Mammalian Sleep and Longevity: An Analysis of the mammalsleep Data Set

Mariska Batavia, Jack Donohue, and Rowdy Dudley

Executive Summary

Daily sleep duration and longevity are important aspects of mammalian biology, and are factors that affect, and are affected by, many other physiological, life history, and ecological traits. Both daily sleep duration and longevity may be hard to accurately measure in practice, due to the logistical constraints inherent in capturing, housing, and/or monitoring rare or non-domesticated species. In this project we sought to model these traits as a function of other, potentially more easily measurable features of species.

Our data for this project came from the mammalsleep data set from the Faraway package in R. This dataset contains data on 62 species of mammals, and includes measures of brain and body mass, lifespan, gestation, daily sleep duration, and three categorical indicies (predation, exposure, and danger indices) that capture various aspects of a species' behavioral ecology. We sought to build models that could predict daily sleep duration and odds of being longliving as accurately as possible, and also wanted to understand the relationship between these outcomes and various predictor variables.

Over the course of the project, we built three models. The first model used multiple linear regression to model daily sleep duration, treating the three categorical indices as additional quantitative variables in the model, rather than as categories. The second model used multiple linear regression to model daily sleep duration, but unlike the first approach, treated the three categorical indices as categorical variables. This second approach generated a better model than the first approach. Our third model used logistic regression to model the odds of being longliving (defined as the top 25% of lifespans).

Data Description and Methods

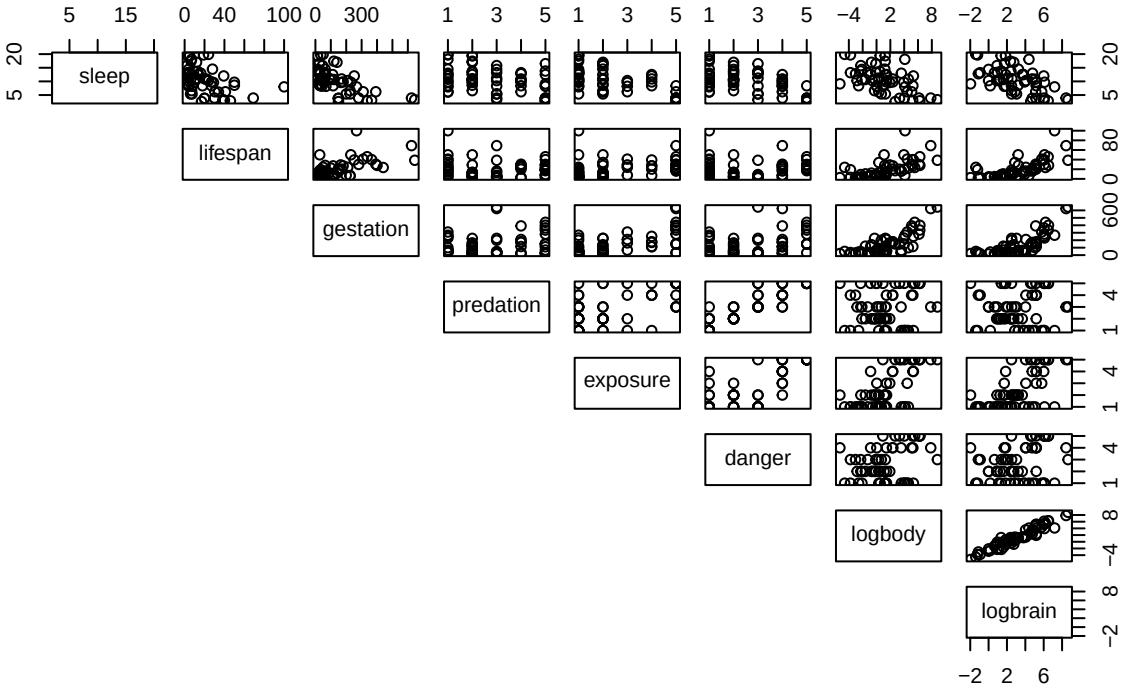
The data source for this project was the mammalsleep dataset from the Faraway package in R. This set included data from 62 species of mammals. Several quantitative variables were available, including brain mass (g), body mass (kg), daily sleep duration (hours), length of gestation (days), and lifespan (years). The dataset also included three categorical variables - each a five-point scale - that characterized species' risk of predation (predation index), exposure during sleep periods (exposure index), and overall level of danger from other animals (danger index).

We performed several manipulations on our data. First, both brain and body mass show a very wide range of values, with the majority of species clustered at the low end of the range. To pull in the extreme values, we added log(body mass) and log(brain mass) to the dataset, and used these in some analyses. Second, all three indices (predation, exposure, and danger), were initially coded as integers; during model building we experimented with these variables by either treating them as quantitative or by converting them to factors and treating them as categorical. Third, the original dataset included lifespan as a quantitative variable; to make this a binary outcome suitable for a logistic regression analysis, we categorized species with lifespan in the 4th quartile (>28 years) as long-living, and species with lifespans shorter than this threshold as not long-living. Finally, we removed points with missing values as needed to perform analyses.

Exploratory Data Analysis

To explore the relationship between daily sleep duration and candidate predictors, we used the pairs function to plot a matrix of scatterplots (Figure 1, below). Note that we used log of body mass and brain mass, as described above, due to the very wide range of values for mass and the consequent difficulty in visualizing relationships to other variables. We also included the predation, exposure, and danger indices in this matrix; even though these variables are fundamentally categorical, we wanted to explore the possibility of treating them as quantitative, since they are each on a scale of 1-5.

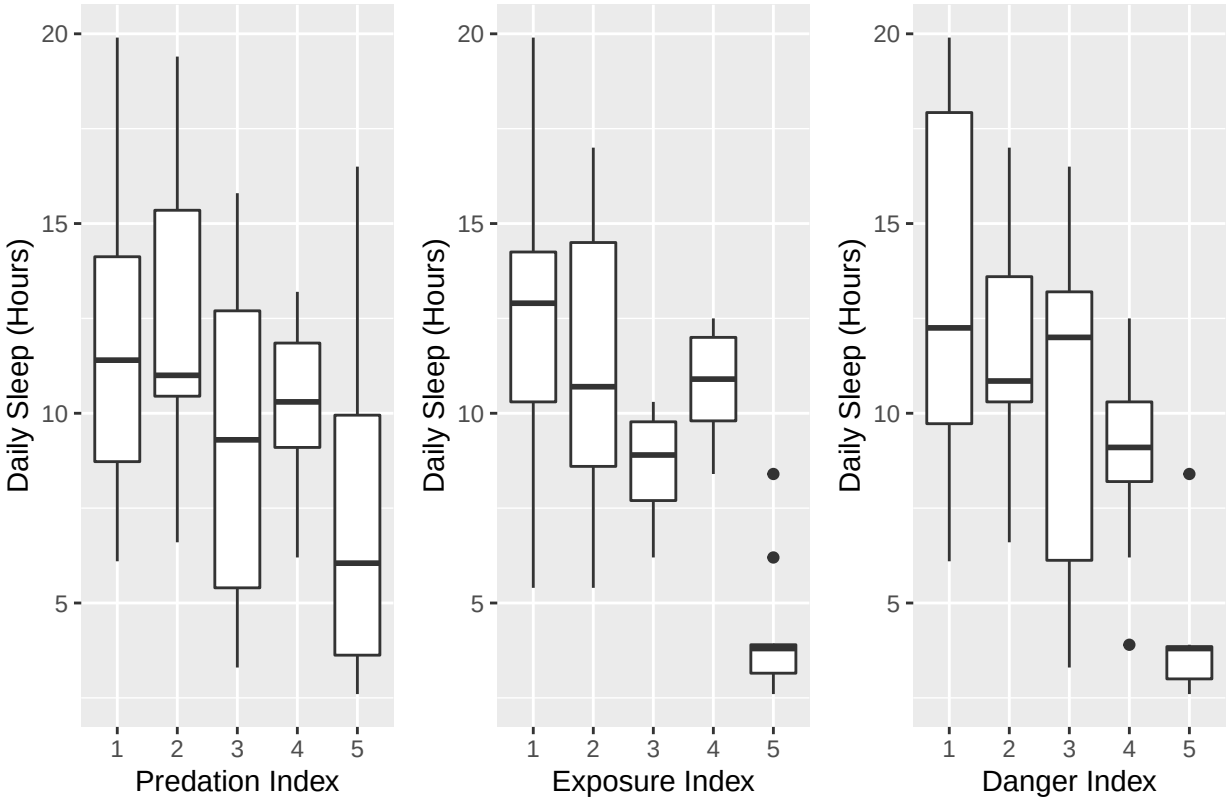
Figure 1: Pairs Plot of Potential Predictors



In addition to showing that sleep is related to all the quantitative predictors, this matrix shows that many of the variables are positively correlated to each other, and we took this into account when we built our models.

We also explored how the three categorical variables were related to daily sleep duration (Figure 2, below).

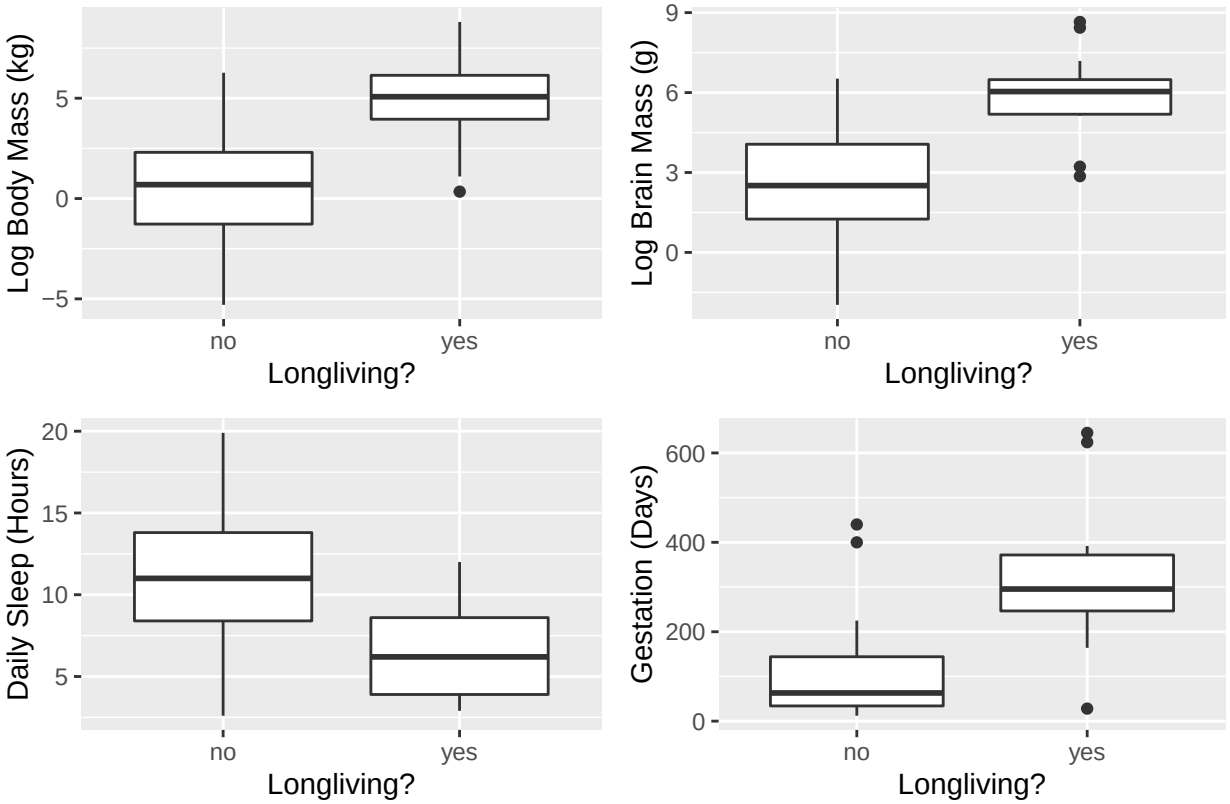
Figure 2: Categorical Predictors Versus Hours of Daily Sleep



Vulnerability to predation, exposure during sleep, and overall danger from other animals seemed like variables that might logically have some relationship to sleeping patterns. From an initial analysis we can see that, despite considerable overlap between levels, the animals that get the most sleep per day generally have lower predation and exposure indices. Animals with the lowest danger index level tend to sleep more hours. This makes sense, as the safest animals will naturally be able to sleep longer hours.

