survival risk is lower. For Predation, Exposure and Danger, we observe variability within the group and variability between the means of each group. Therefore, it is likely that the qualitative variables for these groups are significant.

- 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?

1e. Solution

Based on the exploratory analysis, it is reasonable to assume a linear regression model as the independent quantitative variables show a linear relationship with the response variable NonDreamy, the correlation coefficients confirm direction and strength of relationships with the response, and the independent qualitative variables may be important in explaining and predicting NonDreamy sleep. We should keep all predicting variables in the model to start with.

Question 2: Exploratory Data Analysis - Dreaming

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

R-Code
pairs(data) #or plot(data)

2a. Solution

Because the observations seemingly form a line, it is difficult to assess the linear relationship (direction and form) between Dreaming-BodyWt and Dreaming-BrainWt without transforming these predictors. Recommendation as well to analyze potential outliers.

The relationship between Dreaming-LifeSpan and Dreaming-Gestation appear linear in form and negative in direction, with further analysis on outliers recommended.

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

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

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

2b. Solution

Each continuous predicting variable is negatively correlated with the response, Dreaming. Gestation has the strongest negative correlation at -0.41, LifeSpan is at -0.27, and BodyWt and BrainWt with low correlation at 0.07.

- 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 response variable Dreaming, to improve linearity, although not required.

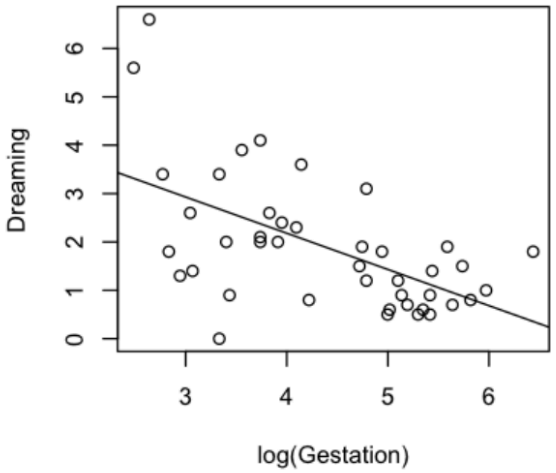
- Visually inspect each continuous predicting variable. Include final plots for each variable in your report. Here is starter code for the first variable, BodyWt.

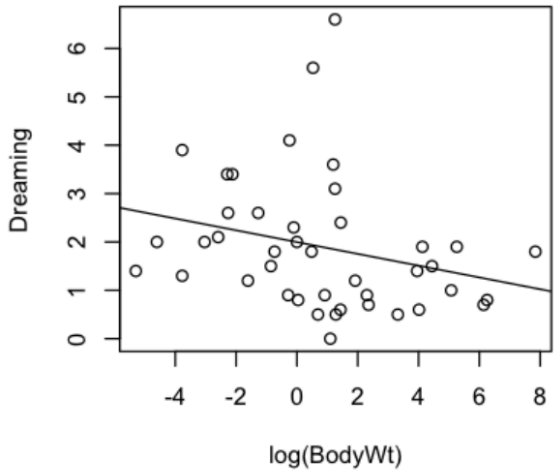
R-Code –

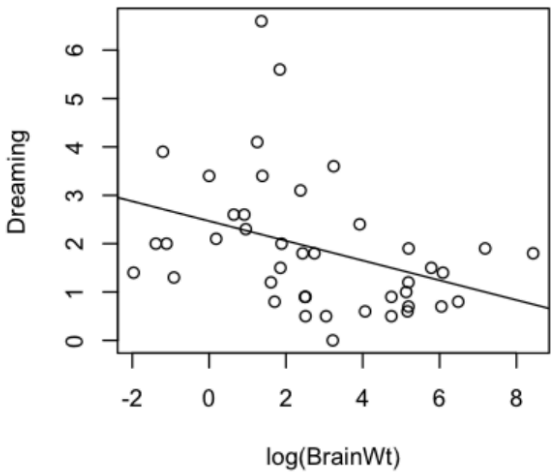
```
#Inspect BodyWt
plot(BodyWt, Dreaming)
abline(lm(Dreaming ~ BodyWt, data = data))
```

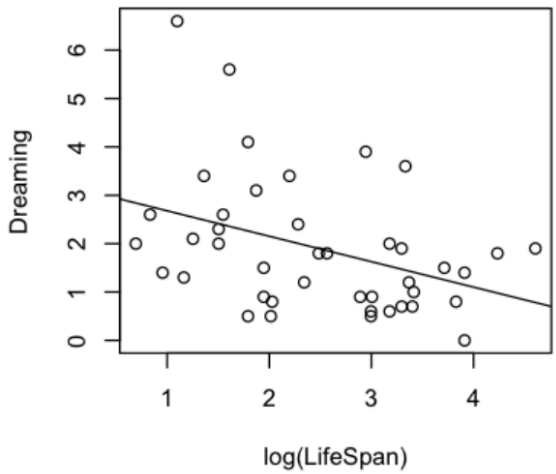
```
# Transform BodyWt
plot(log(BodyWt), Dreaming)
abline(lm(Dreaming ~log(BodyWt), data = data))
```

2c. Solution









In each of the four plots, linear form and negative direction are clearly visible with the log transformation of all quantitative predicting variables.

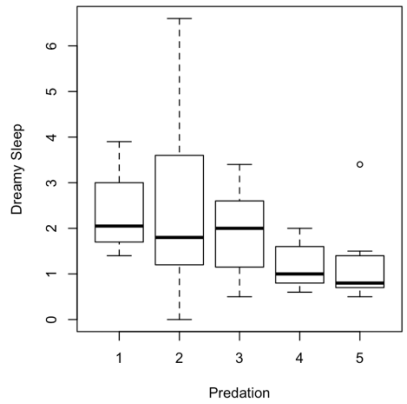
Grading: Interpretation of visual plots can be subjective. Some students may conclude a linear relationship for LifeSpan and Gestation without transformation. Please give full credit to students who at a minimum, concluded that the log transform of both BodyWt and BrainWt improved the linearity of the relationship with the response variable Dreaming.

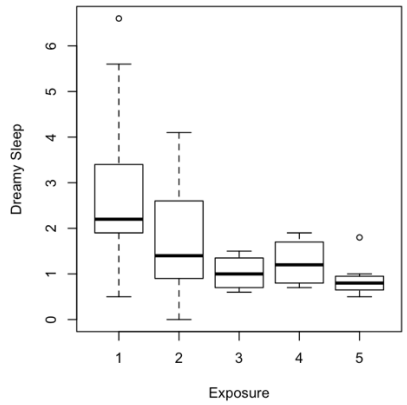
- 2d. Using boxplots, describe the relationship between Dreaming and the categorical independent variables Predation, Exposure, and Danger. Does Dreaming vary with the categorical variables?

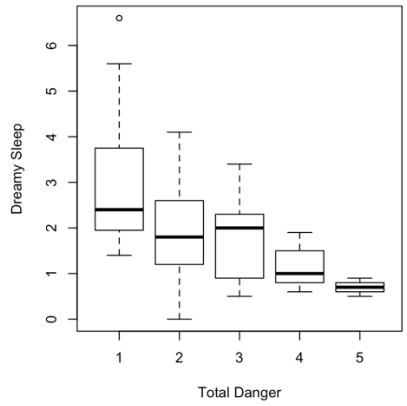
R-Code

```
boxplot(Dreaming ~ Predation, xlab = "Predation", ylab = "Dreamy Sleep")
boxplot(Dreaming ~ Exposure, xlab = "Exposure", ylab = "Dreamy Sleep")
boxplot(Dreaming ~ Danger, xlab = "Total Danger", ylab = "Dreamy Sleep")
```

2d. Solution







Dreaming appears to vary with each qualitative predictor Predation, Exposure and Danger.

We see some degree of overlap in the box plots above. Generally, for each of the categorical variables, Dreamy sleep is higher when survival risk is lower. For Predation, Exposure and Danger, we observe variability within the group and variability between the means of each group. Therefore, it is likely that the qualitative variables for these groups are significant.

- 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?

2e. Solution

Based on the exploratory analysis, it is reasonable to assume a linear regression model as the independent quantitative variables have a linear relationship with the response variable Dreamy, the correlation coefficients indicate some level of negative relationship with the response, and the categorical variables may be important in explaining and predicting Dreamy sleep. We should keep all predicting variables in the model to start with.

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.

R Code

```
model3 <- lm(NonDreaming ~. -Species -Dreaming -TotalSleep, data = data)
summary(model3)
plot(model3, cook.levels = c(4/42,0.5,1))
```

3 Solution

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 7 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 13.331488$
- $\beta_{BodyWt} = 0.003332$
- $\beta_{BrainWt} = -0.001294$
- $\beta_{LifeSpan} = -0.001181$
- $\beta_{Gestation} = -0.013804$
- $\beta_{Predation} = 1.414774$
- $\beta_{Exposure} = 0.224418$
- $\beta_{Danger} = -2.799115$
- $\sigma^2 = 8.162449$

Grading: If the RSE-squared (sigma^2) was not included in the list of parameters with its estimated value, this question should be graded incorrect for one of the five models. Please don’t mark more than one incorrect to avoid doubling up on penalties.

2. What is the equation for the regression line?

$$\hat{NonDreaming} = 13.331 + 0.003BodyWt - 0.001BrainWt - 0.001LifeSpan - 0.014Gestation$$
$$+1.415Predation + 0.224Exposure - 2.799Danger$$

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

Gestation, p-val = 0.0429
Danger, p-val = 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.

