James Scholar Stat200 MLR Report

Max Zhang

2023-04-30

In this Multiple Linear Regression report, I will be analyzing the "mammals" data set from Openintro which includes measures for 39 species of mammals. These data were originally recorded to study the relationship between biological/environmental factors and sleeping in mammals.

```
mammals=read.csv("C:/Users/maxzr/Downloads/mammals.csv")
colnames(mammals)
```

```
##  [1] "species"      "body_wt"      "brain_wt"     "non_dreaming" "dreaming"
##  [6] "total_sleep"  "life_span"    "gestation"    "predation"    "exposure"
## [11] "danger"
```

Biological variables included: body weight, brain weight, life span, and gestation time.

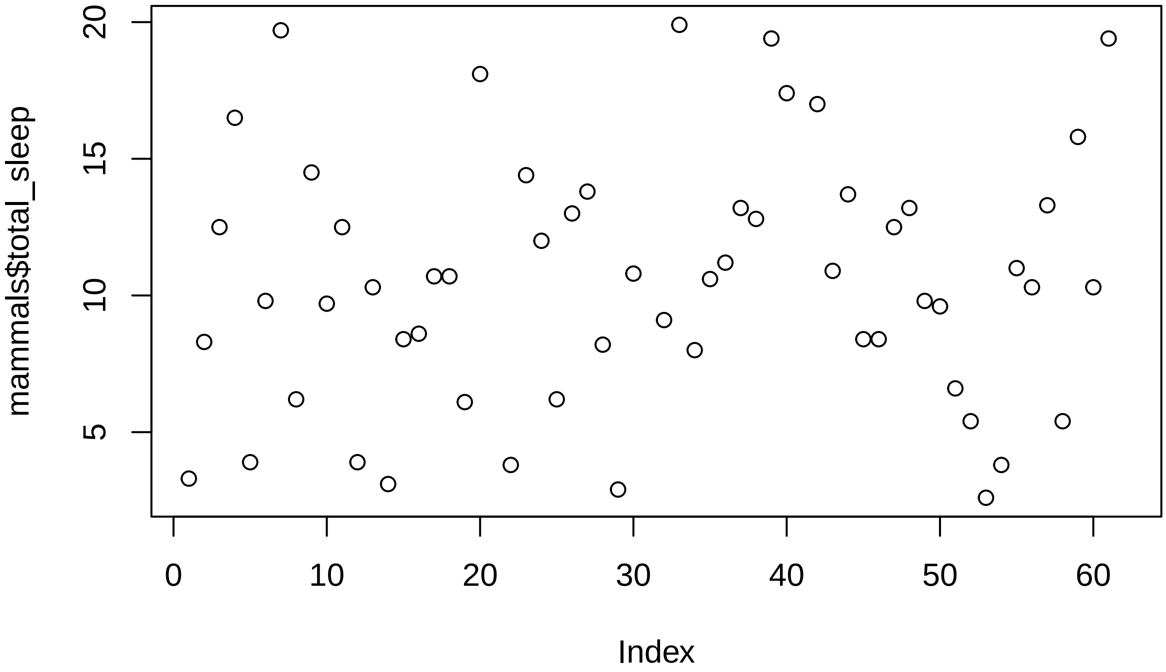
Environmental variables included: predation (likelihood of being preyed upon), exposure (during sleep), and how much danger the mammal encounters from other animals. All 3 of those were operationalized into a scale from 1(least) to 5(most).

Hence, the aim of this report is to use various methods to search for the best fitting model that predicts the total sleep time of the 39 mammals in the data set.

```
head(mammals)
```

```
##                    species  body_wt brain_wt non_dreaming dreaming total_sleep
## 1        Africanelephant 6654.000   5712.0           NA       NA         3.3
## 2 Africangiantpouchedrat    1.000      6.6          6.3      2.0         8.3
## 3              ArcticFox    3.385     44.5           NA       NA        12.5
## 4   Arcticgroundsquirrel    0.920      5.7           NA       NA        16.5
## 5          Asianelephant 2547.000   4603.0          2.1      1.8         3.9
## 6                 Baboon   10.550    179.5          9.1      0.7         9.8
##   life_span gestation predation exposure danger
## 1      38.6       645         3        5      3
## 2       4.5        42         3        1      3
## 3      14.0        60         1        1      1
## 4        NA        25         5        2      3
## 5      69.0       624         3        5      4
## 6      27.0       180         4        4      4
```