To explore longevity (likelihood of being in the top quartile for lifespan) versus potential quantitative predictors, we created boxplots to examine the distribution of predictor values between long-living species and those that are not long-living (Figure 3, below).

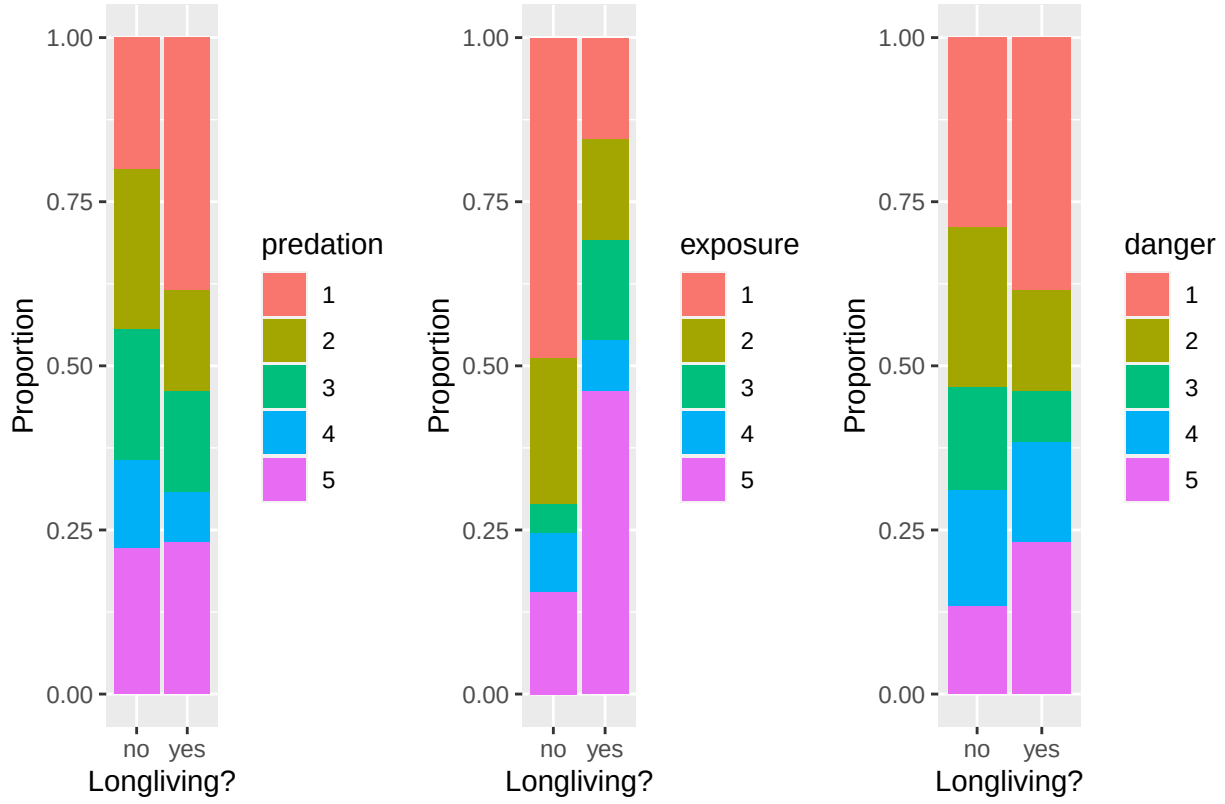
Figure 3: Distribution of Quantitative Predictors by Longevity



From this figure, we can see that long-living species have higher log body and log brain mass, shorter daily sleep duration, and longer gestation periods. Since we knew from Figure 1 that many of these predictors are correlated to each other, we took this into account when building models of longevity.

Finally, to explore longevity (likelihood of being in the top quartile of lifespan) versus potential categorical predictors, we created stacked bar charts to compare proportions of each index level for long-living versus non-long-living species (Figure 4, below).

Figure 4: Distribution of Index Levels by Longevity



On the danger index, we can see that long-living species are more likely to be in extreme levels (1=least danger, 5=most danger) than non-long-living species. Exposure index is markedly different between long-living and non-long-living species, with long-living species proportionally much more exposed; we suspect this may be partially attributable to the generally larger body mass of long-living species, and the relative scarcity of well protected sleeping places for large animals. Lastly, long-living species are at lower risk of predation as compared to their non-long-living counterparts, though most of the differences between long-living species and non-long-living-species were in low risk predation levels. As such, we considered predation to be less robustly related to longevity than the other two indicies.

Model 1: Daily Sleep Duration in Mammals with Quantitative Predictors

Our first objective was to identify a model that predicts how much a mammal sleeps based on multiple predictors, requiring us to perform a multiple linear regression. Our data involved both quantitative and categorical predictors, so we decided to run multiple models to determine if it was better to treat the categorical predictors as categorical, or if it worked better to treat them as quantitative. Initially we decided to try a model with the categorical predictors treated as quantitative variables.

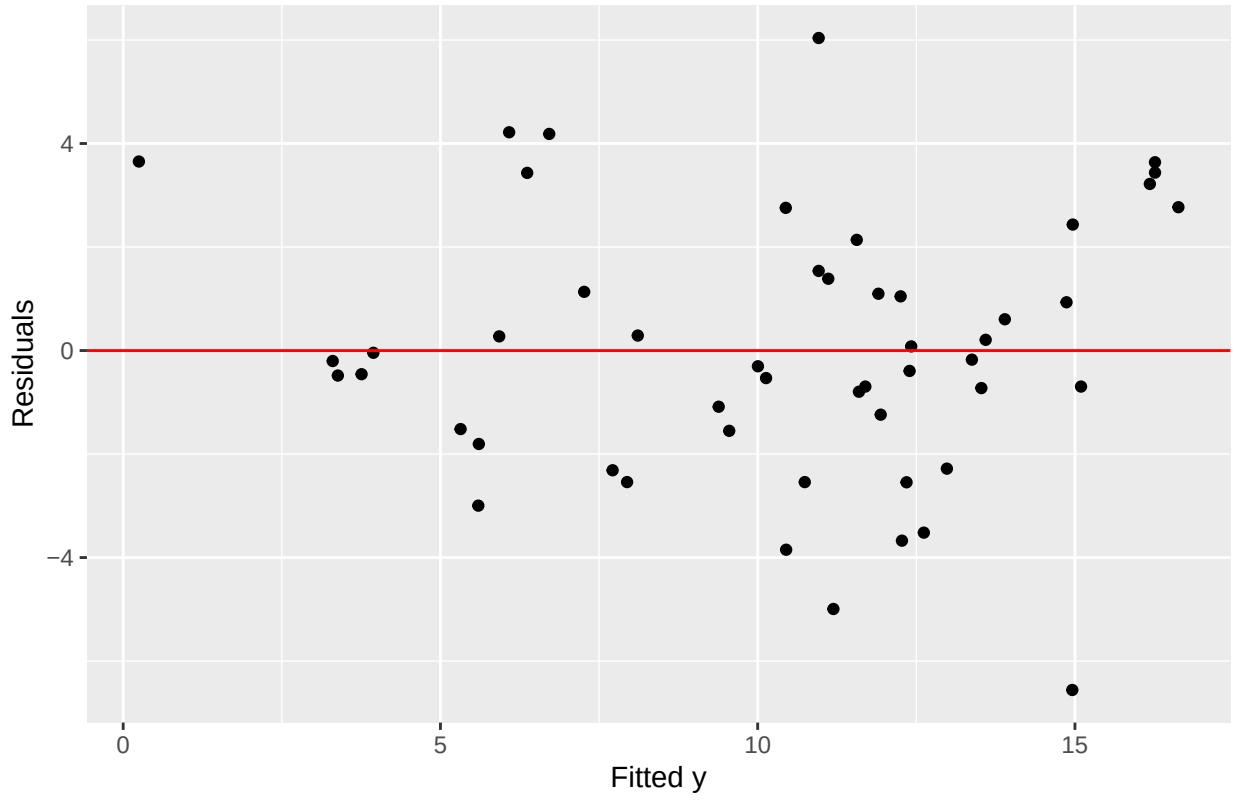
```
##
## Attaching package: 'MASS'

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

##
## Call:
## lm(formula = sleep ~ LogBody + lifespan + gestation + danger +
##     LogBrain + predation + exposure, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5584 -1.5343 -0.2021  1.4630  6.0387
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.148539   1.562319  10.976 4.74e-14 ***
## LogBody      0.128558   0.459446   0.280 0.780964
## lifespan     0.014995   0.032532   0.461 0.647175
## gestation   -0.008116   0.004732  -1.715 0.093559 .
## danger      -4.193364   1.120484  -3.742 0.000536 ***
## LogBrain    -0.841289   0.629720  -1.336 0.188586
## predation    1.949940   0.946696   2.060 0.045510 *
## exposure     0.828047   0.560568   1.477 0.146921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.765 on 43 degrees of freedom
## Multiple R-squared:  0.6997, Adjusted R-squared:  0.6508
## F-statistic: 14.31 on 7 and 43 DF,  p-value: 1.976e-09
```

Figure 5: Residual Plot



On an initial look at the pairs function in (figure above) we can see that the categorical predictors do not take on a linear shape and look atypical when compared to the other quantitative predictors, indicating that treating them as categorical predictors is likely the better option for our model.

The initial residual plot of the model indicated that assumption one was likely met, with what looks to be an

even distribution of points on either side of the line, but assumption two looks not to be met, with large differences in the variation of space between points.

Based on the summary of the model many of the predictors look to be insignificant based on the p-values. This indicated that we should run a partial F test on those predictors with high p-values to determine if we can drop them all from the model. This would hopefully improve our model's predictive ability.

```
## Analysis of Variance Table
##
## Model 1: sleep ~ danger + gestation + predation
## Model 2: sleep ~ LogBody + lifespan + gestation + danger + LogBrain +
##     predation + exposure
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     47 373.65
## 2     43 328.84  4    44.815 1.465 0.2295
```

```
## Analysis of Variance Table
##
## Response: sleep
##           Df Sum Sq Mean Sq F value    Pr(>F)
## LogBody    1 402.97  402.97 52.6934 5.436e-09 ***
## lifespan   1   0.01    0.01  0.0018  0.965959
## gestation  1  77.65   77.65 10.1544  0.002681 **
## danger     1 224.38  224.38 29.3406 2.560e-06 ***
## LogBrain   1  15.90   15.90  2.0792  0.156566
## predation  1  28.43   28.43  3.7175  0.060465 .
## exposure   1  16.69   16.69  2.1820  0.146921
## Residuals 43 328.84    7.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 14.40395
```

```
## [1] 1.231251e-06
```

```
## [1] 2.821628
```