Interpret BodyWt: This problem uses the Level-Level interpretation Y = b0 + b1*X

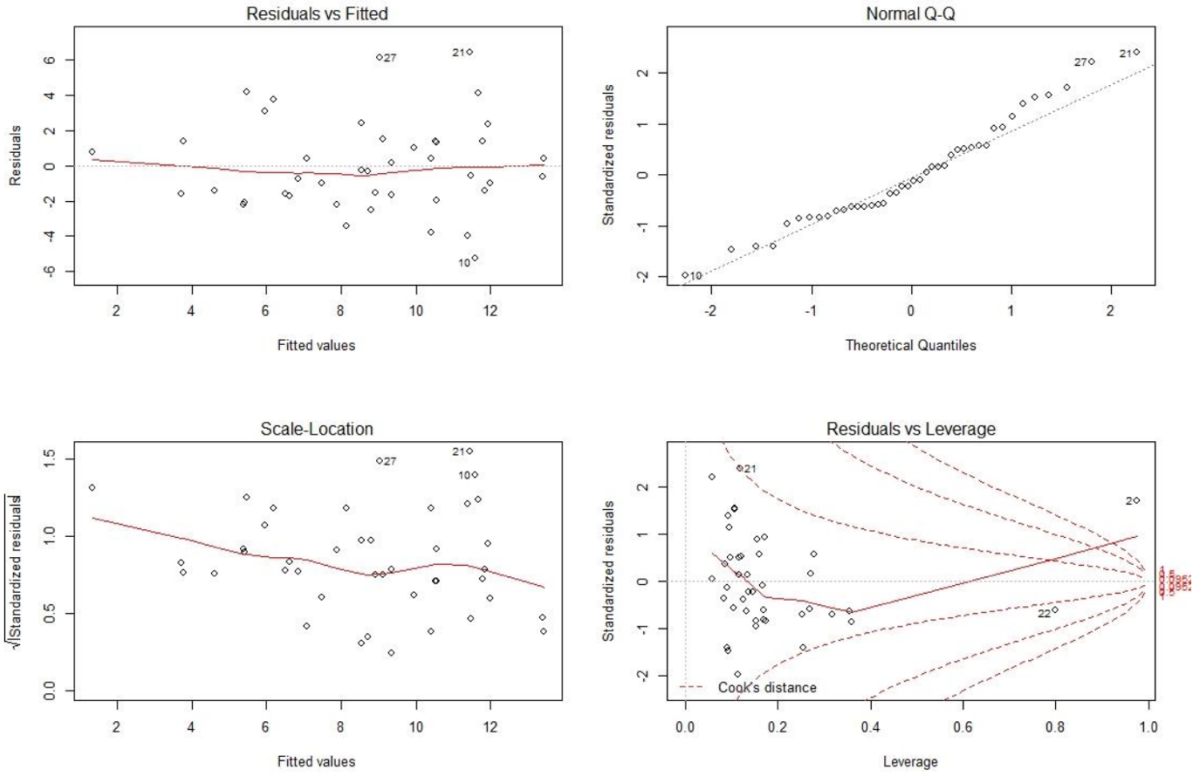
The estimated value for b1 BodyWt = 0.0033. The standard error for *s* BodyWt = 0.0056. The interpretation is that for a one unit increase of BodyWt, the expected value of NonDreaming would ***increase*** by 0.0033 units, holding all other parameters constant. Given the standard error, the value for NonDreaming is likely to increase with a large variation.

Interpret Predation: This problem uses the Level-Level interpretation Y = b0 + b1*X

The estimated value for b5 Predation = 1.415. The standard error for *s* Predation = 1.0273. The interpretation is that for a one unit increase of Predation, the expected value of NonDreaming would ***increase*** by 1.415 units, holding all other parameters constant. Given the standard error, the value for NonDreaming is likely to increase with a large variation.

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

Plots for model3



Assumptions of linearity, constant variance, independence and normality appear to hold. An outlier at observation 2 should be researched and mitigated if deemed necessary. Using a loose cutoff point for Cook's distance = 1, an outlier at observation 2 should be researched and mitigated if deemed necessary. Even if we reduced the cutoff for Cook's distance = 0.05, no additional outliers would need analyzing.

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

R-Code

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

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

```
> summary(model3a)

Call:
lm(formula = log(NonDreaming) ~ BodyWt + BrainWt + LifeSpan +
    Gestation + Exposure + Predation + Danger, 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 *
Exposure     0.0224126  0.0770035   0.291   0.7728
Predation    0.1608420  0.1229089   1.309   0.1994
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
```

The model parameters would be the intercept , the coefficients corresponding to the 7 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 2.6679033$
- $\beta_{BodyWt} = -0.0002651$
- $\beta_{BrainWt} = 0.0001303$
- $\beta_{LifeSpan} = -0.0022691$
- $\beta_{Gestation} = -0.0018214$
- $\beta_{Predation} = 0.1608420$
- $\beta_{Exposure} = 0.0224126$
- $\beta_{Danger} = -0.3207756$
- $\sigma^2 = 0.1168272$

2. What is the equation for the regression line?

$$\hat{log(NonDreaming)} = 2.6679 - 0.0003BodyWt + 0.0001BrainWt - 0.002LifeSpan - 0.0018Gestatio$$
$$+0.1608Predation + 0.0224Exposure - 0.3208Danger$$

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

Gestation, p-val = 0.0265
Danger, p-val = 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.

Two Solutions (A & B) for Interpretation - Please give full credit if student submitted correct solution A or correct solution B.

The reference guide for Log-Linear approximates b1 ~ = 1. The interpretation for formula log(Y) = b0 + b1*X is: As X increases by 1 unit, Y increases by (b1*100)%, holding all other factors constant. This interpretation uses an approximation of b1 ~ = 1, and is an acceptable approximation for values -0.15 > b1 < 0.15. It is less precise for b1 values -0.15 < b1 > 0.15 when interpreting log(Y) models.

The precise and generally preferred formula for log(Y) model interpretation exponentiates the coefficient, and therefore it is accurate for all coefficient values of b1. The precise interpretation for formula log(Y) = b0 + b1*X is: As X increases by 1 unit, Y increases by (e^{b1} – 1)*100%.

In R, this command is (exp(b1) - 1)*100.

Solution A: Using the Reference Guide - not exponentiating the coefficient:

Interpret BodyWt: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X

The estimated value for b1 BodyWt = -0.0003. The standard error for *s* BodyWt = 0.0007. The interpretation is that for a one unit increase of BodyWt, the expected value of NonDreaming would ***decrease*** by 0.03% (this is the same as stating: would *increase* by -0.03%), holding all other parameters constant. For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by |-0.0003*100|%, or *increase* by -0.0003*100%.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Interpret Predation: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X

The estimated value for b6 Predation = 0.1608. The standard error for *s* Predation = 0.1229. The interpretation is that for a one unit increase of Predation, the expected value of NonDreaming would

increase by 16.08% or (0.1608*100)% holding all other parameters constant. Given the standard error, the value for NonDreaming is likely to increase with a large variation.

Solution B: Precise interpretation for log(Y) – exponentiating the coefficient = e^{b1}

Interpret BodyWt: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X using the e^{b1} approach.

The estimated value for b1 BodyWt = -0.0003. The standard error for *s* BodyWt = 0.0007. The interpretation is that for a one unit increase of BodyWt, the expected value of NonDreaming would ***decrease*** by 0.03% holding all other parameters constant. For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by: |(exp(-0.0003) -1)*100| %, or *increase* by: (exp(-0.0003) -1)*100%.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Interpret Predation: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X using the e^{b1} approach.

The estimated value for b6 Predation = 0.1608. The standard error for *s* Predation = 0.1229. The interpretation is that for a one unit increase of Predation, the expected value of NonDreaming would ***increase*** by 17.45% holding all other parameters constant.

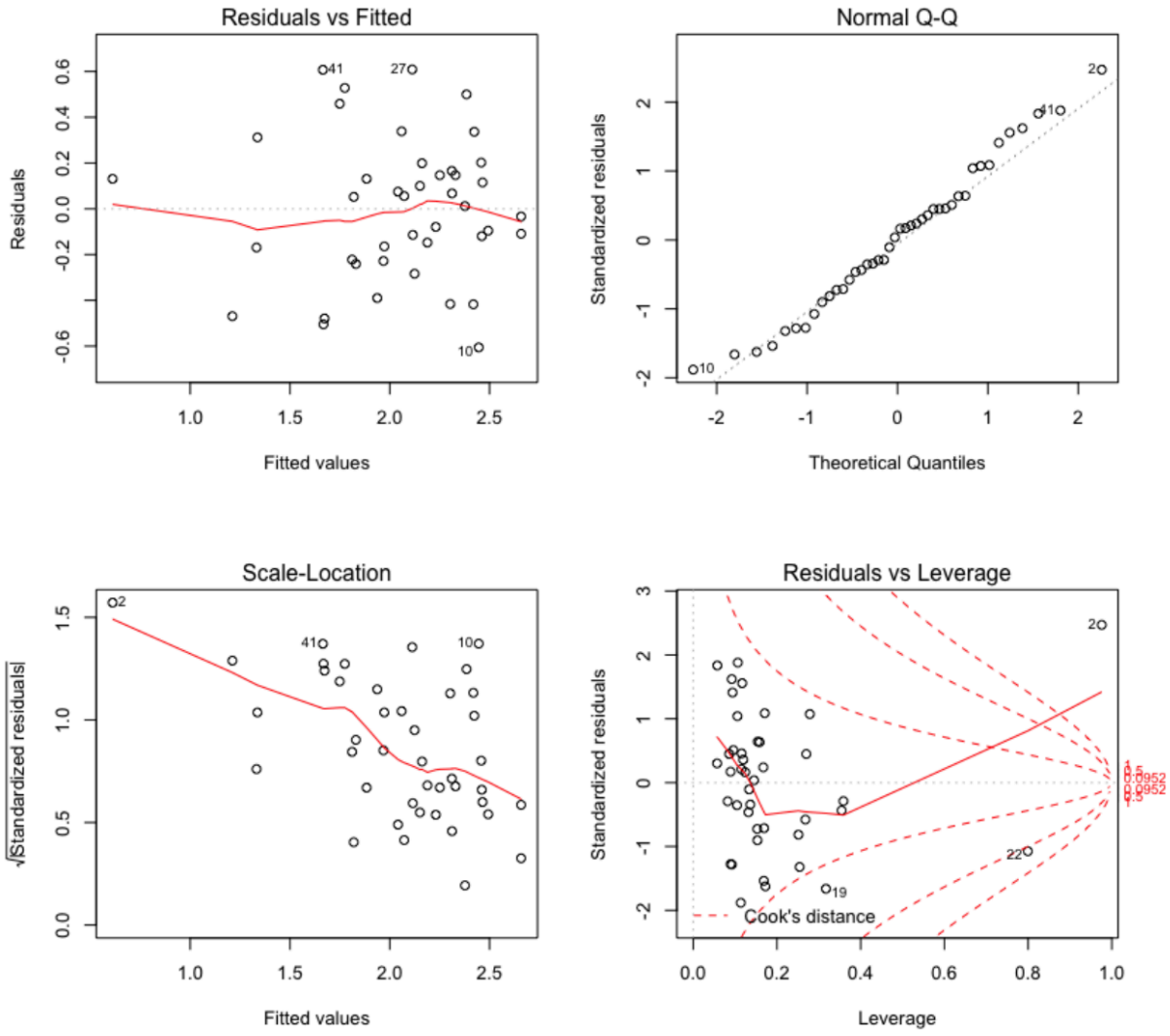
Calculation: (exp(0.1608) -1) *100.

