Question 1

Part 1: We want to fit a linear regression model that can be used to predict TotalSleep. Explain why Dreaming, NonDreaming and Species are BAD variables to include in this regression model.

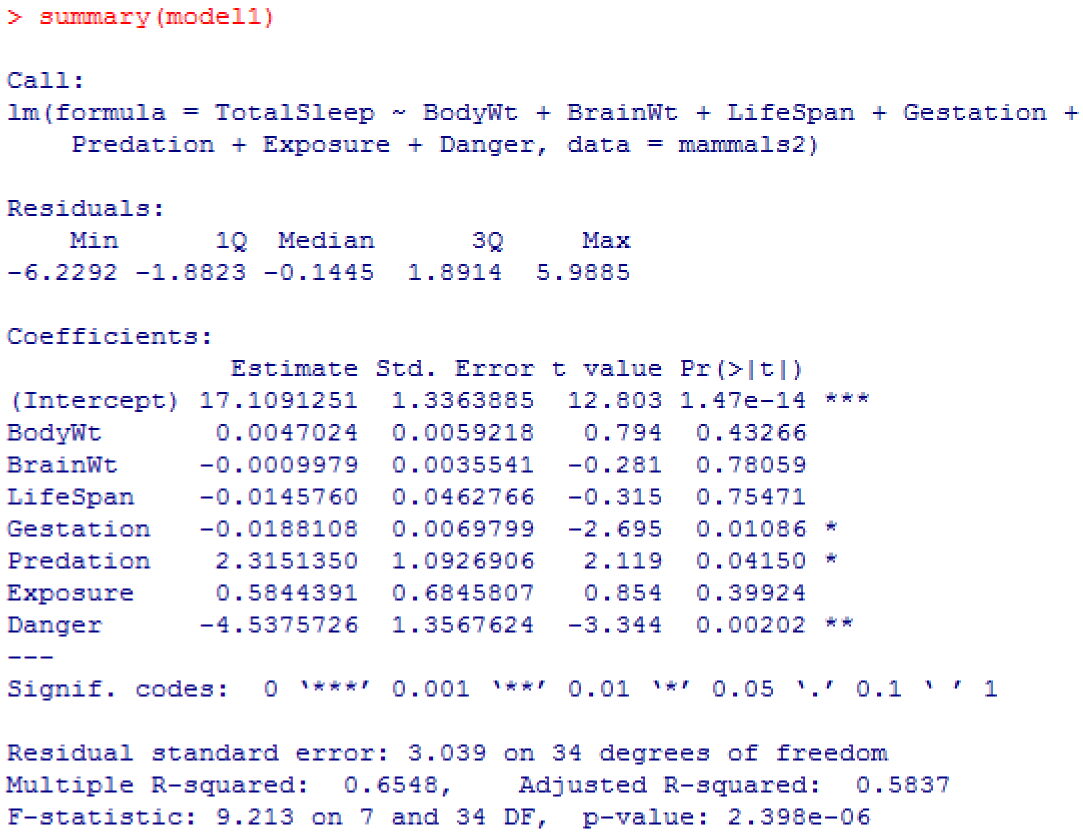
We want to predict TotalSleep which is a sum of Dreaming and NonDreaming i.e Total sleep already accounts for both, so it would be a bad choice for variables. Species consists the name of mammals (in character), which would be very difficult to predict just by names, so it would not be as helpful.

Part 2: Treat Predation, Exposure and Danger as numericals. Run model1, the linear regression model with TotalSleep vs BodyWt, BrainWt, LifeSpan, Gestation, Predation, Exposure and Danger. Clearly show the R command that you use, and include the R's model summary

R command used:

model1 <- lm(TotalSleep ~ BodyWt + BrainWt + LifeSpan + Gestation + Predation + Exposure + Danger,data=mammals)

R's model summary:



```
> summary(model1)

Call:
lm(formula = TotalSleep ~ BodyWt + BrainWt + LifeSpan + Gestation +
    Predation + Exposure + Danger, data = mammals2)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2292 -1.8823 -0.1445  1.8914  5.9885

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.1091251  1.3363885  12.803 1.47e-14 ***
BodyWt       0.0047024  0.0059218   0.794  0.43266
BrainWt     -0.0009979  0.0035541  -0.281  0.78059
LifeSpan    -0.0145760  0.0462766  -0.315  0.75471
Gestation   -0.0188108  0.0069799  -2.695  0.01086 *
Predation    2.3151350  1.0926906   2.119  0.04150 *
Exposure     0.5844391  0.6845807   0.854  0.39924
Danger      -4.5375726  1.3567624  -3.344  0.00202 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.039 on 34 degrees of freedom
Multiple R-squared:  0.6548,    Adjusted R-squared:  0.5837
F-statistic: 9.213 on 7 and 34 DF,  p-value: 2.398e-06
```

Part 3: Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients. Report the R2 and adjusted R2 of your model. What are the meaning of these values?