Our partial F test (see equations above) indicated that we could use the reduced model, with only danger, gestation, and predation as predictors. To be sure about which predictors we should keep and which we should drop we performed a forward selection to determine which predictors would result in the lowest AIC value for our model, hopefully indicating which predictors we can drop from the model.

```
## Start:  AIC=158.39
## sleep ~ 1
##
##             Df Sum of Sq     RSS    AIC
## + exposure   1    489.24  605.62 130.20
## + danger     1    442.50  652.37 133.99
## + gestation  1    433.25  661.61 134.71
## + LogBrain   1    416.10  678.77 136.01
## + LogBody    1    402.97  691.90 136.99
## + predation  1    232.05  862.82 148.25
## + lifespan   1    171.92  922.95 151.68
## <none>                   1094.87 158.40
##
## Step:  AIC=130.2
## sleep ~ exposure
##
```

```
##             Df Sum of Sq    RSS    AIC
## + gestation  1    80.043 525.58 124.97
## + LogBrain   1    76.909 528.71 125.27
## + LogBody    1    62.682 542.94 126.62
## + danger     1    39.928 565.70 128.72
## + lifespan   1    30.527 575.10 129.56
## <none>                   605.62 130.20
## + predation  1     3.111 602.51 131.93
##
## Step:  AIC=124.97
## sleep ~ exposure + gestation
##
##             Df Sum of Sq    RSS    AIC
## + danger     1    98.720 426.86 116.36
## + predation  1    35.904 489.68 123.36
## <none>                   525.58 124.97
## + LogBrain   1    15.995 509.59 125.39
## + LogBody    1     9.499 516.08 126.04
## + lifespan   1     0.306 525.27 126.94
##
## Step:  AIC=116.36
## sleep ~ exposure + gestation + danger
##
##             Df Sum of Sq    RSS    AIC
## + LogBrain   1    64.737 362.12 109.97
## + predation  1    56.814 370.05 111.07
## + LogBody    1    45.579 381.28 112.60
## <none>                   426.86 116.36
## + lifespan   1    11.062 415.80 117.02
##
## Step:  AIC=109.97
## sleep ~ exposure + gestation + danger + LogBrain
##
##             Df Sum of Sq    RSS    AIC
## + predation  1    31.440 330.68 107.34
## <none>                   362.12 109.97
## + LogBody    1     0.435 361.69 111.91
## + lifespan   1     0.204 361.92 111.94
##
## Step:  AIC=107.34
## sleep ~ exposure + gestation + danger + LogBrain + predation
##
##            Df Sum of Sq    RSS    AIC
## <none>                  330.68 107.34
## + lifespan  1   1.24744 329.44 109.14
## + LogBody   1   0.22148 330.46 109.30
##
## Call:
## lm(formula = sleep ~ exposure + gestation + danger + LogBrain +
##     predation, data = data)
##
## Coefficients:
## (Intercept)     exposure    gestation       danger     LogBrain    predation
```

```
##   16.939915     0.835128    -0.007528    -4.176352    -0.634101     1.906272
```

```
##
## Call:
## lm(formula = sleep ~ exposure + gestation + danger + predation +
##     LogBrain, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6732 -1.4997 -0.1465  1.3862  5.7219
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.939915   0.993241  17.055  < 2e-16 ***
## exposure     0.835128   0.544191   1.535 0.131878
## gestation   -0.007528   0.004488  -1.678 0.100374
## danger      -4.176352   1.095886  -3.811 0.000418 ***
## predation    1.906272   0.921601   2.068 0.044374 *
## LogBrain    -0.634101   0.273975  -2.314 0.025264 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.711 on 45 degrees of freedom
## Multiple R-squared:  0.698,  Adjusted R-squared:  0.6644
## F-statistic:  20.8 on 5 and 45 DF,  p-value: 1.044e-10
```

The forward selection (see summary above) indicated that we should only include 5 of our 7 predictors. With this information we then conducted a VIF test to determine which of the predictors the forward selection chose could be dropped due to multicolinearity among the remaining predictors.

```
##              LogBody   LogBrain    lifespan gestation     danger   predation
## LogBody   1.00000000 0.95874091  0.64867546 0.7651316 0.25773981  0.07299113
## LogBrain  0.95874091 1.00000000  0.72671913 0.7733380 0.22583881  0.03300267
## lifespan  0.64867546 0.72671913  1.00000000 0.6374717 0.03804264 -0.12554213
## gestation 0.76513158 0.77333800  0.63747172 1.0000000 0.31119646  0.14079500
## danger    0.25773981 0.22583881  0.03804264 0.3111965 1.00000000  0.94143392
## predation 0.07299113 0.03300267 -0.12554213 0.1407950 0.94143392  1.00000000
## exposure  0.62940686 0.60719483  0.35972080 0.6254329 0.76810506  0.62654891
##            exposure
## LogBody   0.6294069
## LogBrain  0.6071948
## lifespan  0.3597208
## gestation 0.6254329
## danger    0.7681051
## predation 0.6265489
## exposure  1.0000000
```

```
##  exposure gestation    danger predation  LogBrain
##  5.022446  2.781078 16.656999 12.800012  3.263859
```

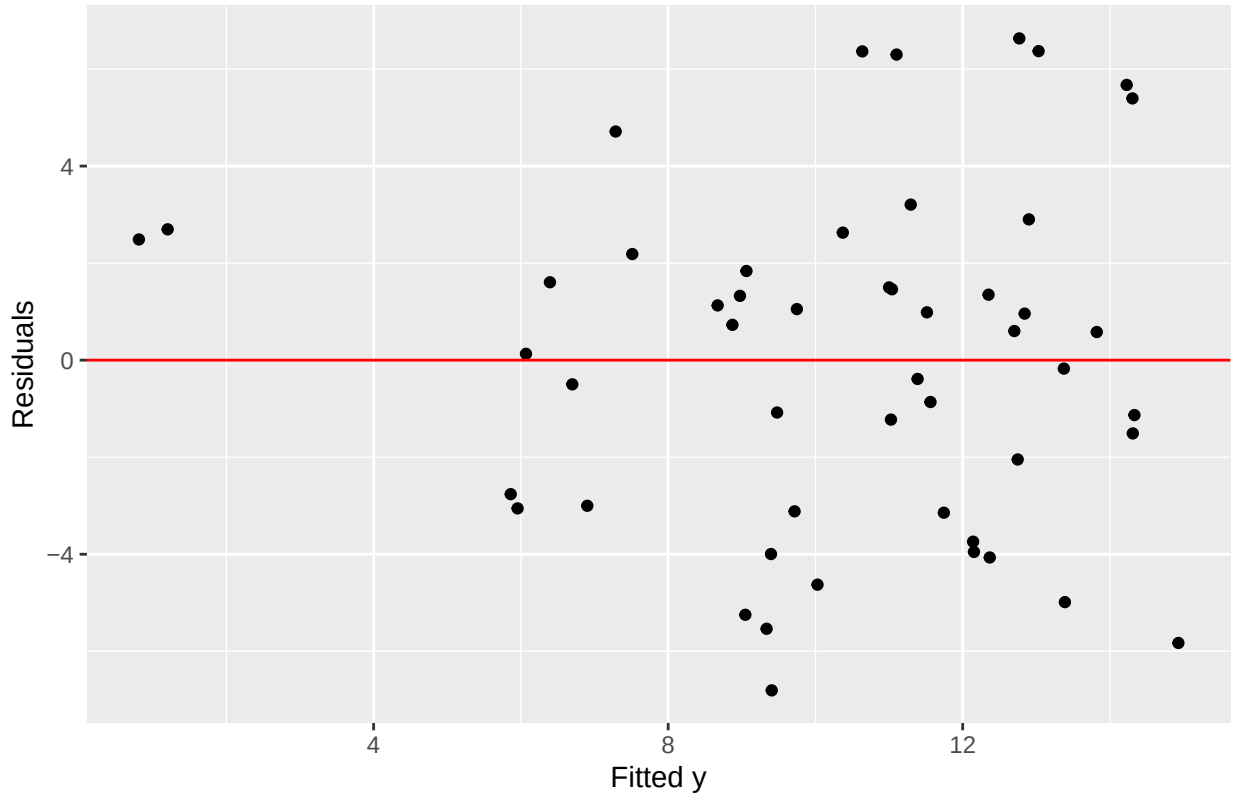
We found that 3 of our 5 predictors had a high VIF score (see output above), meaning they were found to be colinear, allowing us to drop them from the model as well, leaving us with 2 predictors for our final model.

```
##
## Call:
## lm(formula = sleep ~ gestation + LogBrain, data = data)
##
```