Given the standard error, the value for NonDreaming is likely to increase with a large variation.

(Notice the difference in interpretation of Predation between the two methods A and B)

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

Plots for model3a



Assumptions of linearity, constant variance, independence and normality appear to hold. Using a loose cutoff point for Cook's distance = 1, an outlier at observation 2 should be researched and mitigated if deemed necessary. If we reduced the cutoff for Cook's distance = 0.05, an outlier at observation 22 should also be assessed.

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

R-Code

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

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 7 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 11.4819$
- $\beta_{BodyWt} = -0.3917$
- $\beta_{BrainWt} = -0.5591$
- $\beta_{LifeSpan} = 1.2991$
- $\beta_{Gestation} = -0.5083$
- $\beta_{Exposure} = 0.5310$
- $\beta_{Predation} = 1.6141$
- $\beta_{Danger} = -2.9313$
- $\sigma^2 = 6.6564$

2. What is the equation for the regression line?

$$\hat{NonDreaming} = 11.4819 - 0.3917log(BodyWt) - 0.5591log(BrainWt) + 1.2991log(LifeSpan)$$
$$-0.5083log(Gestation) + 0.5310Exposure + 1.6141Predation - 2.9313Danger$$

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

   Danger, p-val = 0.0166

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

Interpret log(BodyWt): This problem uses the Linear-Log interpretation Y= b0 + b1*log(X)

The estimated value for b1 log(BodyWt) = -0.3919. The standard error for *s* log(BodyWt) = 0.4787. The interpretation is that for a 1% increase of BodyWt, the expected value of NonDreaming would ***decrease*** by 0.0039 units (this is the same as stating: would *increase* by -0.0039 units), holding all other parameters constant). For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by |-0.3919/100|units, or *increase* by -0.3919/100 units.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Interpret Predation: This problem uses the Level-Level interpretation Y = b0 + b1*X

The estimated value for b6 Predation = 1.6141. The standard error for *s* Predation = 0.9812. The interpretation is that for a 1 unit increase of Predation, the expected value of NonDreaming would ***increase*** by 1.6141 units, holding all other parameters constant. Given the standard error, the value for NonDreaming is likely to increase with a large variation.

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

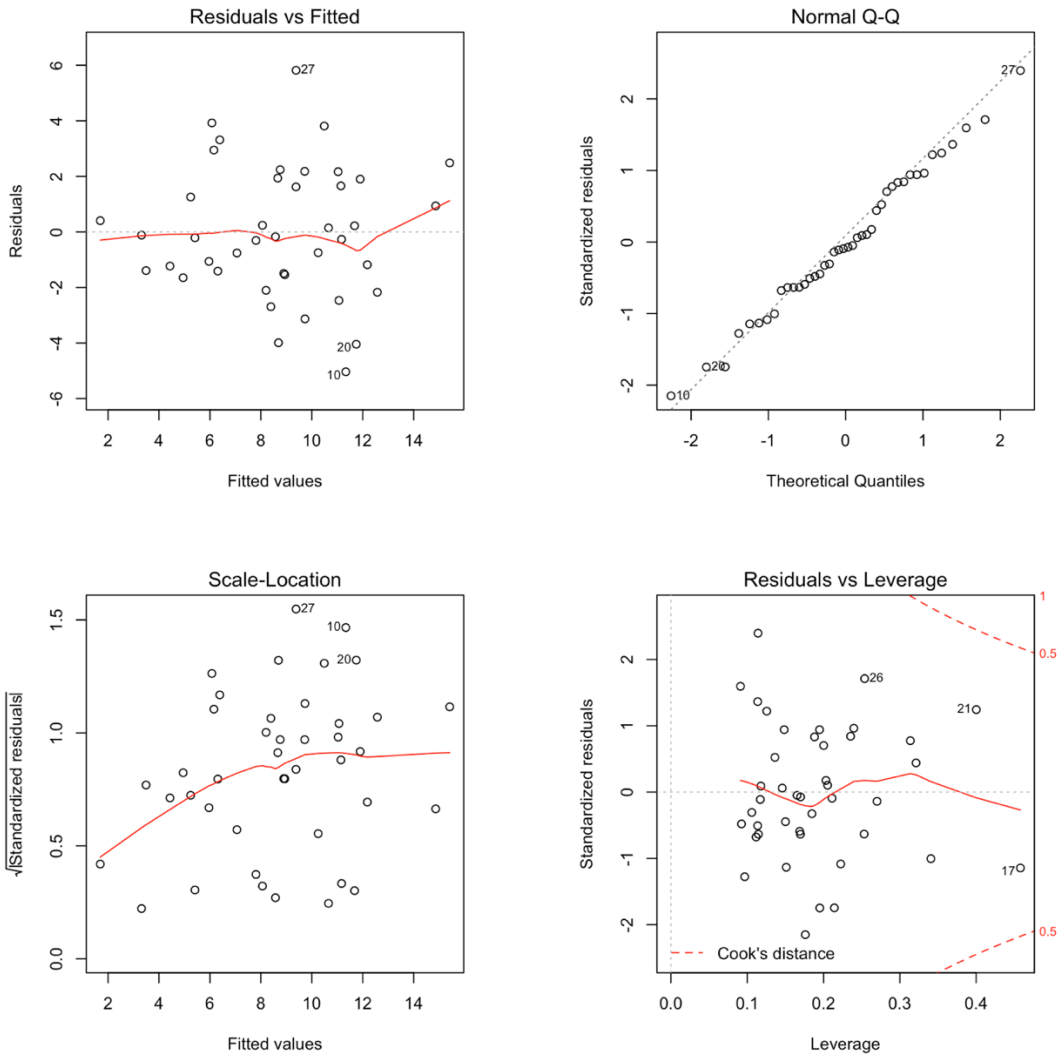
The answer is – we cannot determine this because one of the models uses the transformed response variable log(Y) and the other did not, therefore we cannot compare the models in a statistical sense .

The method we've been taught to confirm the integrity of the model's fit to the data is to review the assumptions of linearity, constant variance, independence and normality, and it seems plausible that we could compare the models through the assumptions.

When we look at the assumptions for model3a and model3b, all assumptions appear to hold for both, but model3b's constant variance, as seen in the Residuals v. Leverage plot, seems more homoscedastic than model3a, and model3b does not show any outliers, where model3a does. This is perhaps as far as we'll be able to compare the models.

Grading: If a student compared the plots and data assumptions between 3a and 3b, please give the student full credit. If a student put thought, analysis and effort into their response, even though their results may not match the solution, please give them full credit.

Plots for model3b



- 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.

R-Code

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

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 5 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 13.0446$
- $\beta_{log(BodyWt)} = -0.4294$
- $\beta_{log(BrainWt)} = -0.4437$
- $\beta_{log(LifeSpan)} = 1.0811$
- $\beta_{log(Gestation)} = -0.6992$
- $\beta_{Danger} = -0.8723$
- $\sigma^2 = 6.911641$

2. What is the equation for the regression line?

$$\hat{NonDreaming} = 13.0446 - 0.4294log(BodyWt) - 0.4437log(BrainWt) + 1.0811log(LifeSpan)$$
$$-0.6992log(Gestation) - 0.8723Danger$$

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

Danger, p-val = 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.

Interpret log(BodyWt): This problem uses the Linear-Log interpretation Y= b0 + b1*log(X)

The estimated value for b1 log(BodyWt) = -0.4294. The standard error for *s* log(BodyWt) = 0.4663. The interpretation is that for a 1% increase of BodyWt, the expected value of NonDreaming would ***decrease*** by 0.0043 units (this is the same as stating: would *increase* by -0.0043 units), holding all other parameters constant). For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by |-0.4294/100|units, or *increase* by -0.4294/100 units.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Interpret Predation: This problem uses the Level-Level interpretation Y = b0 + b1*X

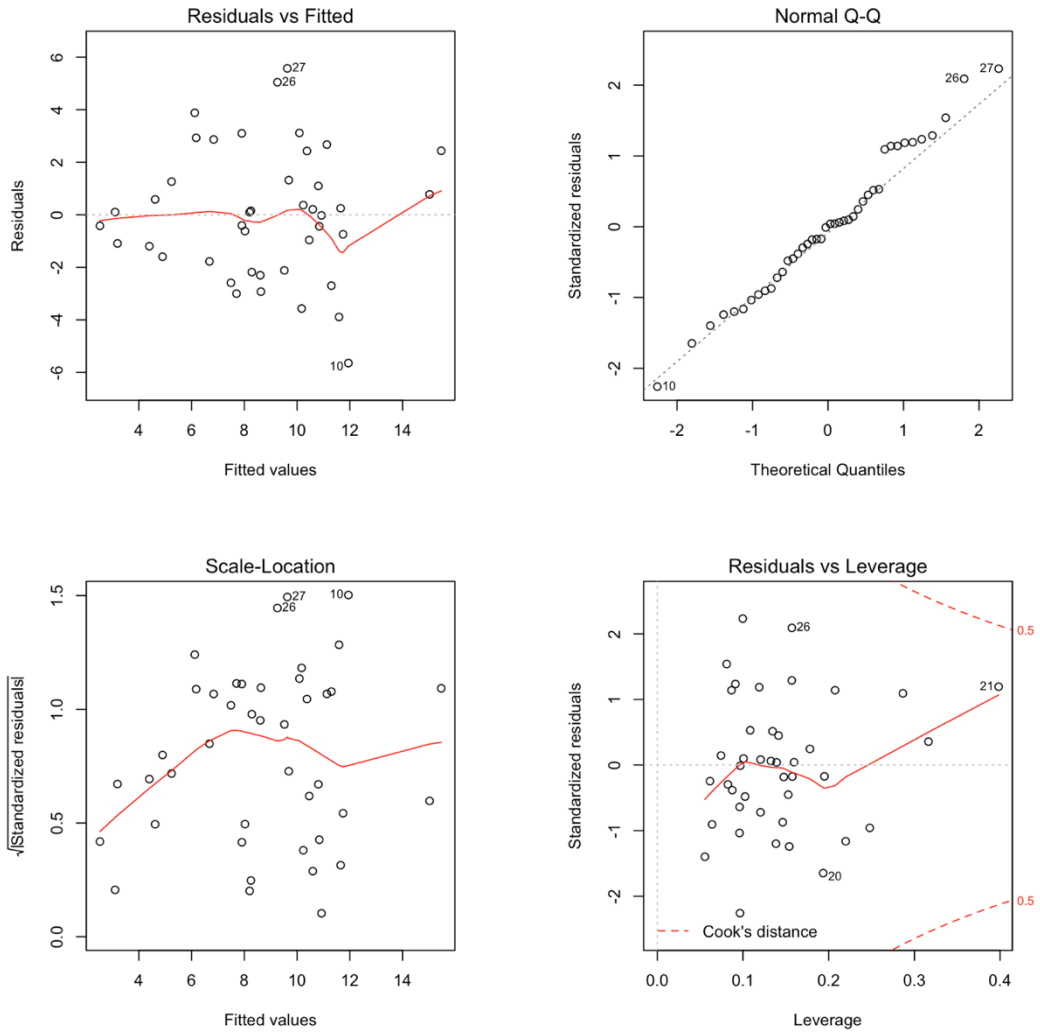
The estimated value for b5 Danger = -0.8723. The standard error for *s* Danger = 0.3249. The interpretation is that for a 1 unit increase of Danger, the expected value of NonDreaming would ***decrease*** by 0.8723 units, holding all other parameters constant. Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

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

Here, we are able to statistically compare the models because the response variables for both are in the same state. Let's use ANOVA to compare model3b with model3c and determine if there is a positive or negative impact on model4c with the removal of Exposure and Danger .

```
> anova(model3b, model3c)
Analysis of Variance Table

Model 1: NonDreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) +
    Exposure + Predation + Danger
Model 2: NonDreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) +
    Danger
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     34 226.24
2     36 248.89 -2   -22.646 1.7017 0.1975
```

We see that the Df of -2 shows the subtraction of the two predictors Exposure and Danger, and the high p-value indicates that the change to model3c reduced the fit of the data, and that model3b is the better choice.