Y-hat (TotalSleep) = $17.10912 + 0.0047 * X_1 – 0.00099 * X_2 - 0.01457 * X_3 - 0.0188 * X_4 + 2.31513 * X_5$

$+ 0.58443 * X_6 - 4.5375726 * X_7$

Where X1 = BodyWt, X2 = BrainWt, X3 = LifeSpan, X4 = Gestation, X5 = Predation, X6 = Exposure, X7 = Danger

R^2 = 0.6548

Adjusted R^2 = 0.5837

R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is a less estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R^2.

The coefficients predict the amount of total sleep increase or decrease with respect to the intercept, and the p values are the probability of such event taking place.

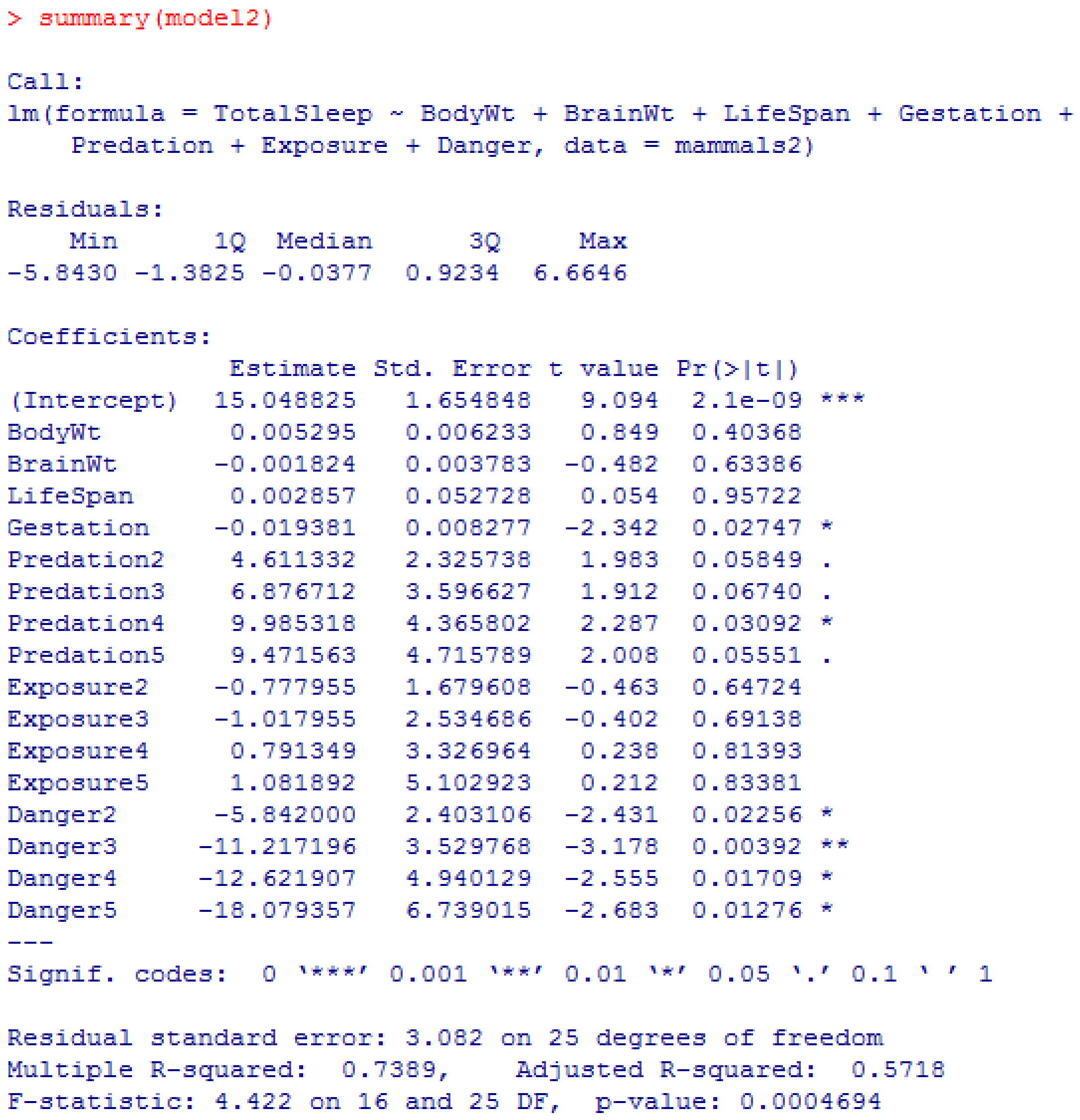
Question 2

Part 1: Treat Predation, Exposure and Danger as categoricals. Run model2, the linear regression model with TotalSleep vs BodyWt, BrainWt, LifeSpan, Gestation, Predation, Exposure and Danger. Clearly show the R command that you use, and include the R's model summary.