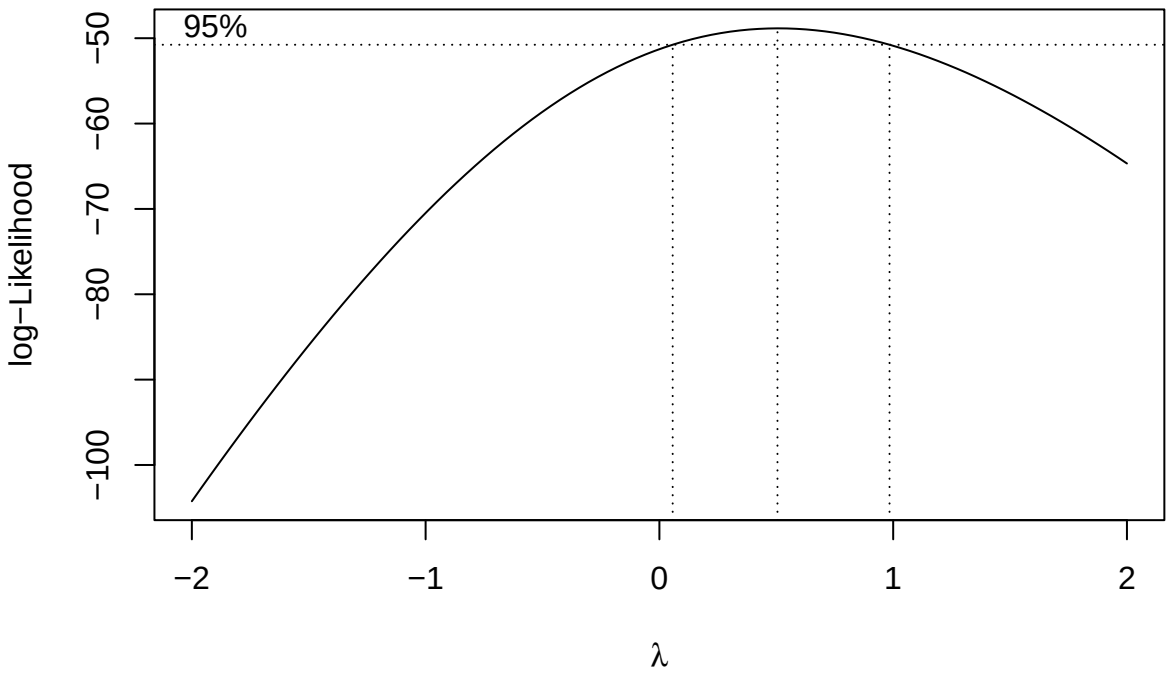
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8064 -3.0284  0.5794  2.0121  6.6316
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.019746   0.810619  17.295   <2e-16 ***
## gestation   -0.012447   0.005607  -2.220   0.0312 *
## LogBrain    -0.598704   0.315995  -1.895   0.0642 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.581 on 48 degrees of freedom
## Multiple R-squared:  0.4378, Adjusted R-squared:  0.4143
## F-statistic: 18.69 on 2 and 48 DF,  p-value: 9.956e-07
```

After viewing the summary of our newest model this left us with 2 predictors of significance and an R2 value of .44. We then ran our residual plot (see chart below) again to determine if our regression assumptions had been met.

Figure 6: Residual Plot



This plot looks better, but the variance among points still seems to be off. This drove us to then transform the y variable in order to constrain the variance among points. Initially we produced an boxcox plot (see chart below) to determine the best adjustment to make in terms of transformation of the y variable.



Since the value 0 was not included in our range within the boxcox plot (see chart above) we determined the best transformation procedure would be to raise the y to 0.5 and apply that to our predictors and see if it impacts our plot.

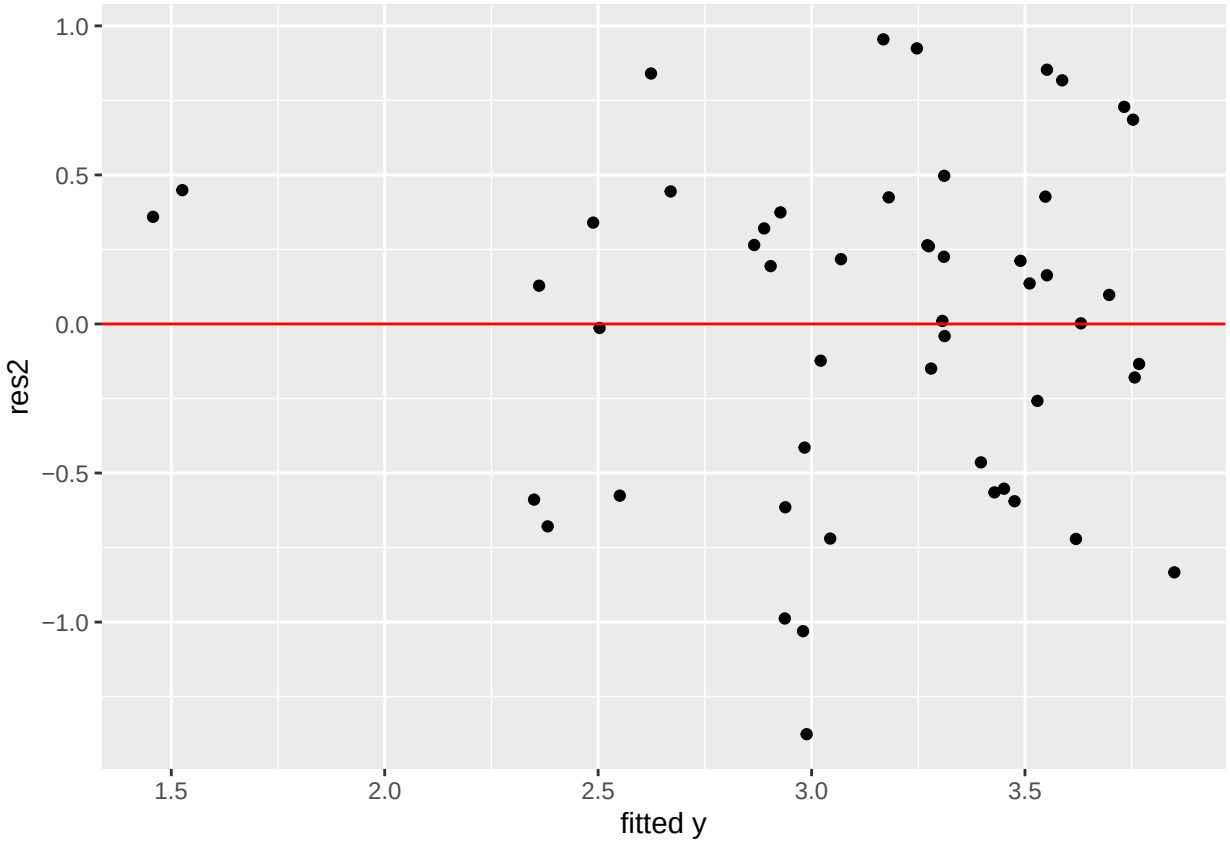
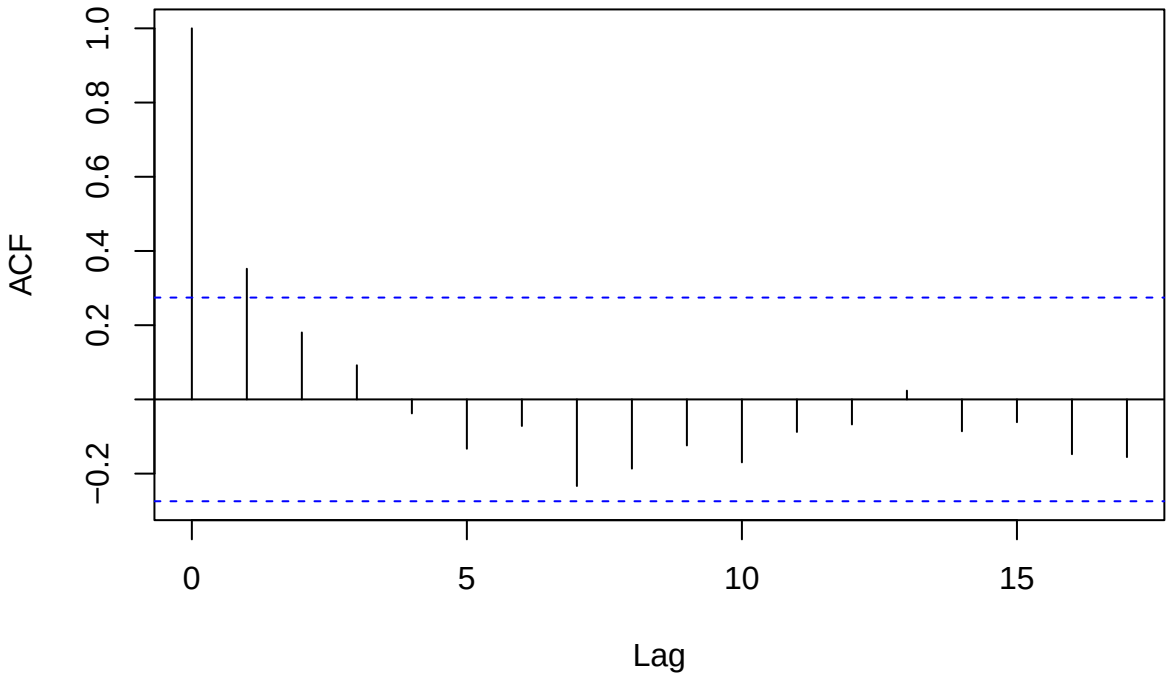
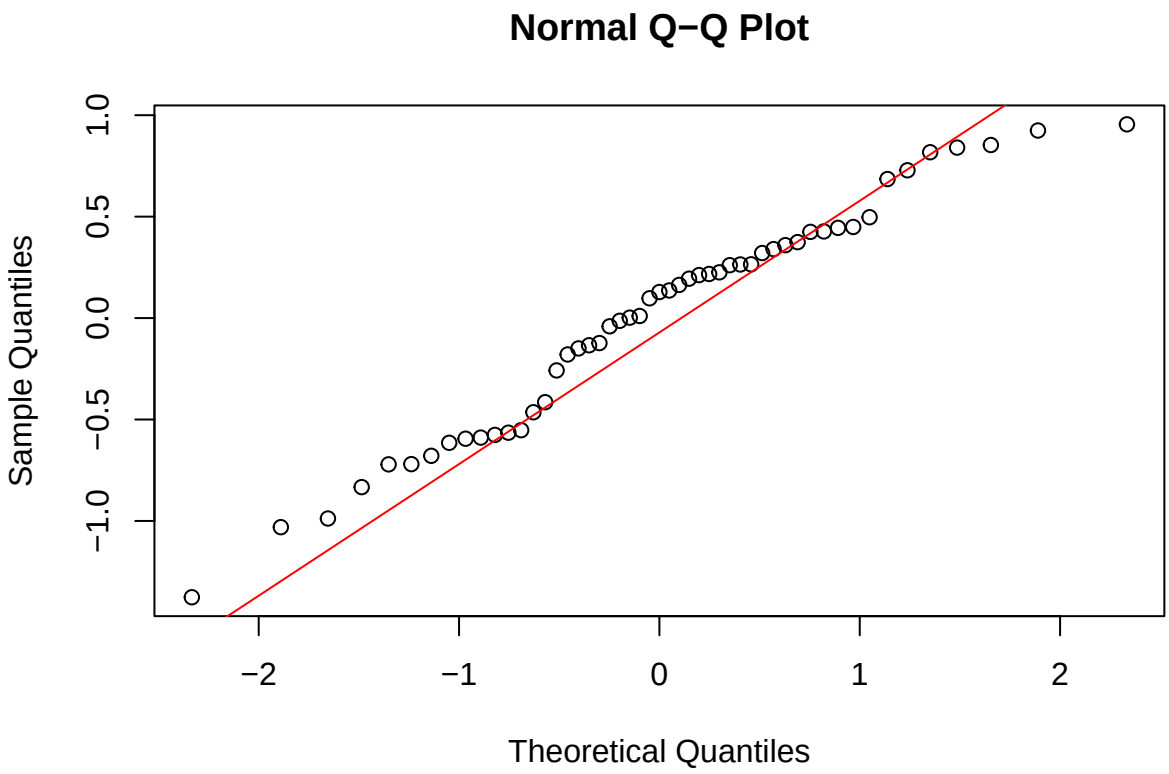

1.0
0.5
0.0
−0.5
−1.0
res2
1.5
2.0
2.5
3.0
3.5
fitted y

Figure 7: ACF Plot of Residuals with ystar





The transformation of y had little impact on the variance issues (see chart above) we were seeing in our previous residual plot, but improved the R^2 value of the model to .57. We then performed an ACF plot (see ACF chart above) and a QQ plot (see QQ plot above) to determine if our model was within tolerance.

```
summary(result.ystar)
```

```
##
## Call:
## lm(formula = ystar ~ gestation + LogBrain, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3760 -0.5084  0.1283  0.3669  0.9549
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7356153  0.1287495  29.015  < 2e-16 ***
## gestation   -0.0024010  0.0008906  -2.696  0.00965 **
## LogBrain    -0.0843632  0.0501890  -1.681  0.09928 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5688 on 48 degrees of freedom
## Multiple R-squared:  0.4699, Adjusted R-squared:  0.4478
## F-statistic: 21.27 on 2 and 48 DF,  p-value: 2.424e-07
```

Our final model was sleep^0.5~gestation + Logbrain, which equates to y = 3.9673871 - 0.0019734(x1) -

0.2287189(x2). Our model indicates that as a mammal's gestation and brain size increase, the total amount they sleep is reduced.

Model 2: Daily Sleep Duration in Mammals with Quantitative and Categorical Predictors

The second multiple linear regression model treats the appropriate variables as categorical rather than quantitative. The initial approach began with ensuring R could read the variables correctly, by transforming them into factor variables. Following this, we dropped variables that were not of interest. The two main variables of non-interest are dream and nondream, as they are derivatives of the dependent variable sleep, and are highly correlated. We also dropped predation and exposure for similar reasons, as the danger index is based off of a combination of predation, exposure, and other information. Animal was also dropped, as that variable is simply a name identifying what specific animal the observation is referring to. This left us with the variables, sleep (y), lifespan, danger, brain, body and gestation. As mentioned previously we log transformed brain and body mass before our analysis, in order to pull in extreme values, and aide in determining linear relationships between the variables in the data set through visual inspection. These log transformed variables were kept and their corresponding original columns were dropped.

Through visual inspection of Figure 1 (above), we found that there are linear relationships between the variables of interest. So we took our next steps in order to identify an appropriate regression model. This began with running all possible regressions, and evaluating the models that the search procedure suggested. In order to cast a wide net, and make the most informed decision in relation to model selection, we also ran other automatic search procedures. For instance, we ran a backward selection procedure using the intercept only model and the full model. This process gave us the model that we finally decided on, $lm(sleep\ danger + log.brain + log.gestation)$. The output of these processes is below:

```
## Subset selection object
## Call: regsubsets.formula(sleep ~ ., data = data2, nbest = 1, really.big = T)
## 8 Variables  (and intercept)
##               Forced in Forced out
## lifespan          FALSE      FALSE
## danger2           FALSE      FALSE
## danger3           FALSE      FALSE
## danger4           FALSE      FALSE
## danger5           FALSE      FALSE
## log.brain         FALSE      FALSE
## log.body          FALSE      FALSE
## log.gestation     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          lifespan danger2 danger3 danger4 danger5 log.brain log.body
## 1  ( 1 ) " "      " "     " "     " "     " "     " "       " "
## 2  ( 1 ) " "      " "     " "     " "     "*"     " "       " "
## 3  ( 1 ) " "      " "     "*"     " "     "*"     " "       " "
## 4  ( 1 ) " "      " "     "*"     "*"     "*"     " "       " "
## 5  ( 1 ) " "      " "     "*"     "*"     "*"     "*"       " "
## 6  ( 1 ) " "      "*"     "*"     "*"     "*"     "*"       " "
## 7  ( 1 ) "*"      "*"     "*"     "*"     "*"     "*"       " "
## 8  ( 1 ) "*"      "*"     "*"     "*"     "*"     "*"       "*"
##          log.gestation
## 1  ( 1 ) "*"
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
```