- 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.

R-Code

Finalmodel3 <- lm(log(NonDreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) + Danger, data = data)

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 5 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 2.78927$
- $\beta_{log(BodyWt)} = -0.11203$
- $\beta_{log(BrainWt)} = 0.03583$
- $\beta_{log(LifeSpan)} = 0.07153$
- $\beta_{log(Gestation)} = -0.13977$
- $\beta_{Danger} = -0.11576$
- $\sigma^2 = 0.1099586$

2. What is the equation for the regression line?

$$\hat{log(NonDreaming)} = 2.78927 - 0.11203 log(BodyWt) + 0.03583 log(BrainWt) + 0.07153 log(LifeSpan)$$
$$- 0.13977 log(Gestation) - 0.11576 Danger$$

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

   Danger, p-val = 0.0077

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

Two Solutions (A & B) for Interpretation - Please give full credit if student submitted correct solution A or correct solution B.

The reference guide for Log-Log approximates b1 ~ = 1. The interpretation for formula log(Y) = b0 + b1*log(X) is: As X increases by 1%, Y increases by (b1) %, holding all other factors constant. This interpretation uses an approximation of b1 ~ = 1, and is an acceptable approximation for smaller values of b1. It is less precise for larger b1 values when interpreting Log-Log models.

The precise and generally preferred formula for Log-Log model interpretation exponentiates the coefficient/100, and therefore it is accurate for all coefficient values of b1. The precise interpretation for formula log(Y) = b0 + b1*log(X) is: As X increases by 1 unit, Y increases by ($e^{b1/100}$ – 1)*100%. In R, this command is (exp(b1/100) - 1)*100.

Solution A: Using the Reference Guide for Log-Log - not exponentiating the coefficient/100:

Interpret log(BodyWt): This problem uses the Log-Log interpretation for log(Y) = b0 + b1*log(X)

The estimated value for b1 log(BodyWt) = -0.1120. The standard error for *s* log(BodyWt) = 0.0588. The interpretation is that for a one % increase of BodyWt, the expected value of NonDreaming would ***decrease*** by 0.1120 % (this is the same as stating: would *increase* by -0.1120%), holding all other parameters constant. For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by |-0.1120|%, or *increase* by -0.1120 %.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Interpret Danger: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X

The estimated value for b5 Danger = -0.1158. The standard error for *s* Danger = 0.0410. The interpretation is that for a one unit increase of Danger, the expected value of NonDreaming would *decrease* by (0.1158*100)%, or **decrease by 11.58%,** holding all other parameters constant. Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Solution B: Precise interpretation for Log-Log - exponentiating the coefficient/100 = $e^{b1/100}$

Interpret log(BodyWt): This problem uses the Log-Log interpretation for log(Y) = b0 + b1*log(X) using the $e^{b1/100}$ approach.

The estimated value for b1 log(BodyWt) = -0.1120. The standard error for *s* log(BodyWt) = 0.0588. The interpretation is that for a 1% increase of BodyWt, the expected value of NonDreaming would ***decrease*** **by 0.1120 %** (this is the same as stating: would *increase* by – 0.1120 %), holding all other parameters constant. For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by: |(exp(-0.1120/100) -1)*100| %, or *increase* by: (exp(-0.1120/100) -1)*100%.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

Interpret Danger: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X using the e^{b1} approach.

The estimated value for b5 Danger = -0.1158. The standard error for *s* Danger = 0.0410. The interpretation is that for a one unit increase of Danger, the expected value of NonDreaming would ***decrease*** **by 10.93%** holding all other parameters constant. For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

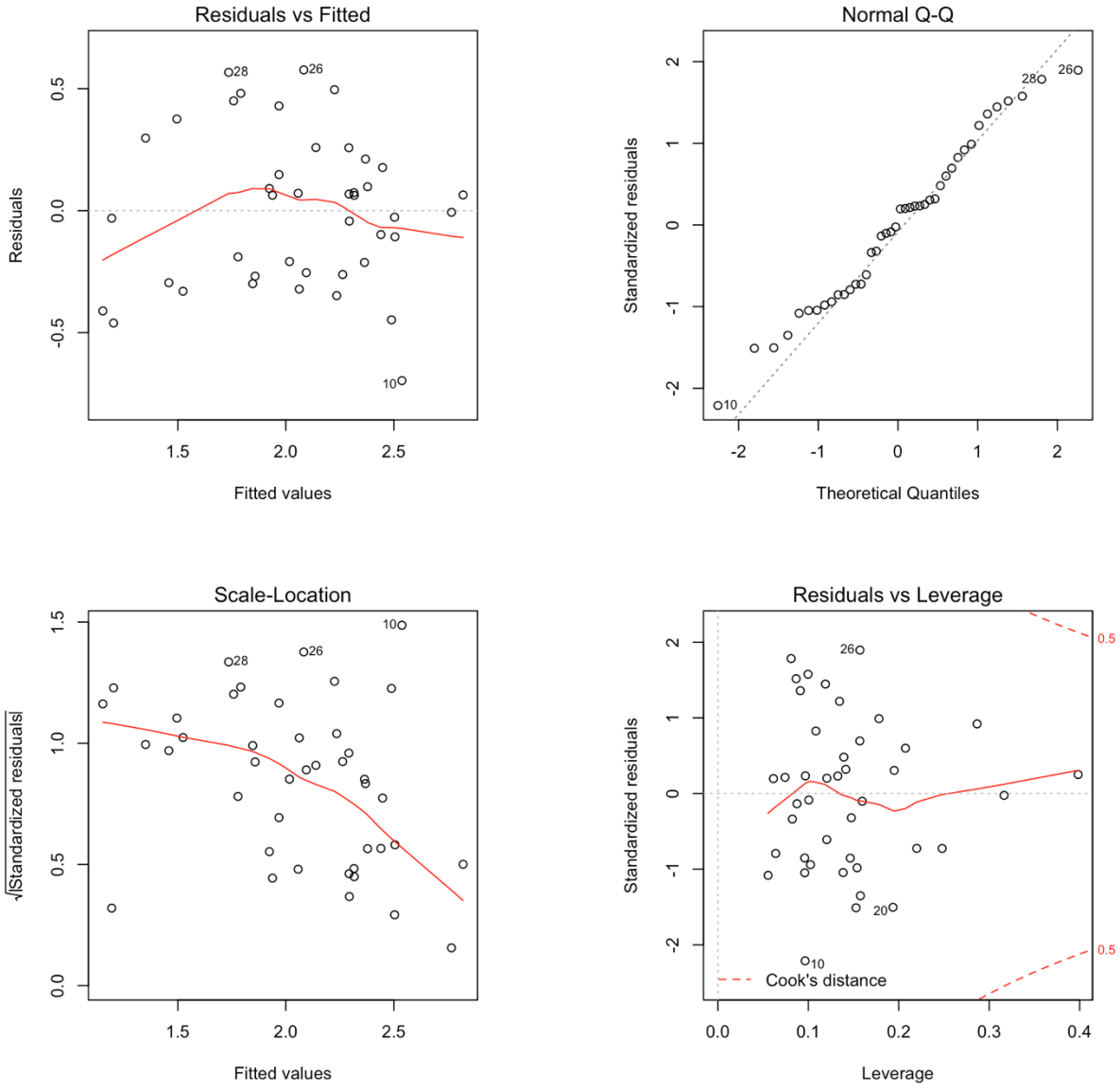
Calculations: *decrease* by: |(exp(-0.1158) -1)*100| %, or *increase* by: (exp(-0.1158) -1)*100 %.

Given the standard error, the value for NonDreaming is likely to decrease with a large variation.

(Notice the difference in interpretation of Danger between the two methods A and B)

5. Did finalmodel3 improve over model3c? Explain how you determined if the model improved or not.

Plots for finalmodel3



Here again, we are not able to compare the statistical values between these models, due to the response log(Y) transform on finalmodel3, and the untransformed response on model3c.

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.

R-Code #remove row 11

data2 = data[-11,] #remove row 11 and use data2 as variable for Q4 and Q6

Question 4: Fitting the Linear Regression Model – Response Variable Dreaming

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.

R Code
```
model4 <- lm(Dreaming ~. -Species -Dreaming -TotalSleep, data = data)
summary(model4)
plot(model4, cook.levels = c(4/41,0.5,1))
```

Q4 Solution

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 7 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 3.820083$
- $\beta_{BodyWt} = 0.003615$
- $\beta_{BrainWt} = -0.001041$
- $\beta_{LifeSpan} = 0.011875$
- $\beta_{Gestation} = -0.007219$
- $\beta_{Predation} = 0.859964$
- $\beta_{Exposure} = 0.295094$
- $\beta_{Danger} = -1.675645$
- $\sigma^2 = 0.7162237$

2. What is the equation for the regression line?

$$\hat{Dreaming} = 3.820083 + 0.003615 BodyWt - 0.001041 BrainWt + 0.011875 LifeSpan - 0.007219 Gestation$$
$$+ 0.859964 Predation + 0.295094 Exposure - 1.675645 Danger$$

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

Gestation, p-val = 0.00147
Predation, p-val = 0.008
Danger, p-val = 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.