R command used:

```
model2 <- lm(TotalSleep ~  BodyWt + BrainWt + LifeSpan + Gestation + newPredation + newExposure + newDanger,data=mammals)
```

Model summary:

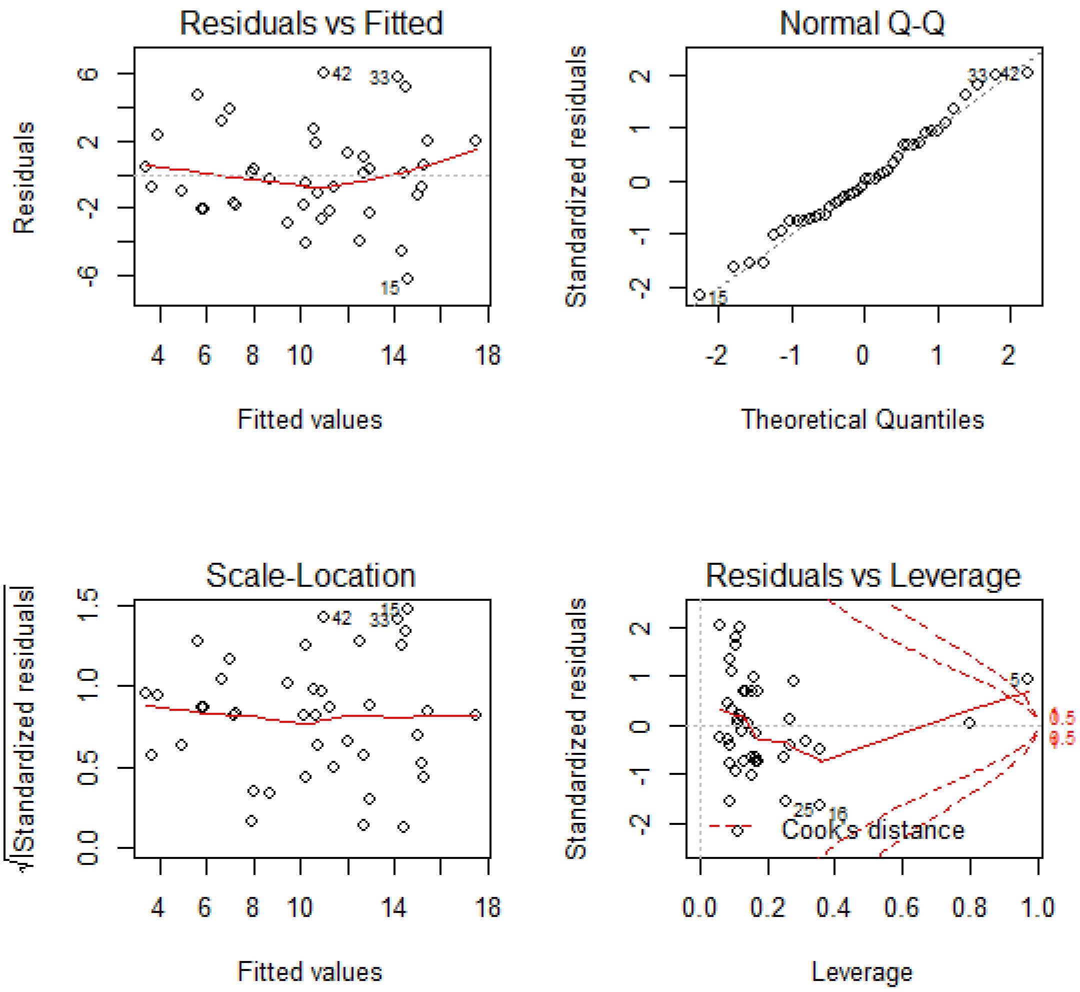


```
> summary(model2)

Call:
lm(formula = TotalSleep ~ BodyWt + BrainWt + LifeSpan + Gestation +
    Predation + Exposure + Danger, data = mammals2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8430 -1.3825 -0.0377  0.9234  6.6646

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.048825   1.654848   9.094  2.1e-09 ***
BodyWt        0.005295   0.006233   0.849  0.40368
BrainWt      -0.001824   0.003783  -0.482  0.63386
LifeSpan      0.002857   0.052728   0.054  0.95722
Gestation    -0.019381   0.008277  -2.342  0.02747 *
Predation2    4.611332   2.325738   1.983  0.05849 .
Predation3    6.876712   3.596627   1.912  0.06740 .
Predation4    9.985318   4.365802   2.287  0.03092 *
Predation5    9.471563   4.715789   2.008  0.05551 .
Exposure2    -0.777955   1.679608  -0.463  0.64724
Exposure3    -1.017955   2.534686  -0.402  0.69138
Exposure4     0.791349   3.326964   0.238  0.81393
Exposure5     1.081892   5.102923   0.212  0.83381
Danger2      -5.842000   2.403106  -2.431  0.02256 *
Danger3     -11.217196   3.529768  -3.178  0.00392 **
Danger4     -12.621907   4.940129  -2.555  0.01709 *
Danger5     -18.079357   6.739015  -2.683  0.01276 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.082 on 25 degrees of freedom
Multiple R-squared:  0.7389,    Adjusted R-squared:  0.5718
F-statistic: 4.422 on 16 and 25 DF,  p-value: 0.0004694
```