```
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

```
## [1] 8
```

```
## [1] 6
```

```
## [1] 3
```

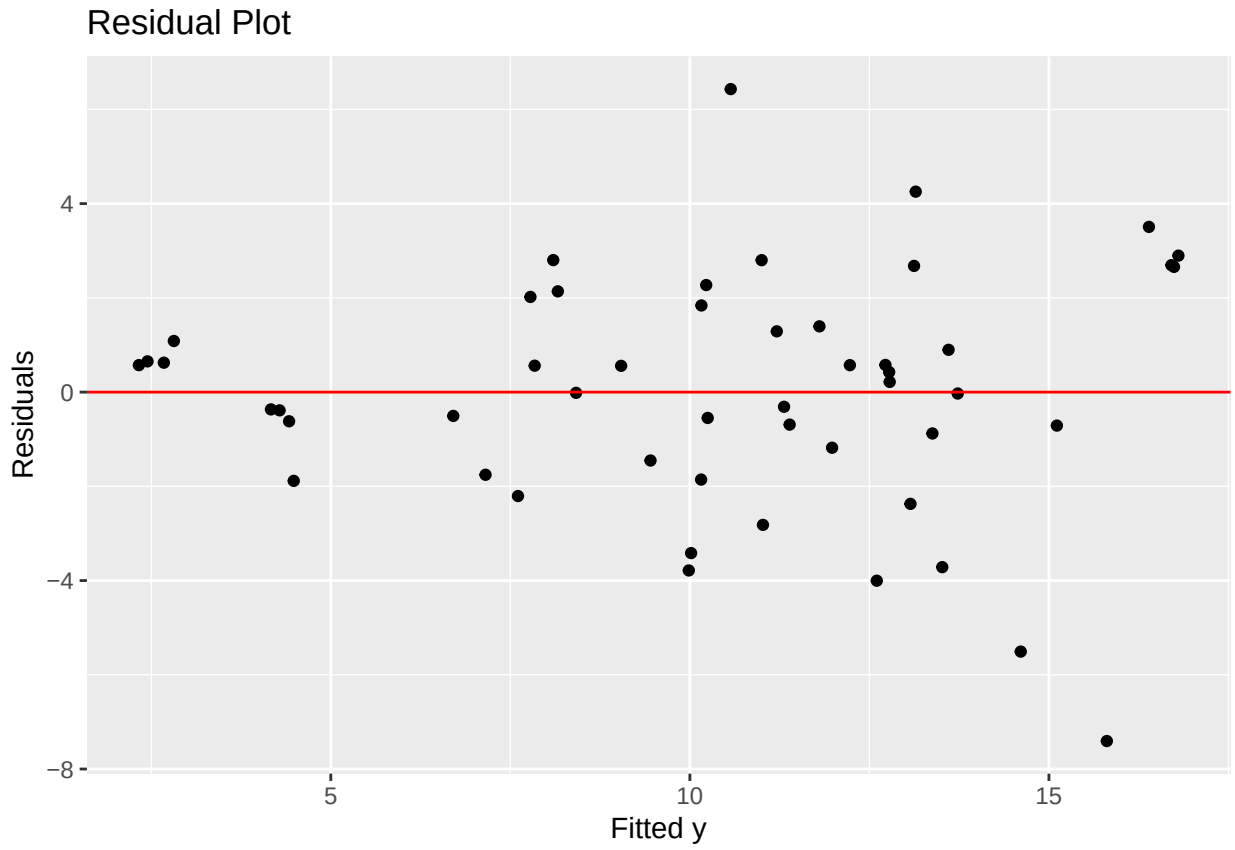
```
## Start:  AIC=111.25
## sleep ~ lifespan + danger + log.brain + log.body + log.gestation
##
##                 Df Sum of Sq    RSS    AIC
## - log.body       1     0.070 317.47 109.26
## - lifespan       1     1.493 318.89 109.48
## - log.brain      1     2.582 319.98 109.66
## <none>                       317.40 111.25
## - log.gestation  1    35.801 353.20 114.70
## - danger         4   260.389 577.79 133.80
##
## Step:  AIC=109.26
## sleep ~ lifespan + danger + log.brain + log.gestation
##
##                 Df Sum of Sq    RSS    AIC
## - lifespan       1     1.424 318.90 107.48
## <none>                       317.47 109.26
## - log.brain      1    17.799 335.27 110.04
## - log.gestation  1    36.090 353.56 112.75
## - danger         4   266.852 584.32 132.37
##
## Step:  AIC=107.48
## sleep ~ danger + log.brain + log.gestation
##
##                 Df Sum of Sq    RSS    AIC
## <none>                       318.90 107.48
## - log.brain      1    30.386 349.28 110.13
## - log.gestation  1    37.615 356.51 111.17
## - danger         4   273.409 592.30 131.06
```

```
##
## Call:
## lm(formula = sleep ~ danger + log.brain + log.gestation, data = data2)
##
## Coefficients:
##   (Intercept)        danger2        danger3        danger4        danger5
##       21.2179        -2.1651        -4.7433        -3.2938        -7.1699
##     log.brain  log.gestation
##       -0.5319        -1.4219
```

```
##
## Call:
## lm(formula = sleep ~ danger + log.brain + log.gestation, data = data2)
##
## Residuals:
```

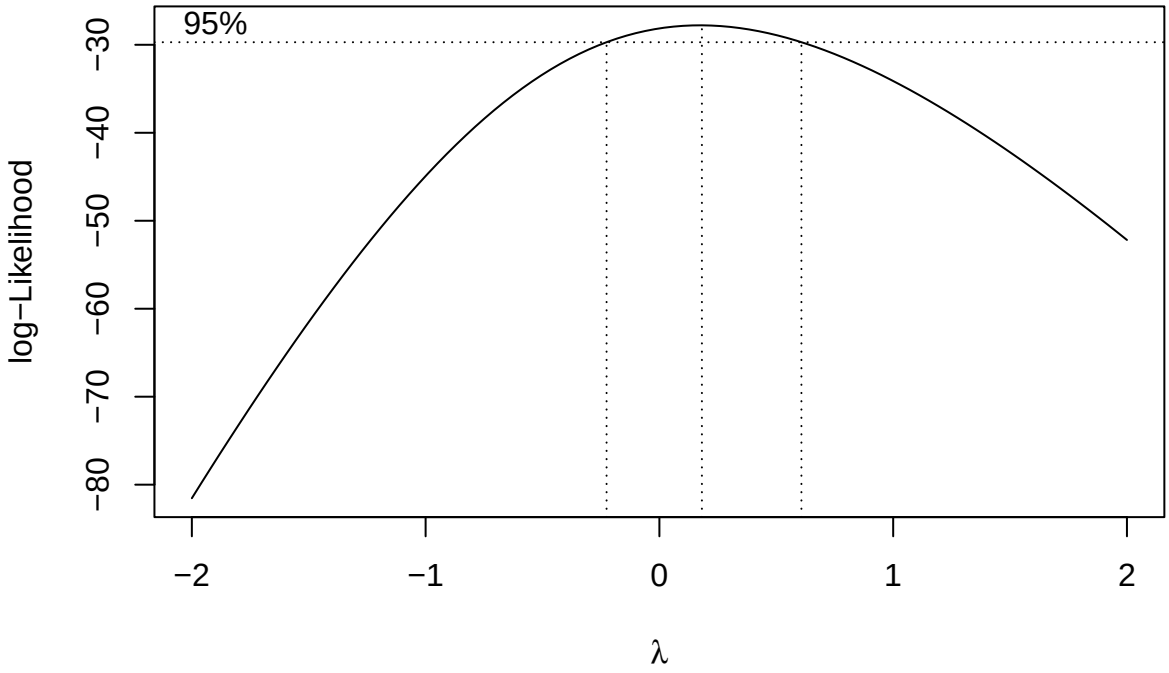
```
##     Min      1Q  Median      3Q     Max
## -7.4064 -1.3164  0.2171  1.6172  6.4313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21.2179     2.2017   9.637 2.06e-12 ***
## danger2        -2.1651     1.0608  -2.041 0.047277 *
## danger3        -4.7433     1.2132  -3.910 0.000316 ***
## danger4        -3.2938     1.1832  -2.784 0.007889 **
## danger5        -7.1699     1.2766  -5.616 1.23e-06 ***
## log.brain      -0.5319     0.2597  -2.048 0.046599 *
## log.gestation  -1.4219     0.6241  -2.278 0.027628 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.692 on 44 degrees of freedom
## Multiple R-squared:  0.7087, Adjusted R-squared:  0.669
## F-statistic: 17.84 on 6 and 44 DF,  p-value: 2.353e-10
```

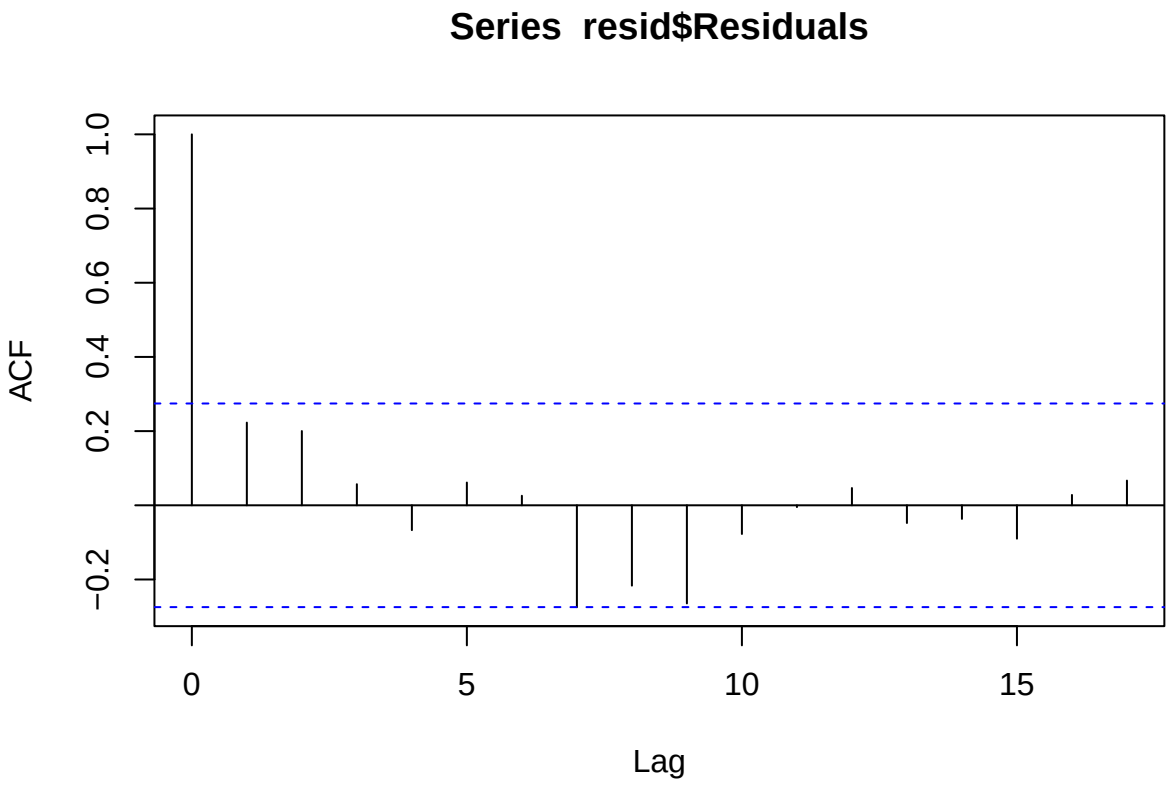
We fit the initial model which had an R^2 of 0.71 and $adj(R^2)$ of 0.67. In order to asses the regression assumptions we displayed a residual plot (below). The mean variance assumption seemed to be met quite nicely, however, the constant variance assumption was not met. This can be seen by the pattern in the residual cloud, as the fitted Y gets larger, the cloud fans out into a con-like shape.