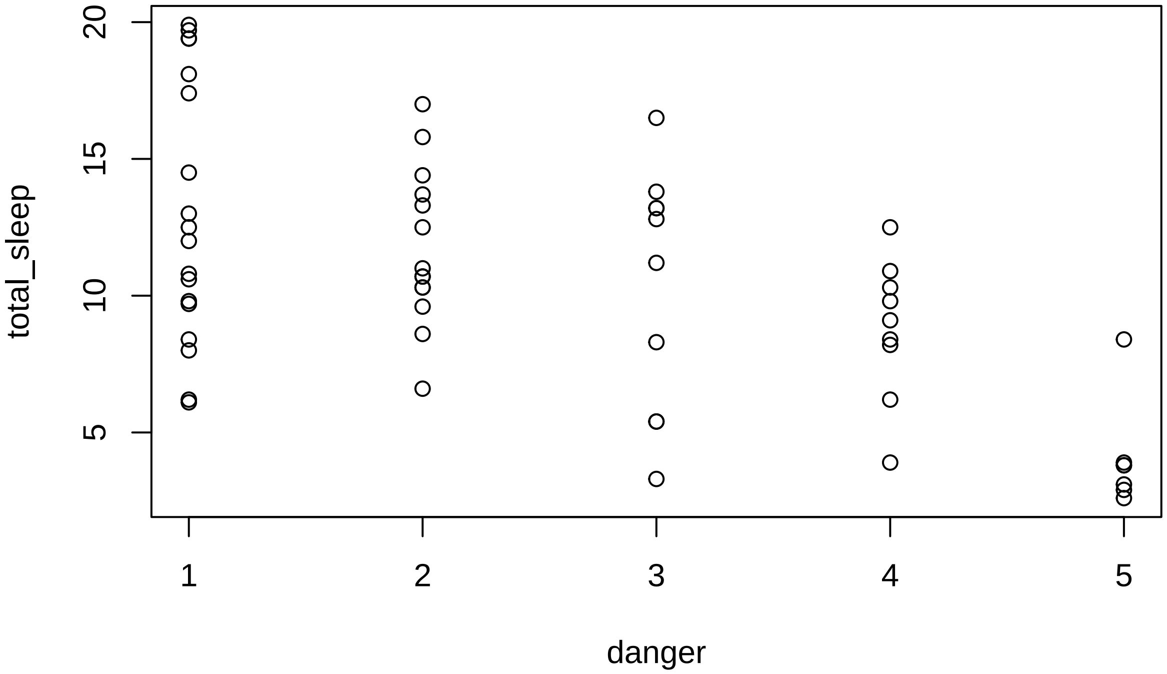
```
plot(mammals$total_sleep)
```



The study actually recorded both dreaming sleep and non-dreaming sleep for all the mammals, however, in this report only the total amount of sleep time is the interested response variable, hence I will only be referring to total_sleep throughout.

I will start developing the MLR model by plotting total sleep with the intuitively most significant variable: predation, as mammals that face more frequent threats from predators will likely spend less time sleeping soundly overall.

```
plot(total_sleep ~ danger, data=mammals)
```



By eye-balling, we can observe a pretty obvious negative correlation, meaning that on average, as the level of danger faced by the mammal increases, their total sleep time decreases.

```
cor(mammals$total_sleep,mammals$danger, use = "complete.obs")
```

```
## [1] -0.5877424
```

```
#a correlation coefficient of -0.588 between the 2 variables, which can be considered as strongly negat
#(complete.obs was used as there are observations in the data set where data is N/A)
```

```
mdanger = lm(total_sleep ~ danger, data=mammals)
summary(mdanger)
```

```
##
## Call:
## lm(formula = total_sleep ~ danger, data = mammals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4177 -2.8070 -0.7225  3.0865  6.8728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.4630     1.0326  14.975  < 2e-16 ***
## danger       -1.9452     0.3578  -5.436 1.23e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.76 on 56 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3454, Adjusted R-squared:  0.3338
## F-statistic: 29.55 on 1 and 56 DF,  p-value: 1.23e-06
```

Looking at the summary of the linear model between total_sleep and danger, we can conclude a significant negative relationship between the total sleep time of a mammal and the level of danger they’re facing. For every single unit increase in the danger level, we would expect a 1.9452 hours reduction on sleep time. - a p-value of 1.23e-06 indicates a highly statistically significant relationship. Based on the adjusted R-squared, ~33.38% of the variation in sleep time can be explained by danger level, despite already having moderate predictive powers, it also suggests that there are other predictors not included in the model that would contribute to a mammal’s total sleep time.