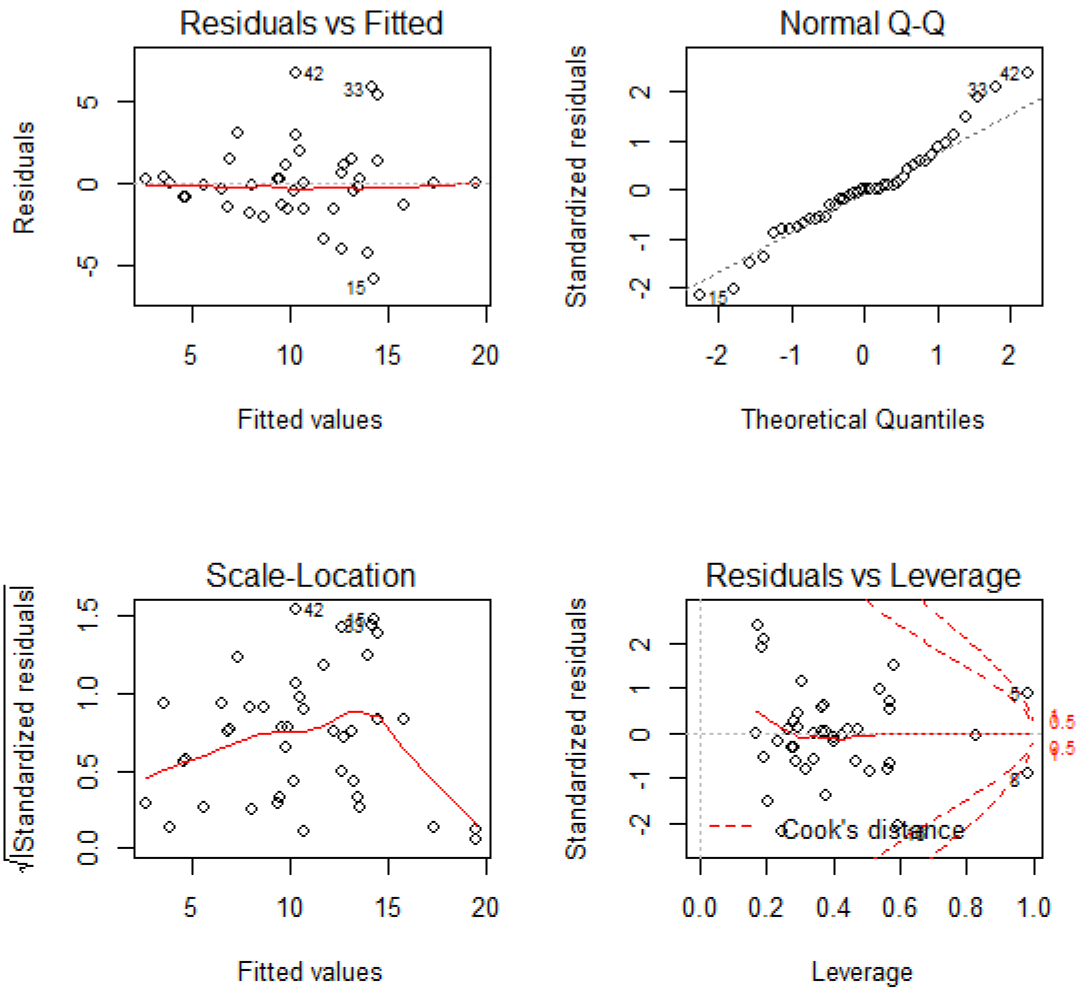
Part 2: Compare model1 and model 2: comment on the coefficients and the diagnostic plots. Say which, if any, of the (a) independence (b) normal distribution and (c) constant variance assumptions are violated.

Diagnostic plots

Model 1:


Residuals vs Fitted
Residuals
6
2
-2
-6
4
6
8
10
14
18
Fitted values
42
33
15
Normal Q-Q
Standardized residuals
2
1
0
-1
-2
Theoretical Quantiles
33
42
15
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
Fitted values
42
33
15
Residuals vs Leverage
Standardized residuals
0.0
0.2
0.4
0.6
0.8
1.0
Leverage
5
25
16
Cook's distance
0.5
0.5

Model 2:



Both R^2 and adjusted R^2 increased, but the adjusted R^2 is okay.

The intercept decreased on model 2, but that is expected with an increase in variables. Adjusted R^2 is fine on both models. In model 1, variables BrainWt, LifeSpan, BodyWt, and Exposure have high p-value. We should eliminate them with the highest p-value first (starting with BrainWt, does not have to be all) until we get a better model. In model 2, LifeSpan, all Exposure (1,2,3,4), BrainWt, and BodyWt have high p value. Similarly like in model1, we should eliminate them with the highest p-value first, starting with LifeSpan to get a better model.