Interpret BodyWt: This problem uses the Level-Level interpretation Y = b0 + b1*X

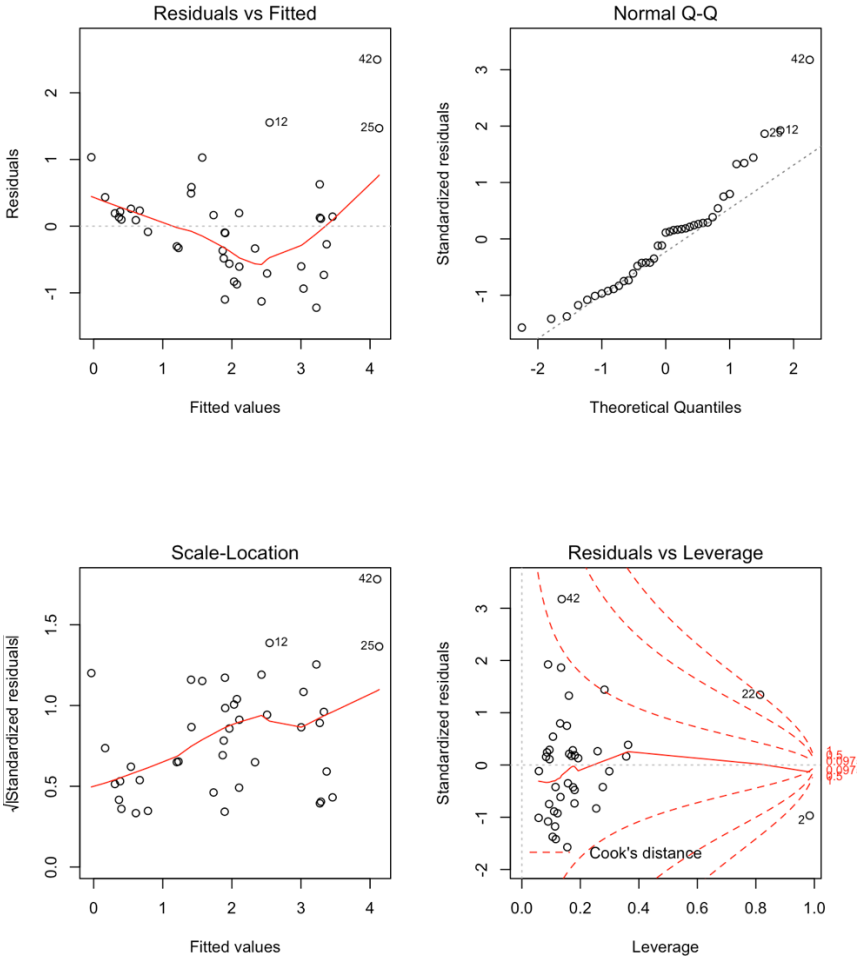
The estimated value for b1 BodyWt = 0.0036. The standard error for *s* BodyWt = 0.001813. The interpretation is that for a one unit increase of BodyWt, the expected value of Dreaming would ***increase*** by 0.0036 units, holding all other parameters constant. Given the standard error, the value for Dreaming is likely to increase with a large variation.

Interpret Predation: This problem uses the Level-Level interpretation Y = b0 + b1*X

The estimated value for b5 Predation = 0.859964. The standard error for *s* Predation = 0.304644. The interpretation is that for a one unit increase of Predation, the expected value of Dreaming would ***increase*** **by 0.86 units**, holding all other parameters constant. Given the standard error, the value for Dreaming is likely to increase with a large variation.

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

Plots for model4



We may have issues with linearity based on the Residuals v. Fitted plot, and the Normality may be in question with skewness. Constant variance and independence appear to hold based on the Residuals v. Leverage plot. Outliers are present at Cook's =1 and at Cook's = 0.05, these would need to be assessed for further action if necessary.

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

R-Code
model4a <- lm(log(Dreaming) ~. -Species -Dreaming -TotalSleep, data = data)

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

```
> summary(model4a)

Call:
lm(formula = log(Dreaming) ~ BodyWt + BrainWt + LifeSpan + Gestation +
    Exposure + Predation + Danger, 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 ***
Exposure     0.1015576  0.0743207   1.366  0.18103
Predation    0.3747033  0.1179789   3.176  0.00323 **
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
```

The model parameters would be the intercept , the coefficients corresponding to the 7 predictors as well as the square of the residual standard error. Parameter estimates:

B0 = 1.4684458
B1BodyWt = 0.0020701
B2BrainWt = -0.0005974
B3LifeSpan = 0.0104864
B4Gestation = -0.0042079
B5Exposure = 0.1015576
B6Predation = 0.3747033
B7Danger = -0.7743707
Sigma^2 = 0.1073872

Grading: If the RSE-squared (sigma^2) was not included in the list of parameters with its estimated value, this question should be graded incorrect for one of the five models in Q4. Please don't mark more than one incorrect to avoid doubling up on penalties.

1. What is the equation for the regression line?

Log(Dreaming)^hat = 1.468 + 0.0021(BodyWt) – 0.0006(BrainWt) + 0.1015(LifeSpan) – 0.0042(Gestation) + 0.1016(Exposure) + 0.3747033(Predation) – 0.7744(Danger)

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

BodyWt, p-val = 0.00582
Gestation, p-val = 9.57e-06
Predation, p-val = 0.00323
Danger, p-val = 8.00e-06

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

Two Solutions (A & B) for Interpretation - Please give full credit if student submitted correct solution A or correct solution B.

The reference guide for Log-Linear approximates b1 ~ = 1. The interpretation for formula log(Y) = b0 + b1*X is: As X increases by 1 unit, Y increases by (b1*100)%, holding all other factors constant. This interpretation uses an approximation of b1 ~ = 1, and is an acceptable approximation for values -0.15 > b1 < 0.15. It is less precise for b1 values -0.15 < b1 > 0.15 when interpreting log(Y) models.

The precise and generally preferred formula for log(Y) model interpretation exponentiates the coefficient, and therefore it is accurate for all coefficient values of b1. The precise interpretation for formula log(Y) = b0 + b1*X is: As X increases by 1 unit, Y increases by (e^{b1} – 1)*100%.

In R, this command is (exp(b1) - 1)*100.

Solution A: Using the Reference Guide - not exponentiating the coefficient:

Interpret BodyWt: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X

The estimated value for b1 BodyWt = 0.0021. The standard error for *s* BodyWt = 0.0007. The interpretation is that for a one unit increase of BodyWt, the expected value of Dreaming would ***increase*** by 0.21% or (0.0021*100)% holding all other parameters constant.

Given the standard error, the value for Dreaming is likely to increase with a large variation.

Interpret Predation: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X

The estimated value for b6 Predation = 0.3747033. The standard error for *s* Predation = 0.1179789. The interpretation is that for a one unit increase of Predation, the expected value of Dreaming would ***increase*** by 37.47% or (0.3747*100)% holding all other parameters constant. Given the standard error, the value for Dreaming is likely to increase with a large variation.

Solution B: Precise interpretation for log(Y) – exponentiating the coefficient = e^{b1}

Interpret BodyWt: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X using the e^{b1} approach.

The estimated value for b1 BodyWt = 0.0021. The standard error for *s* BodyWt = 0.0007. The interpretation is that for a one unit increase of BodyWt, the expected value of Dreaming would ***increase by*** 0.21% holding all other parameters constant.

Calculations: *increase* by: (exp(0.0021) -1)*100%.

Given the standard error, the value for Dreaming is likely to decrease with a large variation.

Interpret Predation: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X using the e^{b1} approach.

The estimated value for b6 Predation = 0.3747033. The standard error for *s* Predation = 0.1179789. The interpretation is that for a one unit increase of Predation, the expected value of Dreaming would ***increase*** by 45.56% holding all other parameters constant.

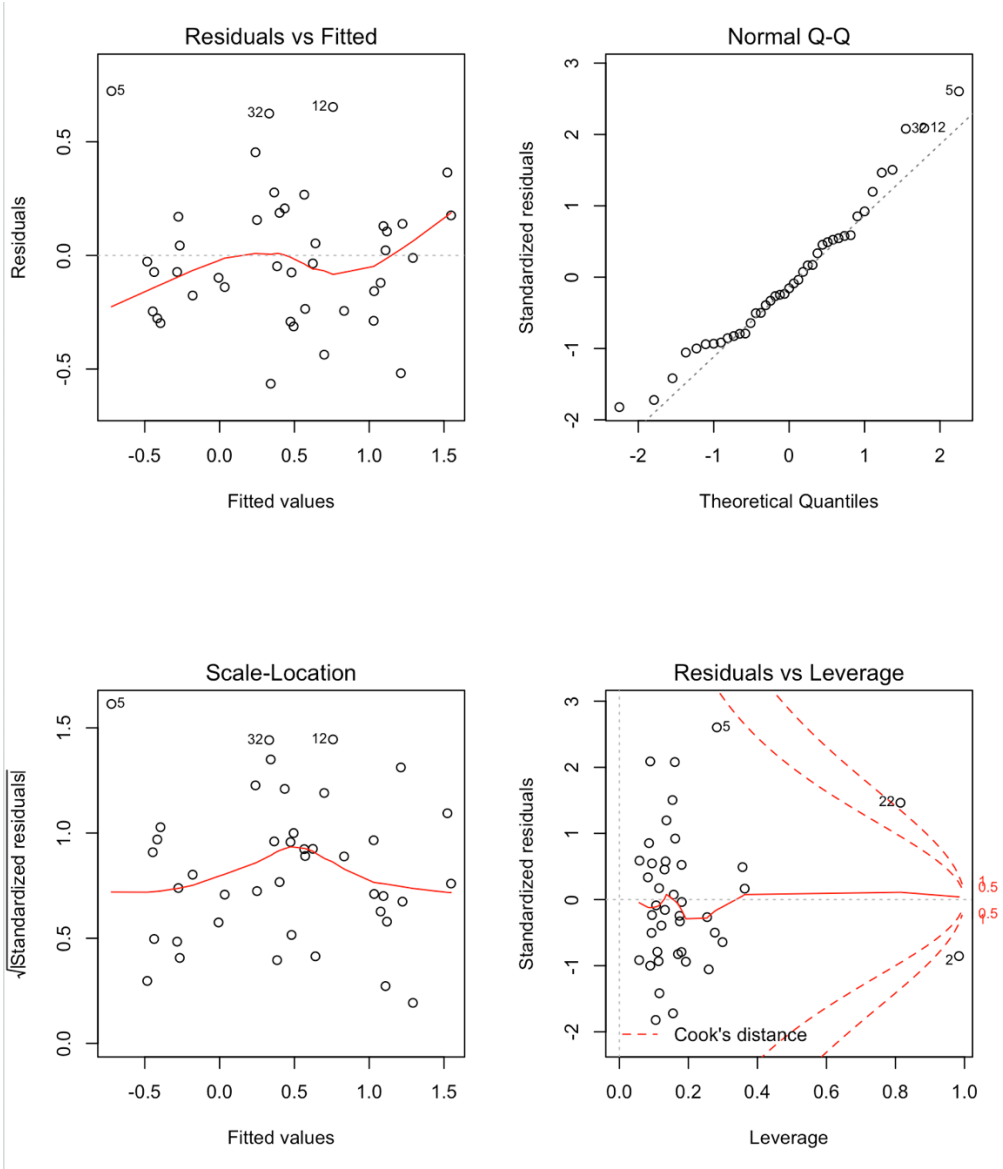
Calculation: (exp(0.3747033) -1) *100.