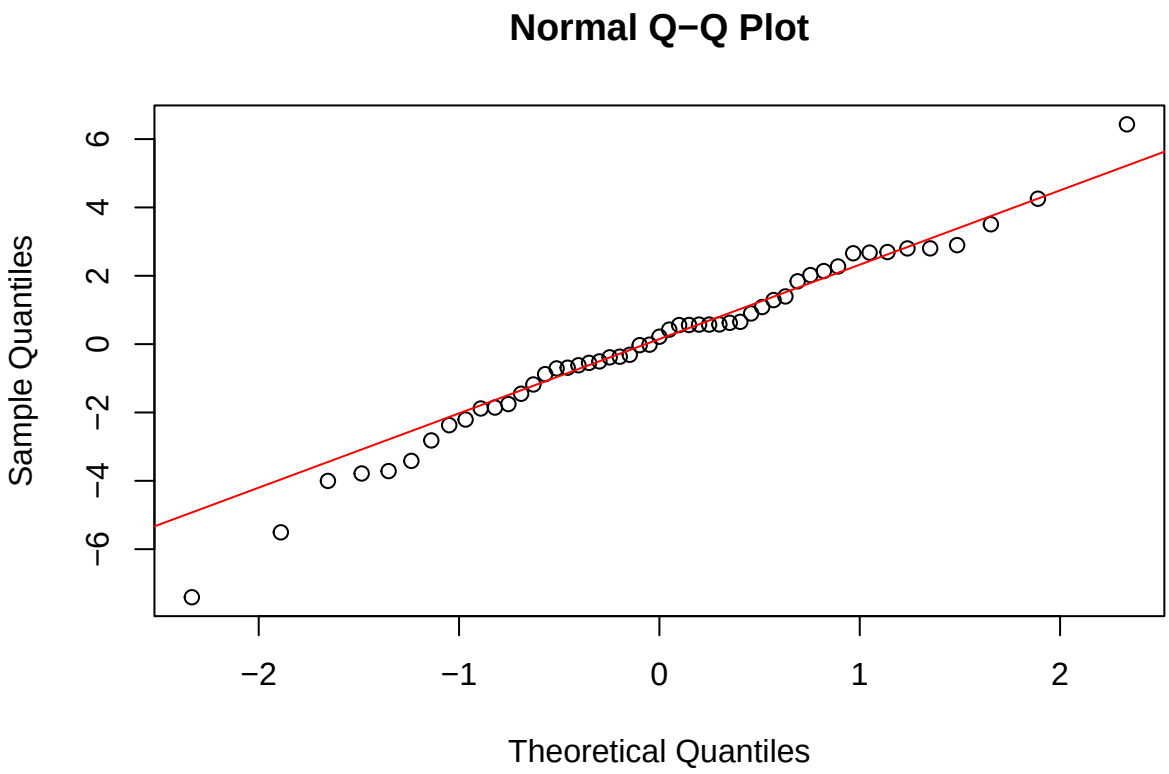


To assess whether or not transformations should be made on the variables, we plotted box cox, ACF, and QQ plots (below). The Normal QQ plot indicated that the data was generally normally distributed, except for the tail ends of the data points which seemed to curve away from the plot line. This could potentially be

an indication of a necessary transformation. The Box Cox plot indicated that a log transformation may be appropriate for the dependent variable. This makes some sense in relation to the fact that other variables in the data set have already been log transformed. the ACF plot also indicated that auto-correlation may be present in the data.



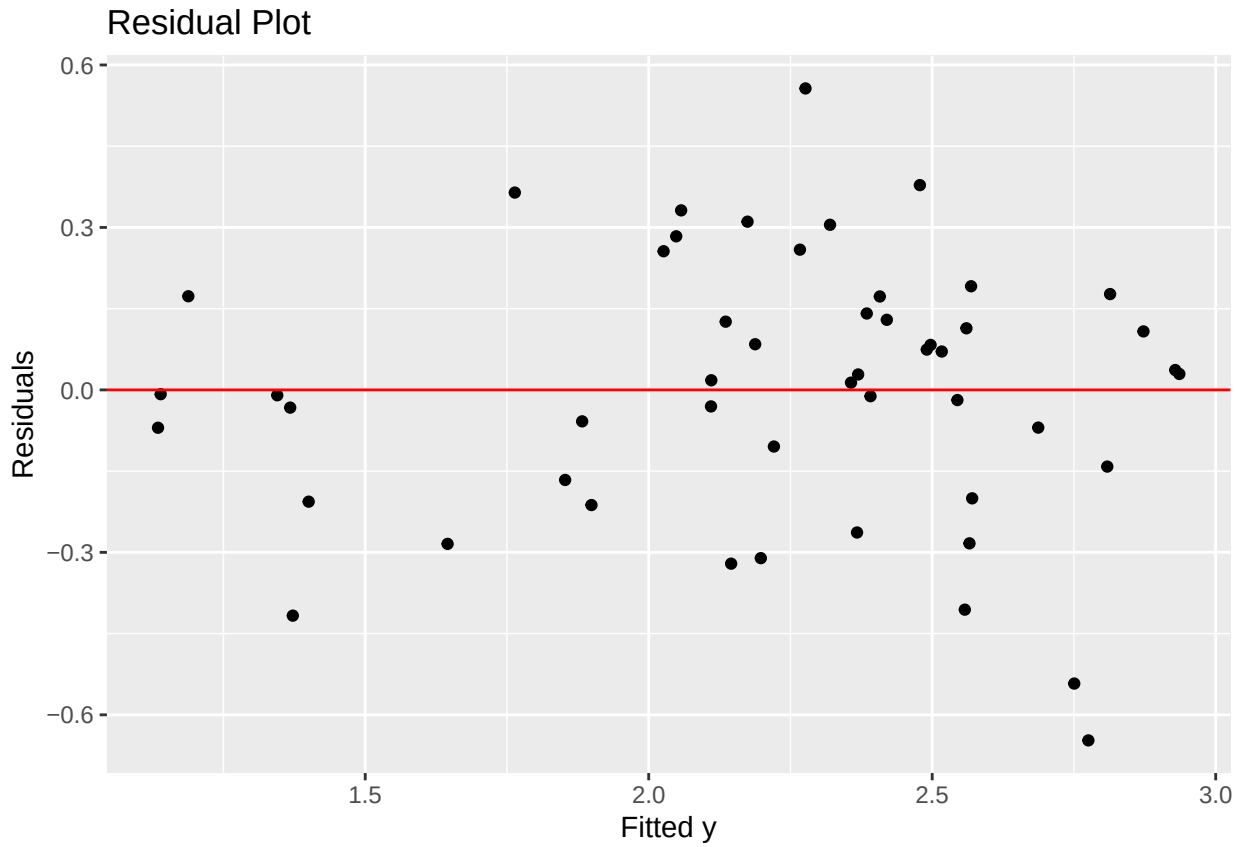

Series resid$Residuals
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0
5
10
15
Lag



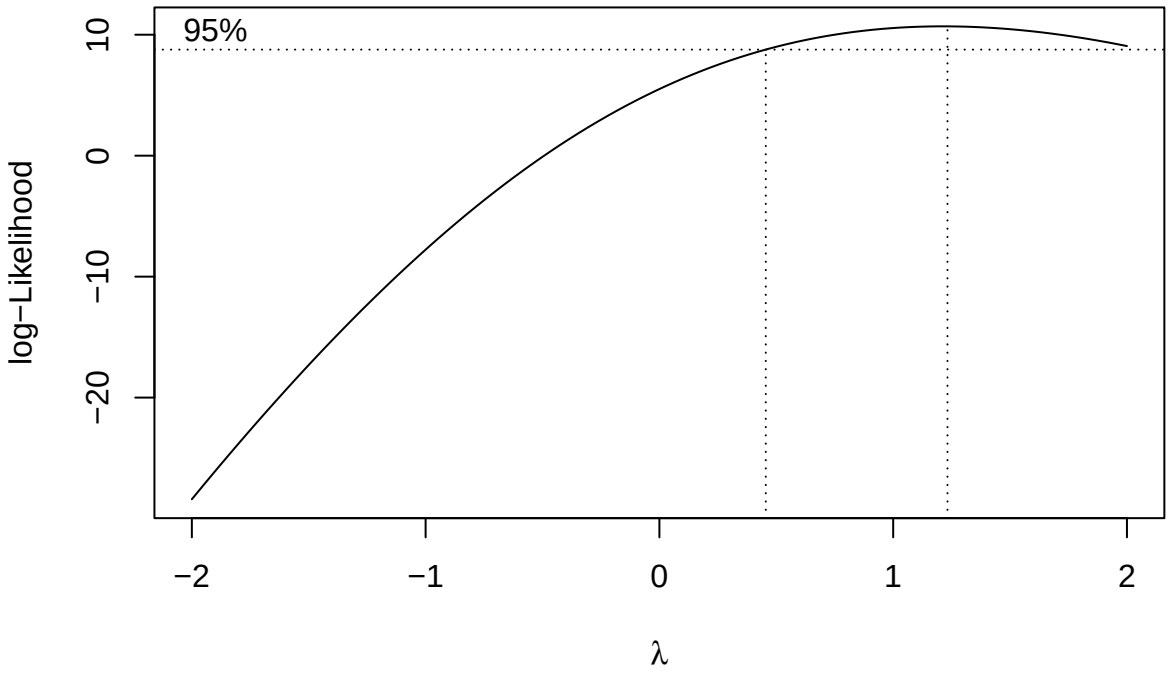
After log transforming the dependent variable, we can see in the output below that the R^2 has improved to 0.79 and the $adj(R^2)$ has improved to .76.

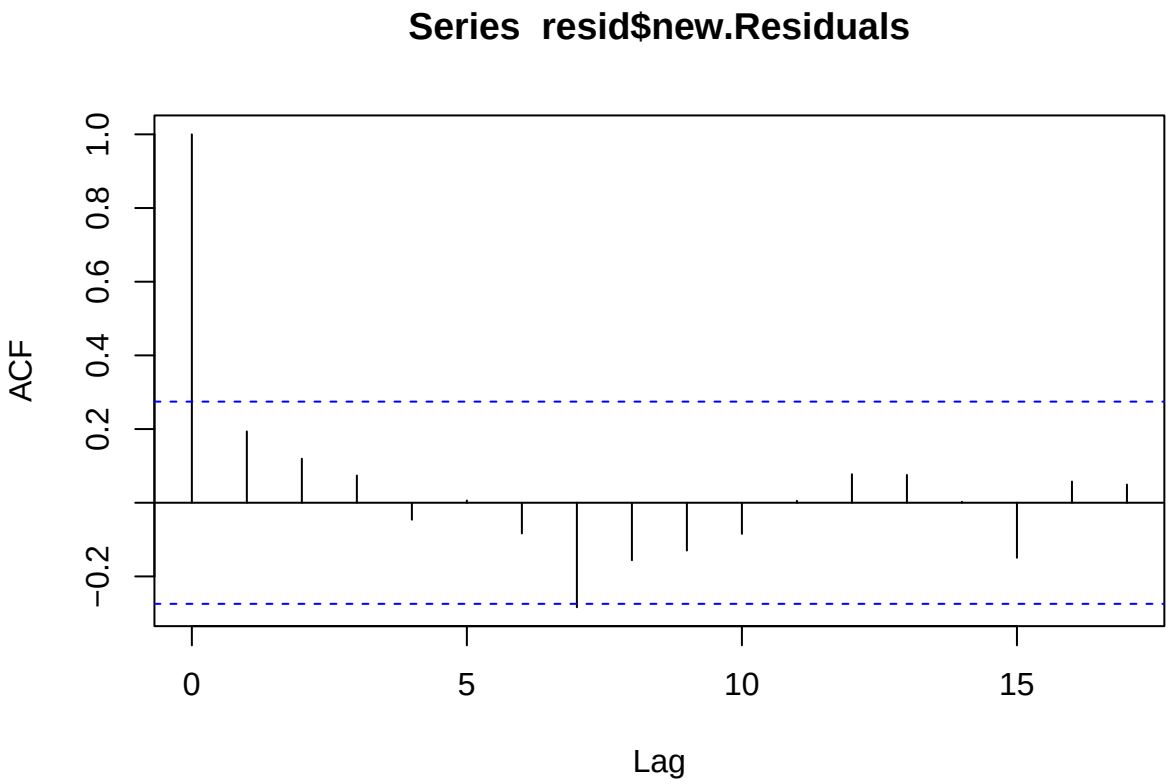
```
##
## Call:
## lm(formula = new.sleep ~ danger + log.brain + log.gestation,
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6472 -0.1539  0.0178  0.1568  0.5567
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.48672    0.21406  16.288  < 2e-16 ***
## danger2       -0.15642    0.10314  -1.517 0.136525
## danger3       -0.47728    0.11796  -4.046 0.000207 ***
## danger4       -0.24829    0.11504  -2.158 0.036399 *
## danger5       -0.96387    0.12412  -7.766 8.78e-10 ***
## log.brain     -0.04536    0.02525  -1.796 0.079375 .
## log.gestation -0.18809    0.06068  -3.100 0.003374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2617 on 44 degrees of freedom
## Multiple R-squared:  0.792,  Adjusted R-squared:  0.7636
## F-statistic: 27.92 on 6 and 44 DF,  p-value: 1.765e-13
```