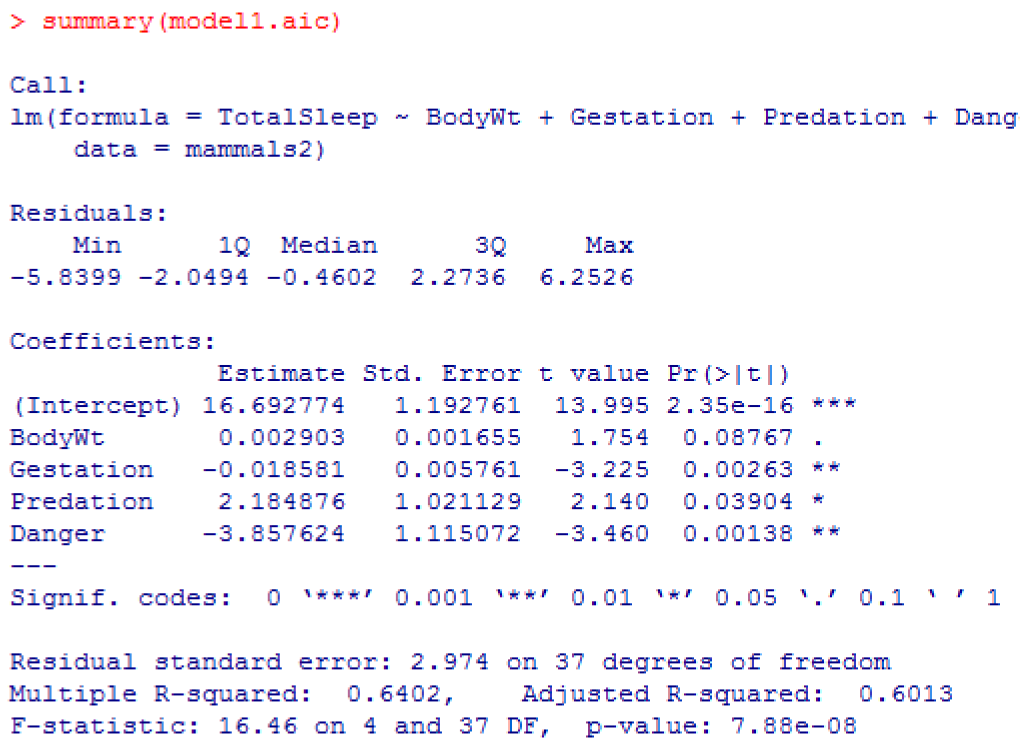
In model 1, the curves on the scale-location, and residual vs fitted are a little bit curvy, but it is acceptable i.e variance and independence are not violated. Normal QQ is good, so normal distribution is not violated.

In model 2, the curves on the scale-location is bent and curvy in the end. So, variance is violated. Residual vs fitted is good i.e independence is not violated. Normal QQ graph has many outliers towards the end, but it is acceptable. Hence, normal distribution is not violated.

Question 3

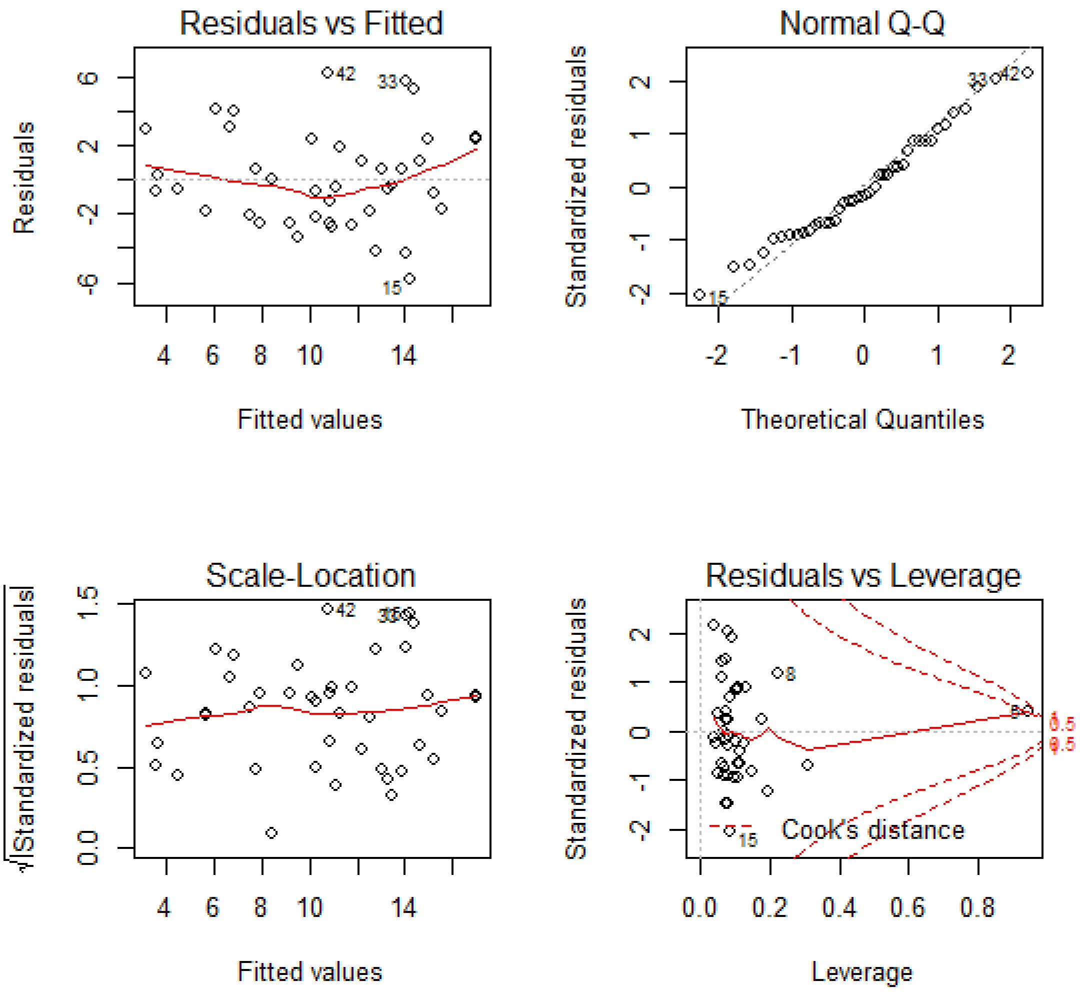
Part 1: Do variable selection with the stepAIC command, starting with model1. Call this model1.AIC. Compare model1.AIC against model1: comment on the coefficients and the diagnostic plots.