Hence, the next step is to conduct multiple linear regression to determine the most significant predictors. I will be using backward selection, starting with all predictors and eliminating the least contributing one successively until the highest adjusted R-squared is obtained.

```
mfull = lm(total_sleep ~ body_wt + brain_wt + life_span + gestation + predation + exposure + danger,
           data = mammals)
summary(mfull)
```

```
##
## Call:
## lm(formula = total_sleep ~ body_wt + brain_wt + life_span + gestation +
##     predation + exposure + danger, data = mammals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8931 -1.9556  0.0842  1.4620  6.5572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.603699   1.078176  15.400  < 2e-16 ***
## body_wt     -0.001601   0.001464  -1.094 0.280039
## brain_wt     0.002319   0.001614   1.437 0.157923
## life_span   -0.039830   0.035366  -1.126 0.266323
## gestation   -0.016465   0.006214  -2.650 0.011230 *
## predation    2.393361   0.971101   2.465 0.017789 *
## exposure     0.632923   0.558573   1.133 0.263448
## danger      -4.508571   1.186118  -3.801 0.000449 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.846 on 43 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.682,  Adjusted R-squared:  0.6302
## F-statistic: 13.17 on 7 and 43 DF,  p-value: 6.364e-09
```

After running a summary on the full model, it’s clear that the not all explanatory variables are contributing to the fitness of the model, therefore I will start eliminating the non-significant ones by the following criterion: 1. p-value: predictors with a p-value too high to suggest a statistically significant relationship will be eliminated as they negatively contribute towards the model’s accuracy. 2. practical impact: predictors with low p-value but also minimal practical impact on total sleep time (very low coefficient) will also be eliminated as they don’t influence the response variable significant enough. Or in other words, based on context knowledge, predictors will be eliminated if they are known from prior knowledge/theory to be insignificant in the context of this observation. 3. model simplicity: only trying to keep the most significant predictors for the simplest model.

Abiding those criterion, the first predictor to be eliminated from the model will be “body_wt”, with a 0.28 p-value and a coefficient of almost less than -0.001, we can confidently say it’s not affecting how much mammals are sleeping.

```
mupdated = lm(total_sleep ~ gestation + predation + danger, data = mammals)
summary(mupdated)
```

```
##
## Call:
## lm(formula = total_sleep ~ gestation + predation + danger, data = mammals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3085 -2.1425 -0.2078  1.5183  6.4208
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.651683   0.907103  17.255  < 2e-16 ***
## gestation   -0.011775   0.003322  -3.545 0.000863 ***
## predation    2.193877   0.800392   2.741 0.008473 **
## danger      -3.777656   0.879825  -4.294 8.07e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.85 on 50 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6415, Adjusted R-squared:  0.6199
## F-statistic: 29.82 on 3 and 50 DF,  p-value: 3.385e-11
```