Given the standard error, the value for Dreaming is likely to increase with a large variation.

(Notice the difference in interpretation of Predation between the two methods A and B)

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

Plots for model4a



Assumptions of linearity, constant variance, independence and normality appear to hold. Using a loose cutoff point for Cook's distance = 1, we find two outliers at observation 2 and 22 that should be assessed.

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

R-Code

model4b <- lm(Dreaming ~ log(BodyWt) +log(BrainWt) +log(LifeSpan) +log(Gestation) +Exposure +Predation +Danger, data = data)

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 7 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 7.43127$
- $\beta_{log(BodyWt)} = 0.44017$
- $\beta_{log(BrainWt)} = -0.35662$
- $\beta_{log(LifeSpan)} = 0.02462$
- $\beta_{log(Gestation)} = -0.82406$
- $\beta_{Exposure} = 0.26488$
- $\beta_{Predation} = 0.59634$
- $\beta_{Danger} = -1.36005$
- $\sigma^2 = 0.509796$

2. What is the equation for the regression line?

$$\hat{Dreaming} = 7.43127 + 0.44017log(BodyWt) - 0.35662log(BrainWt) + 0.02462log(LifeSpan)$$
$$-0.82406log(Gestation) + 0.26488Exposure + 0.59634Predation - 1.36005Danger$$

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

Log(BodyWt), p-val = 0.002195

Log(Gestation), p-val = 0.000145

Predation, p-val = 0.035392

Danger, p-val = 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.

Interpret log(BodyWt): This problem uses the Linear-Log interpretation Y= b0 + b1*log(X)

The estimated value for b1 log(BodyWt) = 0.44017. The standard error for *s* log(BodyWt) = 0.13251. The interpretation is that for a 1% increase of BodyWt, the expected value of Dreaming would ***increase*** by 0.0044 units, holding all other parameters constant).

Calculations: *increase* by 0.44017/100 units.

Given the standard error, the value for Dreaming is likely to increase with a large variation.

Interpret Predation: This problem uses the Level-Level interpretation Y = b0 + b1*X

The estimated value for b6 Predation = 0.59634. The standard error for *s* Predation = 0.27182. The interpretation is that for a 1 unit increase of Predation, the expected value of Dreaming would ***increase*** by 0.59634 units, holding all other parameters constant. Given the standard error, the value for Dreaming is likely to increase with a large variation.

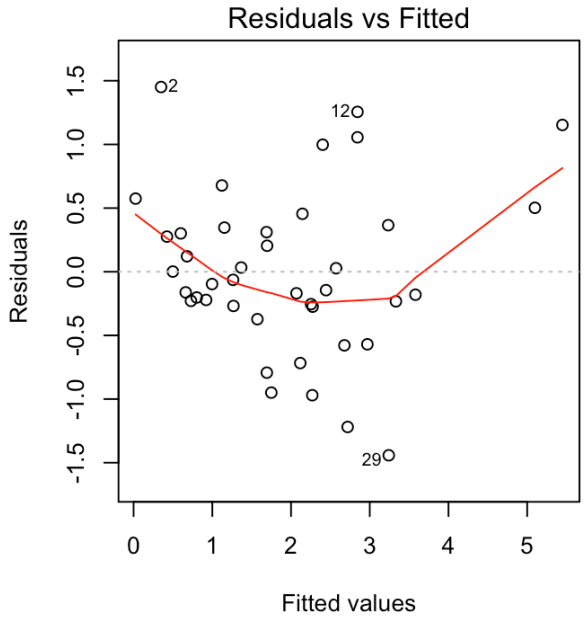
5.Did model4b improve over model4a? Explain how you determined if the model improved or not.

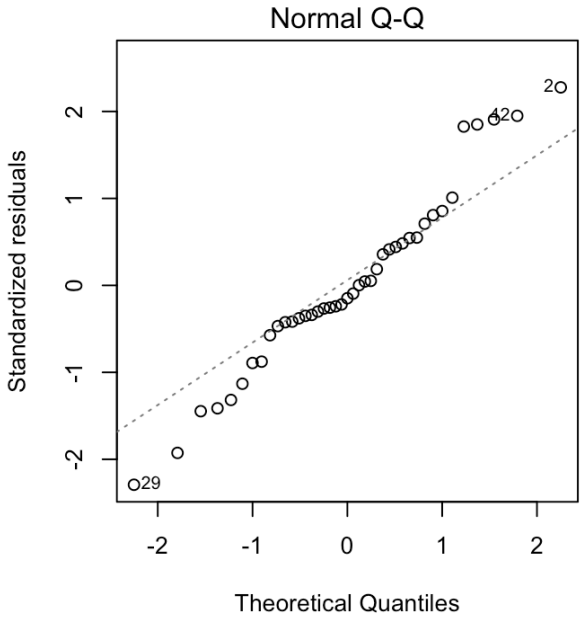
The answer is – we cannot determine this because one of the models uses the transformed response variable log(Y) and the other did not, therefore we cannot compare the models in a statistical sense (at this point in the course).

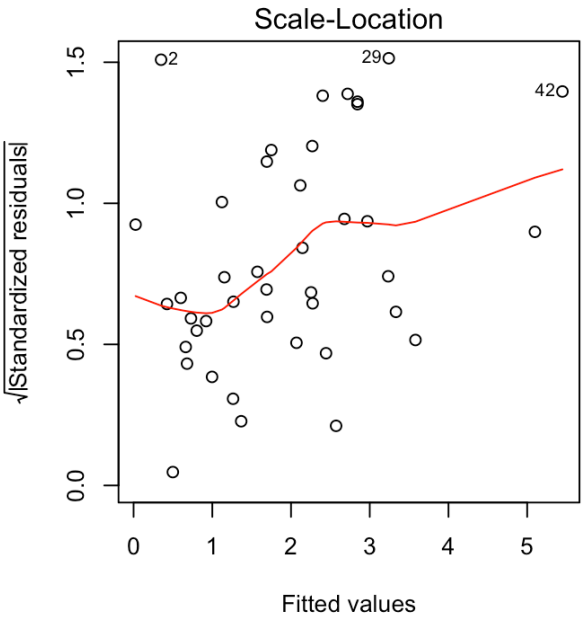
The method we've been taught to confirm the integrity of the model's fit to the data is to review the assumptions of linearity, constant variance, independence and normality, and it seems plausible that we could compare the models through the assumptions. When we look at the plots for model4a and model4b, model4a plots look slightly better. Model4b may have issues with linearity and but all assumptions appear to hold, so there isn't enough evidence to show that either model is significantly better than the other.

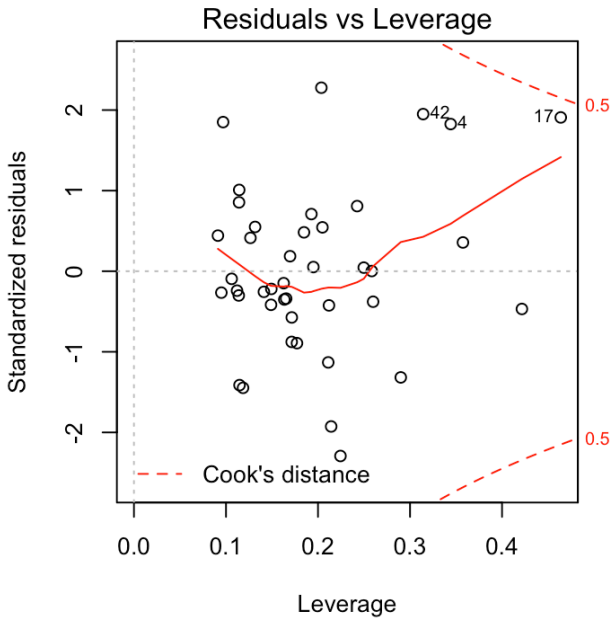
Grading: If a student compared the plots and data assumptions between 4a and 4b, please give the student full credit. If a student put thought, analysis and effort into their response, even though it may not match the solution, please give them full credit.

Plots for model4b









- 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.

R-Code

model4c <- lm(Dreaming ~ log(BodyWt) +log(BrainWt) +log(LifeSpan) +log(Gestation) +Danger, data = data)

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

```
> 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
```

The model parameters would be the intercept , the coefficients corresponding to the 5 predictors as well as the square of the residual standard error. Parameter estimates:

- $\beta_0 = 7.94819$
- $\beta_{log(BodyWt)} = 0.43897$
- $\beta_{log(BrainWt)} = -0.32370$
- $\beta_{log(LifeSpan)} = -0.02089$
- $\beta_{log(Gestation)} = -0.88941$
- $\beta_{Danger} = -0.55009$
- $\sigma^2 = 0.5840016$

2.What is the equation for the regression line?

$$\hat{Dreaming} = 7.94819 + 0.43897log(BodyWt) - 0.32370log(BrainWt) - 0.02089log(LifeSpan)$$
$$-0.88941log(Gestation) - 0.55009Danger$$

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

Log(BodyWt), p-val =0.00263
Log(Gestation), p-val = 5.88e-05
Danger, p-val = 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.

Interpret log(BodyWt): This problem uses the Linear-Log interpretation Y= b0 + b1*log(X)

The estimated value for b1 log(BodyWt) = 0.43897. The standard error for *s* log(BodyWt) = 0.13553. The interpretation is that for a 1% increase of BodyWt, the expected value of Dreaming would ***increase by*** 0.0044 units holding all other parameters constant).

Calculations: *increase* by 0.42897/100 units.

Given the standard error, the value for Dreaming is likely to increase with a large variation.