model1.aic summary:

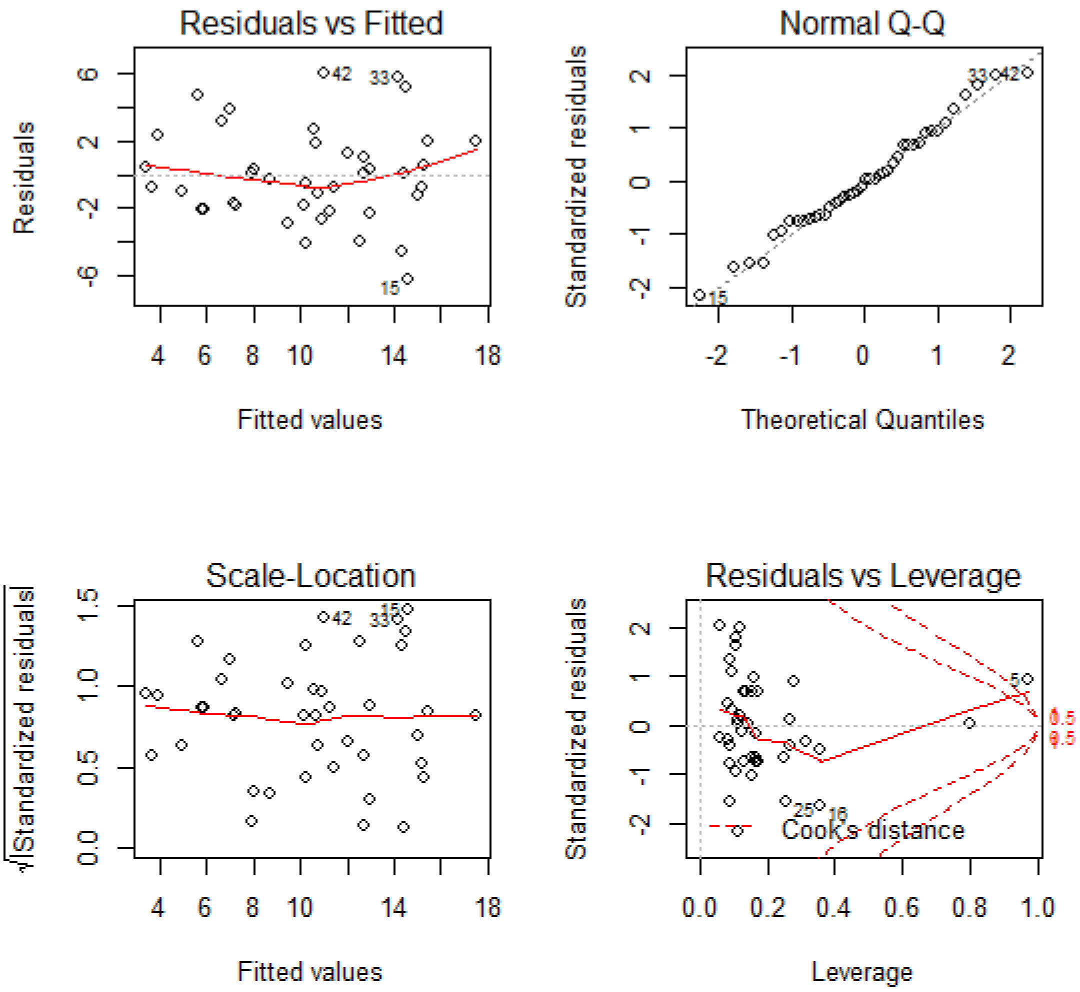


```
> summary(model1.aic)

Call:
lm(formula = TotalSleep ~ BodyWt + Gestation + Predation + Danger,
    data = mammals2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8399 -2.0494 -0.4602  2.2736  6.2526 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.692774   1.192761  13.995 2.35e-16 ***
BodyWt       0.002903   0.001655   1.754  0.08767 .  
Gestation   -0.018581   0.005761  -3.225  0.00263 ** 
Predation    2.184876   1.021129   2.140  0.03904 *  
Danger      -3.857624   1.115072  -3.460  0.00138 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.974 on 37 degrees of freedom
Multiple R-squared:  0.6402,    Adjusted R-squared:  0.6013 
F-statistic: 16.46 on 4 and 37 DF,  p-value: 7.88e-08
```

Diagnostic plot of model1.aic:

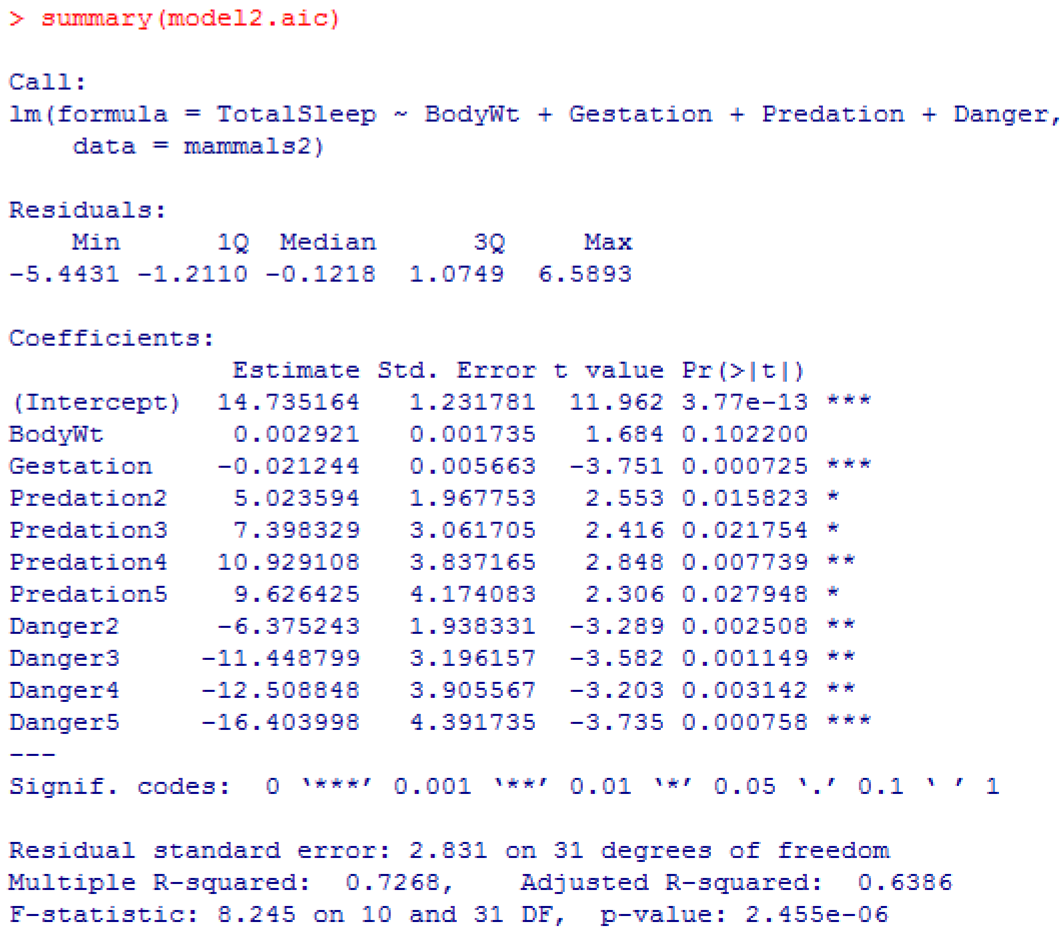

Residuals vs Fitted
Residuals
6
2
-2
-6
4
6
8
10
14
Fitted values
42
33
15
Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles
33
42
15
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
4
6
8
10
14
Fitted values
42
33
15
Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
Leverage
8
15
5
Cook's distance
1
0.5
0.5
1

For model1:



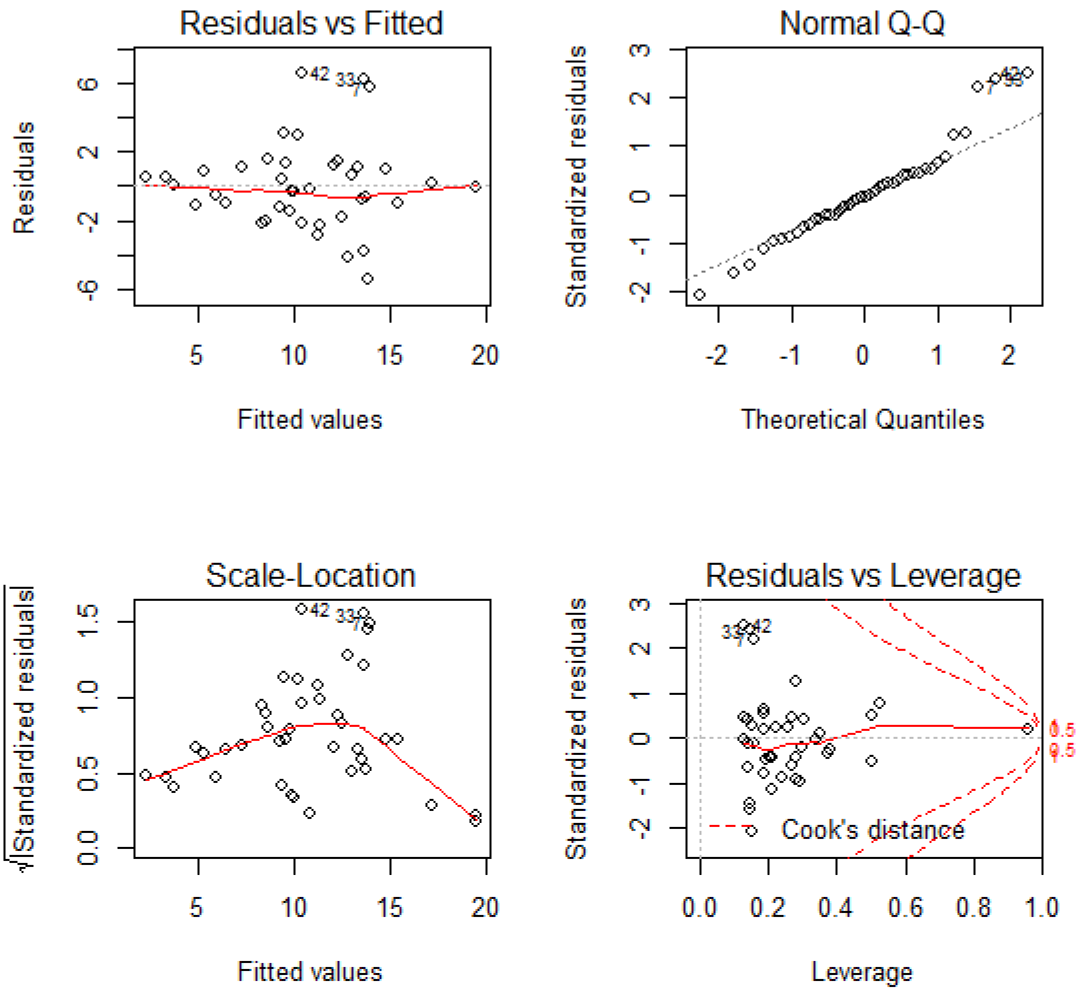
Adjusted R^2, is higher for model1.aic, which is better, and model1.aic has less variables which is also better. The AIC removed BrainWt, LifeSpan, Gestation, and Exposure variables. However, there is not a significant improvement in variance, independence or normality, but regardless there is no violation in any three on both models.

Part 2: Do variable selection with the stepAIC command, starting with model2. Call this model2.AIC. Compare model2.AIC against model2: comment on the coefficients



```
> summary(model2.aic)

Call:
lm(formula = TotalSleep ~ BodyWt + Gestation + Predation + Danger,
    data = mammals2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4431 -1.2110 -0.1218  1.0749  6.5893 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  14.735164   1.231781  11.962 3.77e-13 ***
BodyWt        0.002921   0.001735   1.684 0.102200    
Gestation    -0.021244   0.005663  -3.751 0.000725 ***
Predation2    5.023594   1.967753   2.553 0.015823 *  
Predation3    7.398329   3.061705   2.416 0.021754 *  
Predation4   10.929108   3.837165   2.848 0.007739 ** 
Predation5    9.626425   4.174083   2.306 0.027948 *  
Danger2      -6.375243   1.938331  -3.289 0.002508 ** 
Danger3     -11.448799   3.196157  -3.582 0.001149 ** 
Danger4     -12.508848   3.905567  -3.203 0.003142 ** 
Danger5     -16.403998   4.391735  -3.735 0.000758 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.831 on 31 degrees of freedom
Multiple R-squared:  0.7268,	Adjusted R-squared:  0.6386 
F-statistic: 8.245 on 10 and 31 DF,  p-value: 2.455e-06
```

Diagnostic plot of model2.aic:



The adjusted R^2 increased on model2.aic. The AIC eliminated BrainWt, LifeSpan, Exposure and Danger variables which is also what we wanted to do above with model2. Since, model2.aic has lesser number of variables and similar variance it is a better model than model2.

Part 3: Which model amongst the above 4 is the best? (Give a brief justification). For the better model, summarize the relationship between TotalSleep and other attributes of a mammal.

Comparatively speaking, model1.aic is the best among 4 as no serious violation occurred on the diagnostic plots of model1.aic, adjusted R^2 is larger (better), and has lesser variables compared to model 1. Even though the adjusted R^2 is larger in model2.aic, it has lesser variables than both model2.aic. Both model2 and model2.aic had violation in variance in their diagnostic plot, so between model1 and model1.aic, the later(model1.aic) is the better model.

According to model1.aic the TotalSleep a mamal gets is dependent upon all given variables: BodyWt, Gestation, Predation, Danger.

Part 4: The species Homo Sapiens has the following attributes: BodyWt = 75, BrainWt = 1.4, LifeSpan = 77, Gestation = 268, Predation = 2, Exposure = 2, Danger = 2. Use your model to predict TotalSleep for this species. Is your prediction reasonable? Explain why or why not.

Y-hat (TotalSleep) = 16.69274 + 0.002903 * X_1 – 0.018581 * X_2 + 2.184876 * X_3 – 3.857624 * X_4

TotalSleep = 8.58 hours

My (computer, AIC) prediction is reasonable because there are not many variables compared to the first two models, and diagnostic plots are okay. The results for Homo Sapiens also is reasonable because that is close to the recommended amount of sleep for humans.