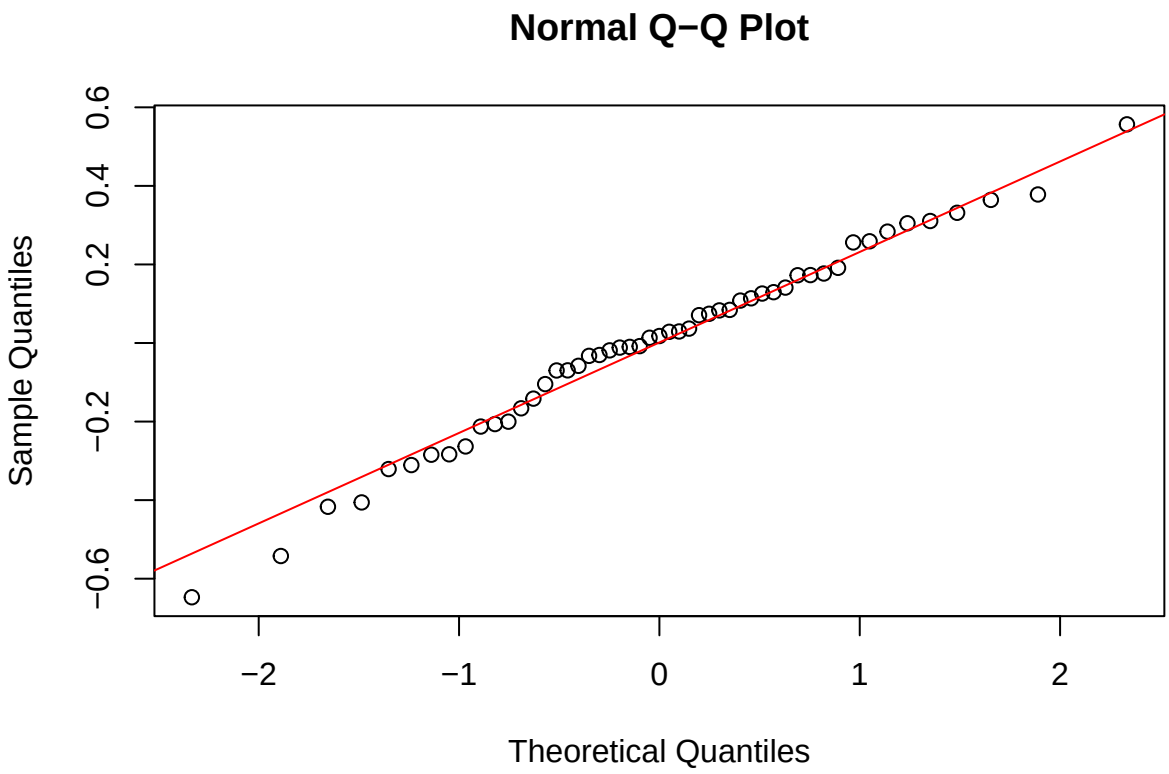
The constant variance assumption also seems to be met based on the residual plot (below), however, the mean variance = 0 assumption is questionable because of an imbalance of negative residuals for the low fitted y values.



The new boxcox plot implies that no more transformations on the dependent variable are necessary. The ACF plot was improved and only indicates auto-correlation at one lag point rather than multiple. Finally, the normal QQ plot seems unchanged. It indicates that the data is generally normally distributed, except at the tails. These plots are below.


95%
log−Likelihood
10
0
−10
−20
−2
−1
0
1
2
λ


Series resid$new.Residuals
ACF
1.0
0.8
0.6
0.4
0.2
−0.2
0
5
10
15
Lag

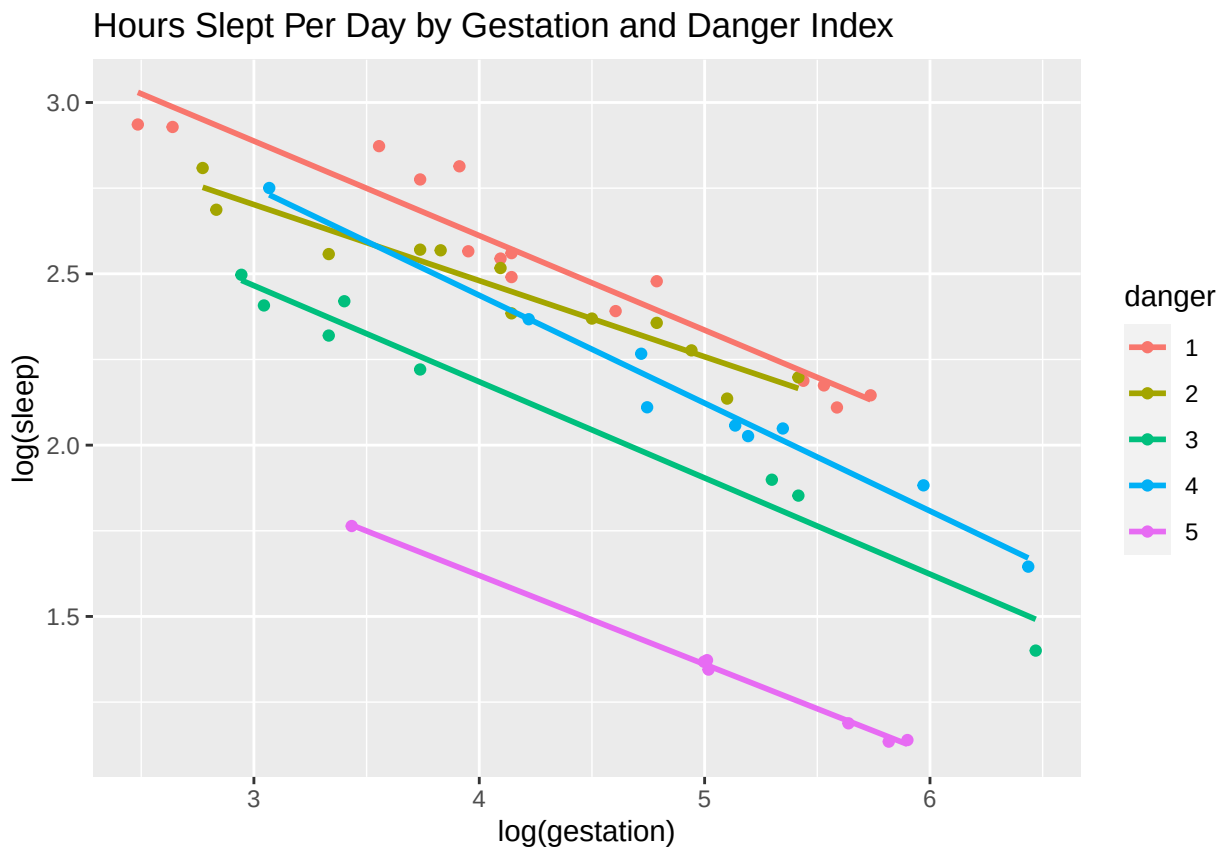


Finally, we checked the variance inflation factors of the independent variables, and no multicollinearity was present based on a threshold of five.

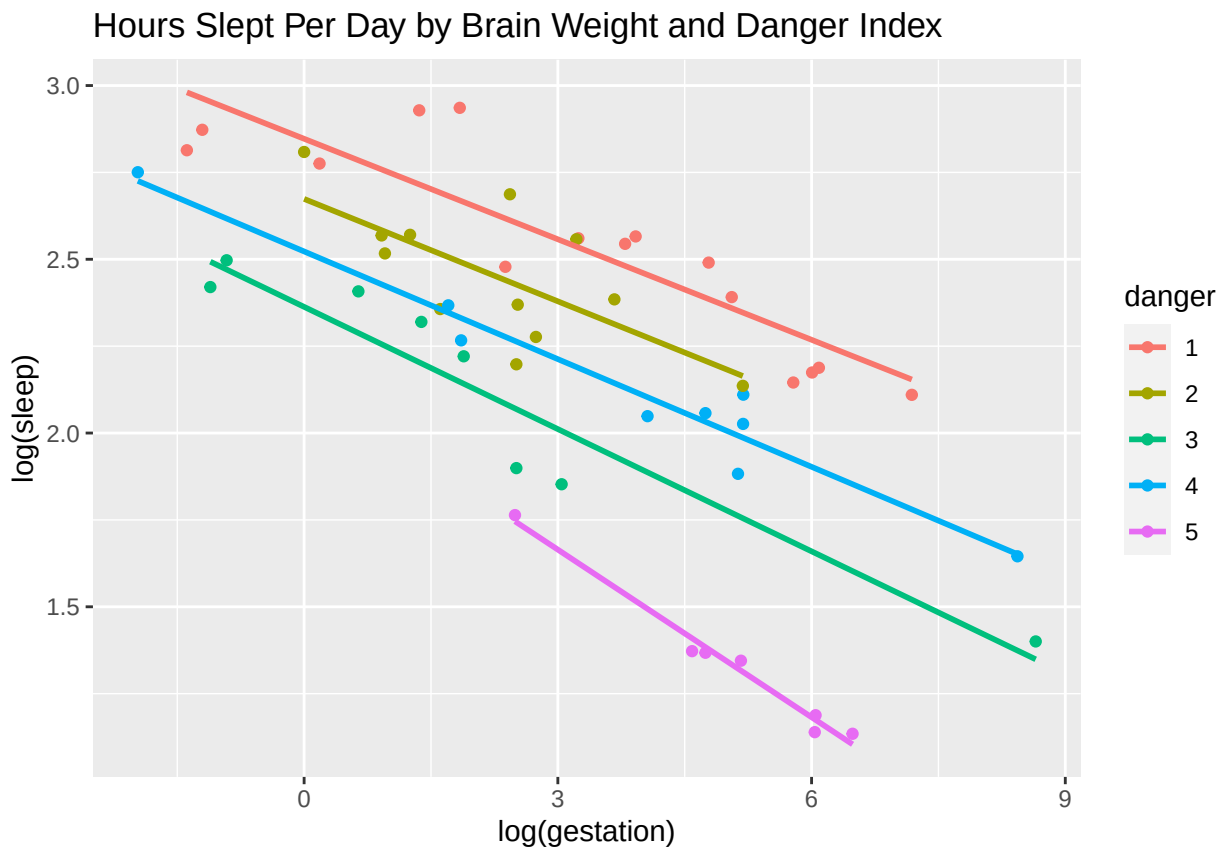
```
##       danger2       danger3       danger4       danger5     log.brain
##      1.424762      1.369840      1.431666      1.358001      2.974526
## log.gestation
##      2.923665
```

The next step we took, after confirming that our regression model was on the right track, was to check for interaction between the independent variables. Based on the scatter plots below, we can see that there are no signs of interaction between sleep and brain (by Danger Index), however there seemed as if there could be interaction between sleep and gestation (by Danger Index). The plot shows that the regression lines for danger 4 an 2 intersect, indicating that an interaction term may be necessary. An interesting note about these graphs is that they indicate that animals with the highest Danger Index, coupled with the highest gestation time get the least amount of sleep per day.