Interpret Predation: This problem uses the Level-Level interpretation Y = b0 + b1*X

The estimated value for b5 Danger = -0.55009. The standard error for *s* Danger = 0.09443. The interpretation is that for a 1 unit increase of Danger, the expected value of Dreaming would ***decrease*** by 0.55 units, holding all other parameters constant. Given the standard error, the value for Dreaming is likely to decrease with a large variation.

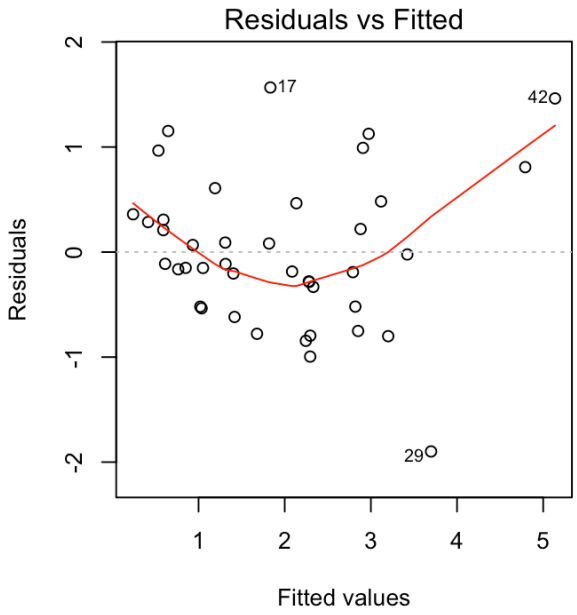
5.Did model4c improve over model4b? Explain how you determined if the model improved or not.

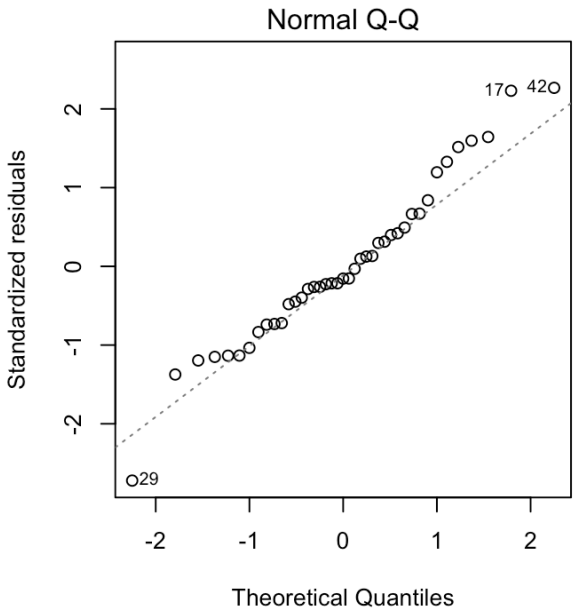
Here, we are able to statistically compare the models because the response variables for both are in the same state. Let's use ANOVA to compare model4b with model4c and determine if there is a positive or negative impact on model4c with the removal of Exposure and Danger .

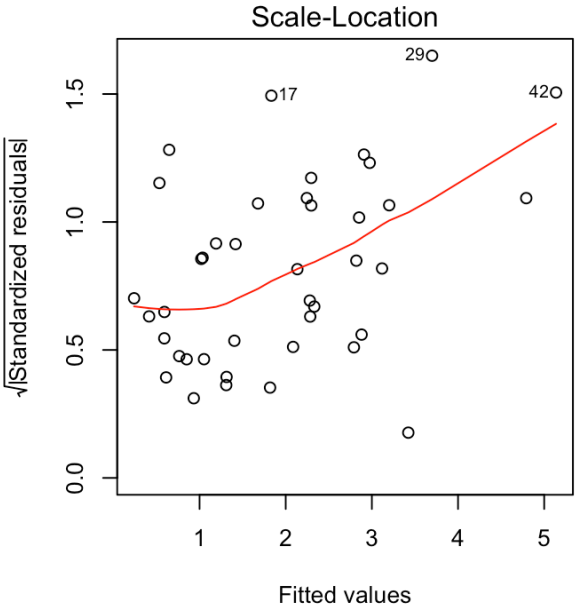
```
> anova(model4b, model4c)
Analysis of Variance Table

Model 1: Dreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) +
    Exposure + Predation + Danger
Model 2: Dreaming ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) +
    Danger
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     33 16.824
2     35 20.440 -2   -3.6164 3.5468 0.04025 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

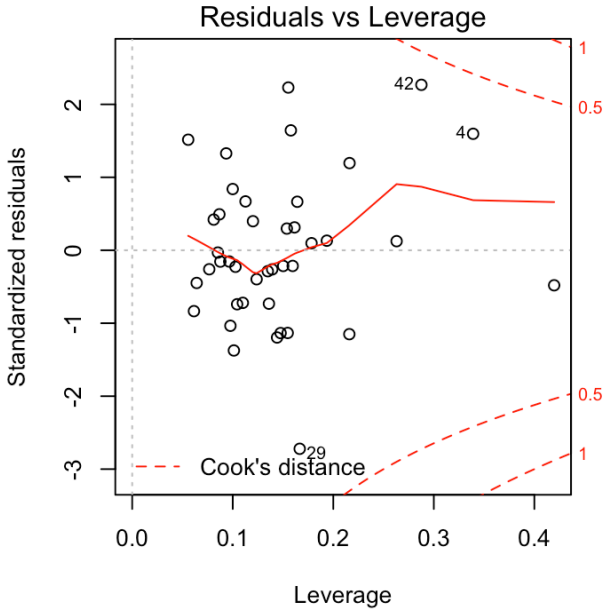
We see that the Df of -2 shows the subtraction of the two predictors Exposure and Danger, and the p-value is significant at alpha = 0.05, but not at alpha =0.01.

Plots for model4c









- 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 model4c as your baseline.

R-Code

Finalmodel4 <- lm(log(Dreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) + Danger, data = data)

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

```
> summary(finalmodel4)