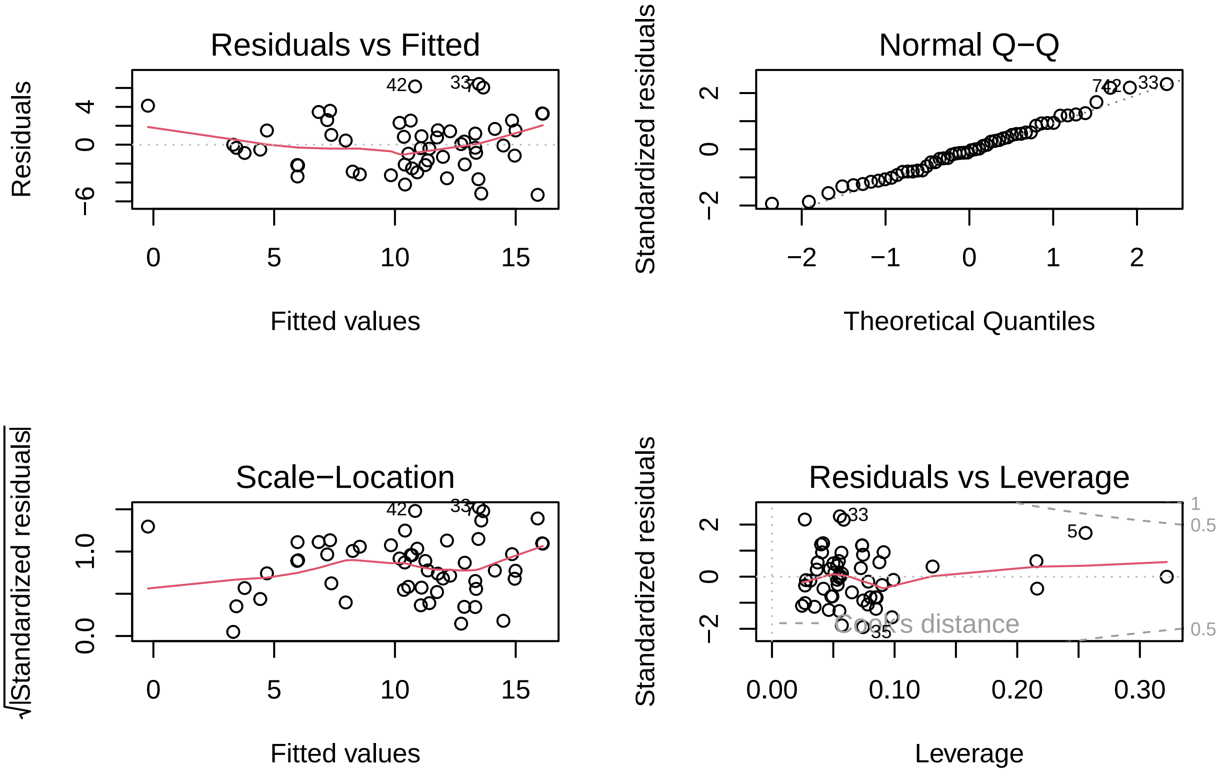
After repeating the selection process 3 times, body_wt, brain_wt, and life span have been abandoned from the model, the updated model consists of 3 predicting variables remaining, as indicated, are all significant at the 99% confidence level. Both predation and danger also has a high coefficient, which means for each unit change in these categories, total sleep time is influenced by a lot. Gestation, on the other hand, despite having an extremely low p-value, also has a low coefficient, but after conducting some research online, I found ample concrete evidence from previous studies that gestation is proved to be negatively linked to sleep duration of mammals, especially the paper “Negative correlation between gestation and sleep duration in mammals” published on Dovepress. (Gonfalone A, 2016). Therefore, I’ve decided to keep it in the final model.

Another factor that impacts the results is the fact that danger and predation, unlike gestation, are not technically continuous data, both predation and danger are operationalized into a scale of 1-5 to fit into the data set.

The final model has an adjusted R-squared of 61.99, explaining ~62% of the variability observed in mammal sleep time, this is an almost 30% improvement in predictive power compared to the simple linear regression model previously where danger was the only predictor.

Lastly, I will use some diagnostic plots to verify whether the model is reasonable and accurate.

```
par(mfrow=c(2,2));plot(mupdated)
```



Residuals vs. Fitted: residuals almost all randomly scatter around the horizontal axis, since there isn't a clear pattern we can say there's a linear relationship between the predictors and the response. However, it's still slightly heterostatistic as the residuals spread out more from 0 towards the right side.

Normal Q-Q: almost all points are lying on or close to the 45-degree line, which suggests that the distribution of residuals follows a normal distribution per the assumption of a linear regression model.

Residuals vs. Leverage: we see most of the points scattering around the horizontal line at zero, which is a good sign. Despite there are no points outside of the cook's distance in this particular graph, point #33 is indicated in all 4 graphs and point 42# in 3 graphs, both of those points deviate from where majority of the other points falls and are potential influential points.