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




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



After seeing that there may be interaction between the independent variables, we fit a new regression model to include the interaction terms:

```
##   2 3 4 5
## 1 0 0 0 0
## 2 1 0 0 0
## 3 0 1 0 0
## 4 0 0 1 0
## 5 0 0 0 1
```

```
##
## Call:
## lm(formula = sleep ~ log.brain * danger + log.gestation * danger,
##     data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4216 -1.0138  0.0845  1.4400  5.9090
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.2153     3.7646   6.167 4.16e-07 ***
## log.brain              -0.8162     0.3949  -2.067    0.046 *
## danger2                -6.8210     5.5247  -1.235    0.225
## danger3                 0.1587     6.9337   0.023    0.982
## danger4               -12.0827    10.8349  -1.115    0.272
## danger5                -3.8710    15.3393  -0.252    0.802
```

```
## log.gestation          -1.6708     1.0550  -1.584    0.122
## log.brain:danger2       0.1701     0.7663   0.222    0.826
## log.brain:danger3       0.9549     0.8193   1.165    0.252
## log.brain:danger4       0.4273     0.9610   0.445    0.659
## log.brain:danger5       2.4109     4.4599   0.541    0.592
## danger2:log.gestation   0.9560     1.5040   0.636    0.529
## danger3:log.gestation  -1.7124     1.9663  -0.871    0.390
## danger4:log.gestation   1.5027     2.8413   0.529    0.600
## danger5:log.gestation  -2.8976     7.1916  -0.403    0.689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.732 on 36 degrees of freedom
## Multiple R-squared:  0.7546, Adjusted R-squared:  0.6591
## F-statistic: 7.906 on 14 and 36 DF,  p-value: 2.882e-07
```

This model had a lower R^2 and the $adj(R^2)$ seemed to indicate over fitting, as it was approximately 0.1 lower. We conducted a hypothesis test in order to check if the interaction terms were significant. By writing a function that would perform the appropriate arithmetic and return the p-value associated with a hypothesis test for interaction terms.

$$H_0\text{: } \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$$

$$H_a\text{: Not all Coefficients in Null} = 0$$

```
## Analysis of Variance Table
##
## Response: sleep
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## log.brain              1 416.10  416.10 55.7465 8.104e-09 ***
## danger                 4 322.26   80.56 10.7935 7.464e-06 ***
## log.gestation          1  37.62   37.62  5.0394   0.03101 *
## log.brain:danger       4  33.08    8.27  1.1080   0.36777
## danger:log.gestation   4  17.10    4.28  0.5729   0.68402
## Residuals             36 268.71    7.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

```
## [1] 416.09930 322.25744  37.61507  33.08237  17.10469 268.70859
```

```
## [1] 0.6322315
```

```
## [1] 0.3159819
```

Both of the p-values that were returned were insignificant, indicated that the interaction terms could be dropped from the model and that we should favor the reduced, first order additive model over the model with interaction terms. In order to see which model may be better at predicting on new data, we use dour function that returns a PRESS statistic for a regression model. Our model had, by far, the lowest PRESS statistic. This indicated that it would most ;likely perform better when predicting on new data compared to the other two models.

```
## [1] 431.3236
```

```
## [1] 4.13968
```

```
## [1] 506.8601
```

Overall, we found that "new.model" seemed to be the strongest model when treating danger as a categorical variable rather than a quantitative variable. We had a relatively strong R^2 of 0.79, and the $adj(R^2)$ did not seem to indicate over fitting. The regression assumptions seemed to be met well, except for variance mean= 0 which was harder to judge because of the skewed values at the lower end of fitted y values. Our diagnostics indicated transformations were needed, and after they were performed our model improved and seemed to indicate good model fit.

Model 3: Longevity in Mammals

Our second objective was to identify a model that predicts and explains the odds of being a long-living species. To do this, we split the original dataset into training (75%) and testing (25%) sets, and used the training set for all model-building.

Knowing that many of the quantitative predictors were strongly correlated to each other, our first step was to narrow down the list of candidate quantitative predictors. We started with a model for longevity that included all four quantitative predictors: log(body), log(brain), sleep, and gestation. The output of this model is shown below.

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
## glm(formula = longliving ~ logbody + logbrain + sleep + gestation,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.323e-04  -2.000e-08  -2.000e-08   0.000e+00   3.623e-04
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2361.598 348994.591  -0.007    0.995
## logbody       -165.062  25025.407  -0.007    0.995
## logbrain       456.119  67168.745   0.007    0.995
## sleep           -5.498   1505.165  -0.004    0.997
## gestation        2.331    361.617   0.006    0.995
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.4403e+01  on 38  degrees of freedom
## Residual deviance: 2.4248e-07  on 34  degrees of freedom
##   (7 observations deleted due to missingness)
## AIC: 10
##
## Number of Fisher Scoring iterations: 25
```

We calculated a ΔG^2 statistic to test the hypothesis the the model was better than an intercept only model. The null hypothesis was that all coefficients in the model equaled 0, and the alternative hypothesis was that at least one coefficient was not 0. We found $\Delta G^2 = 44.40295$ and P = 5.291002e-09; in other words, the model was significantly better at predicting longevity than an intercept only model. Yet, from the model output, we noted that none of the individual coefficients had a significant z-value, confirming our suspicion that multicollinearity among predictors was an issue we needed to address. We also noted the two warning messages, which seemed to indicate that this model perfectly classified species. We generated a confusion matrix using our training set to see if this was the case:

```
##
##       FALSE TRUE
##   no     29    0
##   yes     0   10
```

As suspected, this model perfectly classified species in the training set, which in turn made coefficient variances very high, and therefore it was difficult to relate predictors to the response. We also knew that multicollinearity was an issue, so we proceeded to test different combinations of these four quantitative predictors to see which could be eliminated from the model. A model of longevity as a function of log(body) and log(brain), for example, gave the following output:

```
##
## Call:
## glm(formula = longliving ~ logbody + logbrain, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.00136  -0.23617  -0.04210  -0.00081   1.76619
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.8636     5.1784  -2.484   0.0130 *
## logbody      -0.2737     0.5955  -0.460   0.6458
## logbrain      2.5203     1.2469   2.021   0.0433 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.164  on 43  degrees of freedom
## Residual deviance: 18.937  on 41  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 24.937
##
## Number of Fisher Scoring iterations: 8
```

From this, we saw that, in the presence of log(brain), log(body) did not add anything to the model, given that its p-value was 0.6458. We proceeded to try a total of 8 models, combining different sets of quantitative predictors. Several showed the same warning messages that we got in our first model, and also showed the significant ΔG^2 but insignificant individual predictors. In the end, we determined that the only quantitative predictor we needed to keep was log(brain). The output of this model is below:

```
##
## Call:
## glm(formula = longliving ~ logbrain, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05852  -0.25440  -0.05089  -0.00110   1.59000
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.7018     4.2301  -2.766  0.00567 **
## logbrain      2.0994     0.7691   2.730  0.00634 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.164  on 43  degrees of freedom
## Residual deviance: 19.153  on 42  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 23.153
##
## Number of Fisher Scoring iterations: 8
```

With quantitative variables narrowed down to just log(brain), we next considered whether to add any of the categorical variables (predation index, exposure index, or danger index) to the model. We also considered whether to add them as additional quantitative variables (since they were each on a scale of 1-5), or whether to add them as true categorical predictors. Since the danger index was based on the predation and exposure indices combined with other information, we first considered the addition of this index to the model. When danger index was treated as a quantitative variable, we got the following output:

```
##
## Call:
## glm(formula = longliving ~ logbrain + danger, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84725  -0.19440  -0.02704  -0.00038   1.84576
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.4364     4.7918  -2.595  0.00945 **
## logbrain      2.5042     0.9698   2.582  0.00982 **
## danger       -0.4780     0.4000  -1.195  0.23207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.164  on 43  degrees of freedom
## Residual deviance: 17.443  on 41  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 23.443
##
## Number of Fisher Scoring iterations: 8
```

When danger index was treated as a categorical predictor, we saw the following output:

```
##
## Call:
## glm(formula = longliving ~ logbrain + danger2, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.87722  -0.07170  -0.00218   0.00000   1.35801
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.1475    10.3792  -2.134   0.0329 *
## logbrain      4.1070     1.8968   2.165   0.0304 *
## danger22      3.9417     4.2138   0.935   0.3496
## danger23     -1.9207   217.6468  -0.009   0.9930
## danger24      0.6792     2.0166   0.337   0.7363
## danger25     -3.0652     2.2667  -1.352   0.1763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.164  on 43  degrees of freedom
## Residual deviance: 14.047  on 38  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 26.047
##
## Number of Fisher Scoring iterations: 14
```

The outputs from both these models indicated that including the danger index did not significantly improve the model beyond just including log(brain), so we did not move forward with this predictor.

Based on our exploratory data analysis, the index that seemed most clearly related to longevity was exposure index, so we tried this next. We used the same approach as with the danger index, first trying it as a quantitative variable and then as a categorical variable. Our output for these models was very similar to the outputs obtained when we added the danger index; neither showed a clear benefit of keeping the additional predictor.

Our conclusion from analyzing all these models was that, due to the relationships among the different predictors, we only needed one of these variables to model longevity, and the best predictor was log(brain).The equation for this model (output shown previously) is:

$$log(\frac{\pi}{1-\pi}) = -11.7018 + 2.0994(x_1)$$

where π = the probability of being longliving and x_1= log(brain mass). This equation tells us that for a 10% increase in brain mass, the log odds of being longliving increase by a factor of 2.0994*log(1.1) = 0.2001. Thus, the odds of being longliving increase by a factor of exp(0.2001) = 1.22 for a 10% increase in brain mass.

We calculated a ΔG^2 statistic to test the hypothesis that this model was useful. The null hypothesis was that $\beta_1 = 0$, and the alternative hypothesis was that β_1 was not 0. We found $\Delta G^2 = 28.01163$ and P = 1.205883e-07; in other words, the model was significantly better at predicting longevity than an intercept only model. We generated a confusion matrix with threshold set at 0.5 to test the predictive ability of this model on our test set, shown below:

```
##
##       FALSE TRUE
##   no     11    0
##   yes     2    1
```

All predicted probabilities are essentially 0 or 1, so changing the threshold did not improve the overall accuracy. When we applied our first model to the test set, we obtained the following confusion matrix:

```
##
##       FALSE TRUE
##   no     10    0
##   yes     1    1
```

This model - with all quantitative predictors - had a 0% false positive rate and a 50% false negative rate, but due to multicollinearity among predictors and the consequent high variances and wide confidence intervals of all the coefficients, our ability to infer relationships between the predictors and the response variable was limited. Our final model also has a 0% false positive rate, but it has a 67% false negative rate. We have thus sacrificed some classification ability with this reduced model, but we can relate log(brain mass) to odds of being longliving.

Conclusion

Based on the conclusions drawn from the MLR treating categorical variables as quantitative, specifically none of the categorical predictors were maintained in the final model and the R^2 value was fairly low, we likely should not pursue a model treating categorical variables in this manner. This is further supported by the findings from the model where categorical variables were treated as categorical, where the R^2 value was significantly higher.

In modeling the odds of being longliving, we discovered that, while it was possible to build a model that perfectly classified the training set, this led to very high confidence intervals for coefficients and consequently, none of the coefficients were individually significant. If our goal were to predict the odds of being longliving as accurately as possible, we could use this model, however, it was difficult to relate different predictors to the odds of being longliving. A simpler model sacrificed some accuracy but allowed us to relate log(brain mass) to the odds of being longliving.