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
```

The model parameters would be the intercept , the coefficients corresponding to the 5 predictors as well as the square of the residual standard error. Parameter estimates:

B0 = 3.10346
B*log*(BodyWt) = 0.14340
B*log*(BrainWt) = -0.11064
Blog(LifeSpan) = 0.05317
Blog(Gestation) = -0.41163
BDanger = -0.28973
sigma^2 = 0.1441

2.What is the equation for the regression line?

$$\hat{log(Dreaming)} = 3.10346 + 0.14340log(BodyWt) - 0.11064log(BrainWt) + 0.05317log(LifeSpan)$$
$$-0.41163log(Gestation) - 0.28973Danger$$

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

Log(BodyWt), p-val = 0.040252
Log(Gestation), p-val = 0.000148
Danger, p-val = 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.

Two Solutions (A & B) for Interpretation - Please give full credit if student submitted correct solution A or correct solution B.

The reference guide for Log-Log approximates b1 ~ = 1. The interpretation for formula log(Y) = b0 + b1*log(X) is: As X increases by 1%, Y increases by (b1) %, holding all other factors constant. This interpretation uses an approximation of b1 ~ = 1, and is an acceptable approximation for smaller values of b1. It is less precise for larger b1 values when interpreting Log-Log models.

The precise and generally preferred formula for Log-Log model interpretation exponentiates the coefficient/100, and therefore it is accurate for all coefficient values of b1. The precise interpretation for formula log(Y) = b0 + b1*log(X) is: As X increases by 1 unit, Y increases by ($e^{b1/100}$ – 1)*100%. In R, this command is (exp(b1/100) - 1)*100.

Solution A: Using the Reference Guide for Log-Log - not exponentiating the coefficient/100:

Interpret log(BodyWt): This problem uses the Log-Log interpretation for log(Y) = b0 + b1*log(X)

The estimated value for b1 log(BodyWt) = 0.14340. The standard error for *s* log(BodyWt) = 0.06732. The interpretation is that for a one % increase of BodyWt, the expected value of Dreaming would ***increase*** by 0.1434 holding all other parameters constant. Given the standard error, the value for Dreaming is likely to decrease with a large variation.

Interpret Danger: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X

The estimated value for b5 Danger = -0.28973. The standard error for *s* Danger = 0.04690. The interpretation is that for a one unit increase of Danger, the expected value of Dreaming would *decrease* by (0.2897*100)%, or **decrease by 28.97%,** holding all other parameters constant. Given the standard error, the value for Dreaming is likely to decrease with a large variation.

Solution B: Precise interpretation for Log-Log - exponentiating the coefficient/100 = $e^{b1/100}$

Interpret log(BodyWt): This problem uses the Log-Log interpretation for log(Y) = b0 + b1*log(X) using the $e^{b1/100}$ approach.

The estimated value for b1 log(BodyWt) = 0.14340. The standard error for *s* log(BodyWt) = 0.06732. The interpretation is that for a 1% increase of BodyWt, the expected value of Dreaming would ***increase*** **by 0.1435%** holding all other parameters constant.

Calculations: *increase* by: (exp(0.14340/100) -1)*100%.

Given the standard error, the value for Dreaming is likely to increasee with a large variation.

Interpret Danger: This problem uses the Log-Linear interpretation for log(Y) = b0 + b1*X using the e^{b1} approach.

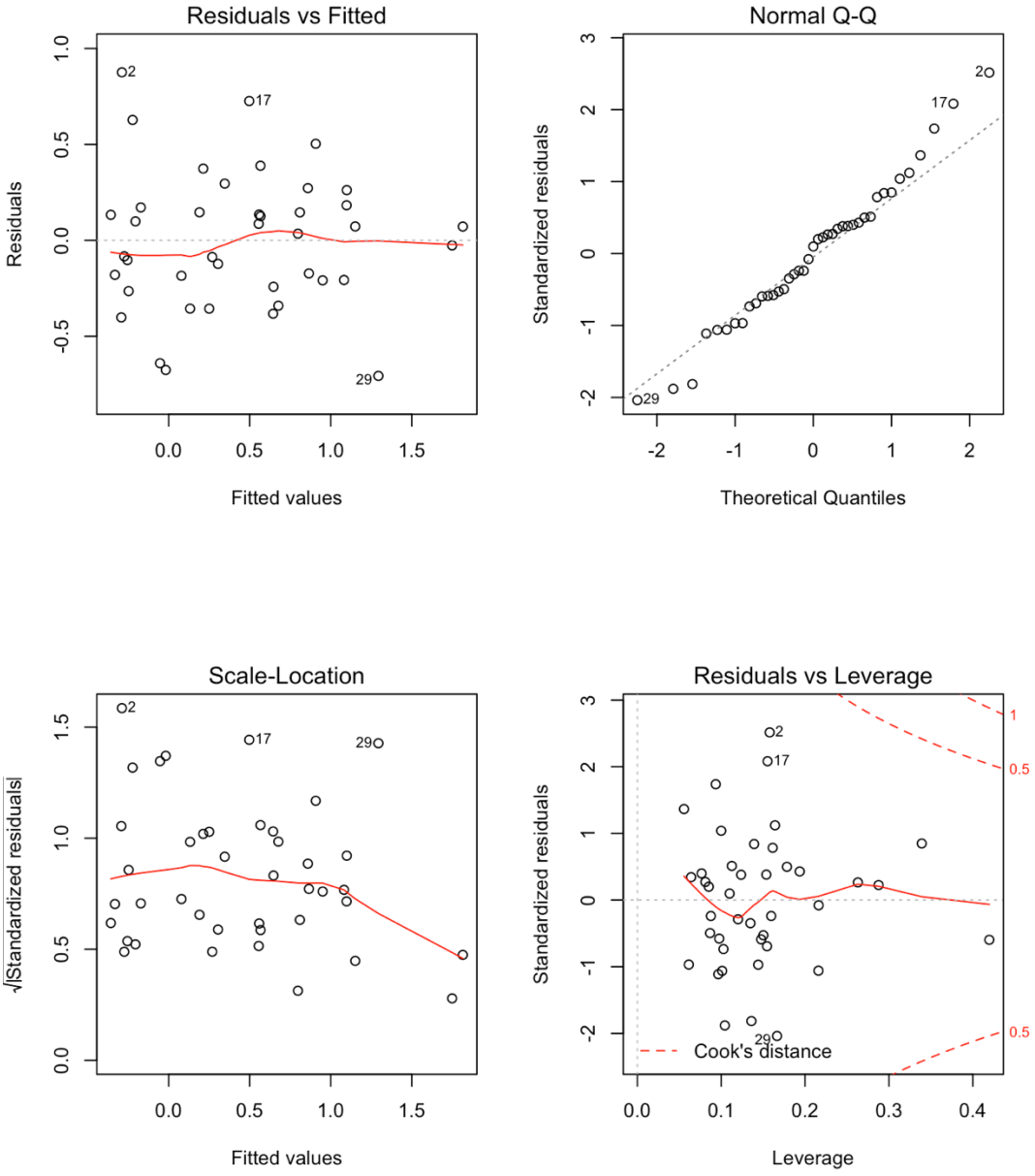
The estimated value for b5 Danger = -0.28973. The standard error for *s* Danger = 0.04690. The interpretation is that for a one unit increase of Danger, the expected value of Dreaming would ***decrease*** **by 25.15%** holding all other parameters constant. For model interpretation, we attempt to minimize confusion, as such, it is simpler to understand the "decrease by" value.

Calculations: *decrease* by: |(exp(-0.2897) -1)*100| %, or *increase* by: (exp(-0.2897) -1)*100 %.

Given the standard error, the value for Dreaming is likely to decrease with a large variation.

(Notice the difference in interpretation for Danger between the two methods A and B)

Plots for finalmodel4



5.Did finalmodel4 improve over model4c? Explain how you determined if the model improved or not.

Here again, we are not able to compare the statistical values between these models, due to the response log(Y) transform on finalmodel4, and the untransformed response on model4c.

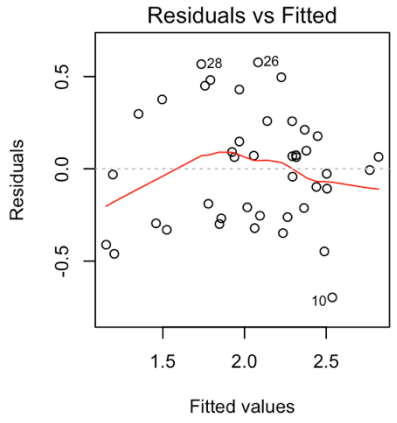
Question 5 - Checking the Assumptions of the Model (**response is NonDreaming**)

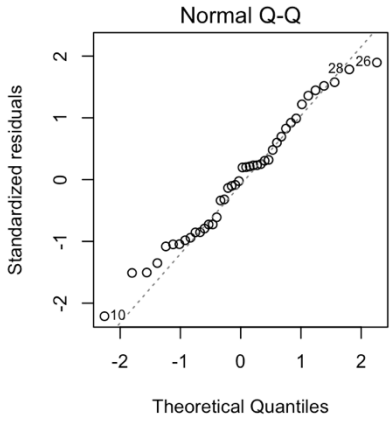
- 5. Plot the relevant residual plots to check the final model assumptions. Individual predictors were assessed for linearity in Question 1 and Question 2. Model linearity should be assessed along with constant variance, independence and normality. You should have 3 plots. Enumerate the assumptions and 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.

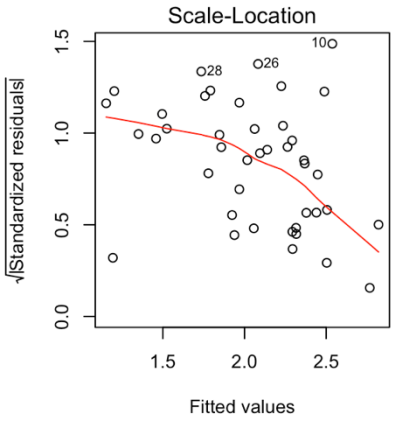
R-Code

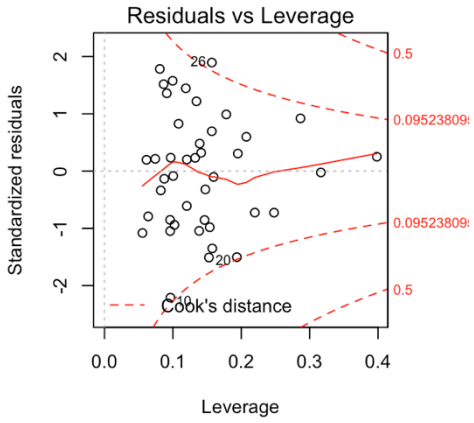
```
Finalmodel3 <- lm(log(NonDreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) + Danger, data = data)
summary(finalmodel3)
par(mfrow = c(2, 2), pty = "s")
plot(finalmodel4, cook.levels = c(4/42,0.5,1))
```

Plots for finalmodel3









With a small dataset of 42 observations, it can be a bit challenging to determine adherence to assumptions due to few data points. While a couple of observations come into question, we do not see large trends or patterns in the plots, and we don't find any remaining outliers with the finalmodel3. Because we don't find departure from the assumptions, we can conclude that linearity, constant variance, independence and normality appear to hold.

Grading: In an earlier version of the instructions, we asked students to ignore model linearity from the assumptions. Please do not take away points if student did not address model linearity in this section.

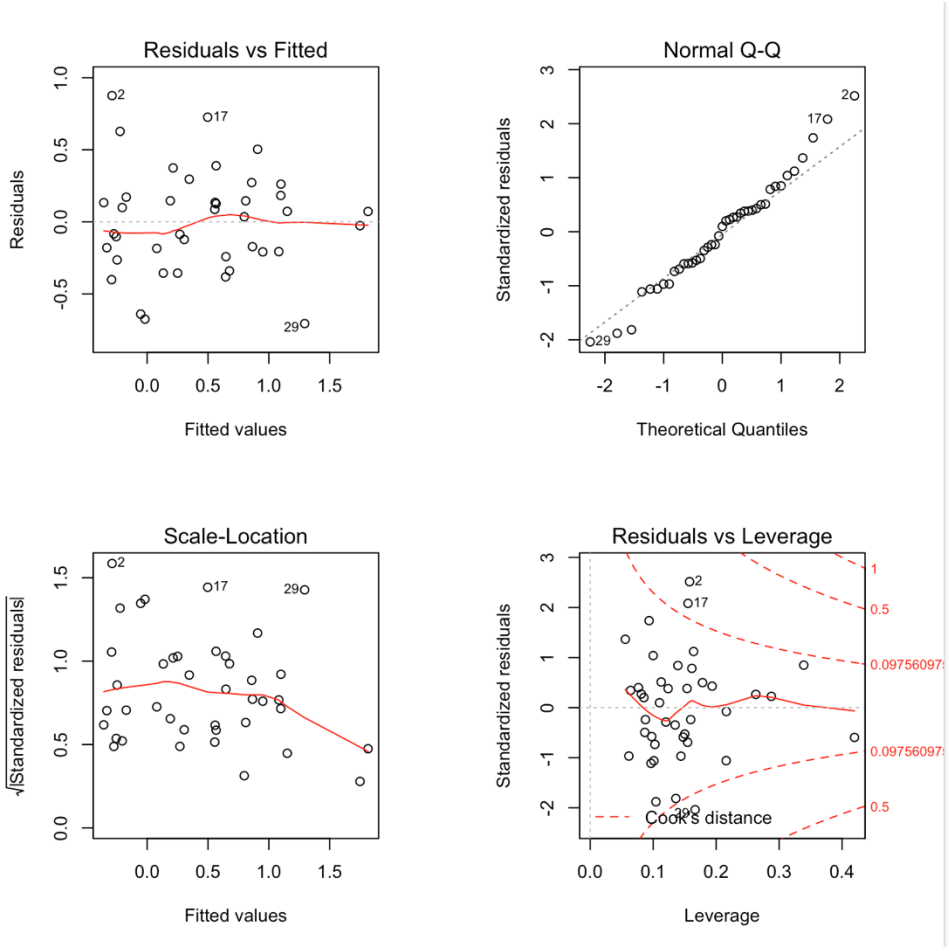
Question 6 - Checking the Assumptions of the Model (response is Dreaming)

- 6. Plot the relevant residual plots to check the final model assumptions. Individual predictors were assessed for linearity in Question 1 and Question 2. Model linearity should be assessed along with constant variance, independence and normality. 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.

R-Code

```
Finalmodel4 <- lm(log(Dreaming) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) + log(Gestation) + Danger, data = data)
summary(finalmodel4)
par(mfrow = c(2, 2), pty = "s")
plot(finalmodel4, cook.levels = c(4/41,0.5,1))
```

Plots for finalmodel4



With a small dataset of 41 observations, it can be a bit challenging to determine adherence to assumptions due to few data points. While a couple of observations come into question, we do not see large trends or patterns in the plots, and we don't find any remaining outliers with the finalmodel4. Because we don't find departure from the assumptions, we can conclude that linearity, constant variance, independence and normality appear to hold.

Grading: In an earlier version of the instructions, we asked students to ignore model linearity from the assumptions. Please do not take away points if student did not address model linearity in this section

---

END (Whew)!!!!