MATH4432: Project 1

Regression: Animal Species Sleeping Hours

Sarah Catherine James
20501098

Alexander Moellers
20501050

Nora Eliza Norden
20500343

Thursday 15th March, 2018

1 Introduction

While there are some key theories, such as restoration and inactivity, for the biological function of sleep in mammals, none are conclusive. To better understand this function, we aim to explore what factors affect the sleep requirements of mammals. To do this, we analyze the interrelationships between sleep and exogenous and endogenous variables and their impact on sleep behavior. Once we better understand the biological purpose of sleep in mammals, assuming there is one at all, we aim to create a statistical model that can predict this behavior for any mammal.

The data set investigates variables related to the inactivity theory of sleep, which suggests that animals are inactive at night in order to protect them from the danger of being vulnerable in the dark. Specifically, the data set contains four constitutional characteristics- lifespan (in years), gestation time (in months), brain weight (in grams) and body weight (in kilograms)- and three ecological variables- sleep exposure, predation (likelihood of being preyed upon) and danger. Sleep exposure and predation were measured using a 5-point index where 1 and 5 indicate the species is least/most exposed and least/most likely to be preyed upon respectively. Danger was also measured on a 5-point index, based on the previous two variables and further information, with 1 indicating the lowest danger and 5 indicating the highest.

In this report we will state our findings on the significance of the above variables on the sleep habits of 62 species of mammal, and explore what our findings tell us about the sleep requirements of mammals. The study was performed using R, and the full code can be found on Github.

2 Data Exploration

2.1 Dealing with missing values

Upon examination of the data set it became clear that 20 instances (where an instance is a species) were missing values for one or more variables. However, just omitting these instances was not an option as it would have significantly reduced the data available for training and testing of the model. Therefore, the team decided to impute the missing data. A quick check of the data revealed that 14 values of the feature *SlowWaveSleep*, 12 of *dreamSleep* and 4 of *sleep*, *danger* and *life* were not available. Furthermore, the features *slowWaveSleep*, *dreamSleep* and "Sleep" depend on each other, because sleep is simply calculated as the sum of *slowWaveSleep* and *dreamSleep*. To impute data in these three features the affected instances were first examined more closely.

It turned out that in ten instances *dreamSleep* and *SlowWaveSleep* were missing, but *Sleep* was available and in two instances, *dreamSleep* was available, but *Sleep* and *SlowWaveSleep* were missing. To deal with these an additional column *frac_SlowWave_dream* was added to the data frame. It was then filled with the numbers obtained by dividing the value of *SlowWaveSleep* by *dreamSleep*. By running a correlation and a summary command on *frac_SlowWave_dream* the mean and the correlations with other variables can be obtained. As there was no significant correlation between *frac_SlowWave_dream* and any variable other than *dreamSleep*, we deem it reasonable to use the mean of *frac_SlowWave_dream*, 5.469, to fill in missing values for *dreamSleep* and *SlowWaveSleep* where possible. We are then left with another two instances where all three sleep features were not filled in, which we omit as they provide no information to train our model for predicting sleep.

Similarly, we had four instances in which *gestation* was missing. After running the correlation command, we find a strong a correlation between *gestation* and *slowWaveSleep* and *brain*. Therefore, we used linear regression to fill in the four missing values for *gestation*. Finally, for the four instances in which there was no data for *life*, we used linear regression to fill in these values using *brain* and *gestation*, due to the strong correlation found between these variable.

We now have a total of 60 species with complete data to train our model for predicting the amount of sleep a mammal requires.

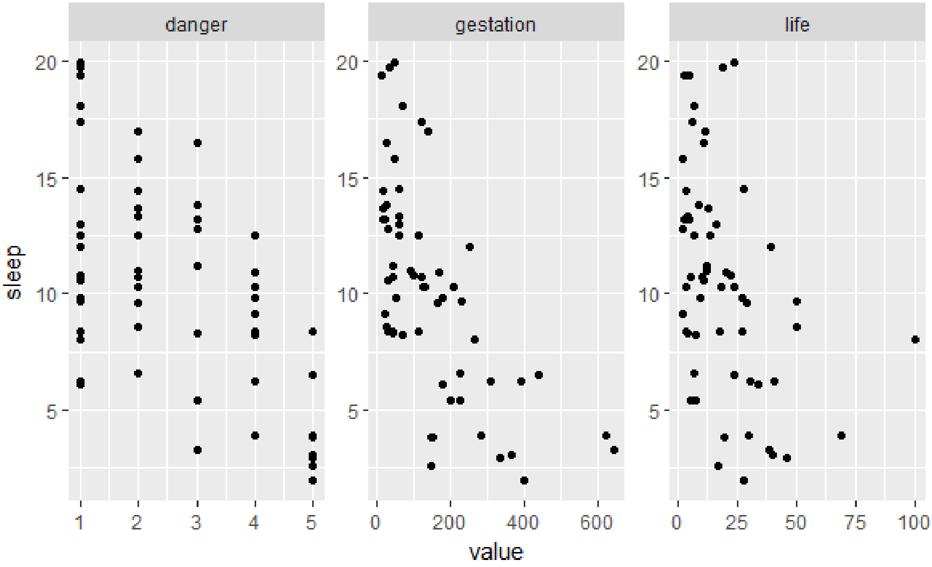

Figure 1: Relationships between sleep and danger, gestation and life.

2.2 Relationships between variables

Now the initial preparations are finished and technically a model can be built. Nevertheless, before training the model it is recommendable to get to know the data and explore relationships between features. This knowledge gained is of great help when interpreting the result and when deciding which features to use for training the model. By creating a scatter plot for *sleep* vs. all other variables, an initial sense of these relationships can be gained. Specifically, a negative correlation between *sleep* and the variables *gestation*, *danger*, and *life* can be observed, suggesting that being asleep is disadvantageous to animals that are more likely to be in danger, have a long gestation time, and live for longer. This is expected as sleep increases a mammals vulnerability, so those most likely to be in danger would be at high risk when asleep and so sleep less. Furthermore, this may be linked to the correlation between gestation time and sleep. Mammals more exposed to danger may have longer gestation times so that their offspring is more developed at birth, and hence more likely to survive, and such mammals sleep less. In addition to this, short lifespan may be an effect of a large amount of sleep due to the vulnerability sleep causes, and so it is not clear which variable is causal in this relationship. However, this explanation is speculative and would require further research to support it. The negative correlations between these variables can be seen in Figure 1.

From running the correlation command, strong correlations between other variables are observed that give further insight into the sleep behaviour of mammals. From Figure 2, we can see that mammals that are more exposed when they sleep require more slow-wave sleep than dream sleep. This suggests that different mammals will not only require different amounts of sleep, but different types of sleep. This is also to be expected, as mammals are less responsive when in dream sleep, so this would be disadvantageous for animals with high sleep exposure as they would need to react quickly to danger. Relationships between variables and the implications will be further addressed in section 3.

3 Regression

3.1 Variable reduction

Before fitting a regression model to the data, the variables are studied. As the objective is to predict the total hours of sleep different mammals need, and the response, sleep, is simply the sum of slow wave sleep and and dream sleep, it is possible to immediately discard these two variables. Looking at the correlation values of the remaining ones, it is observed that the danger and predation levels are highly correlated. The same holds for body and brain weight. Since collinearity reduces the accuracy of the regression coefficient estimations [1], one variable from each pair is eliminated.

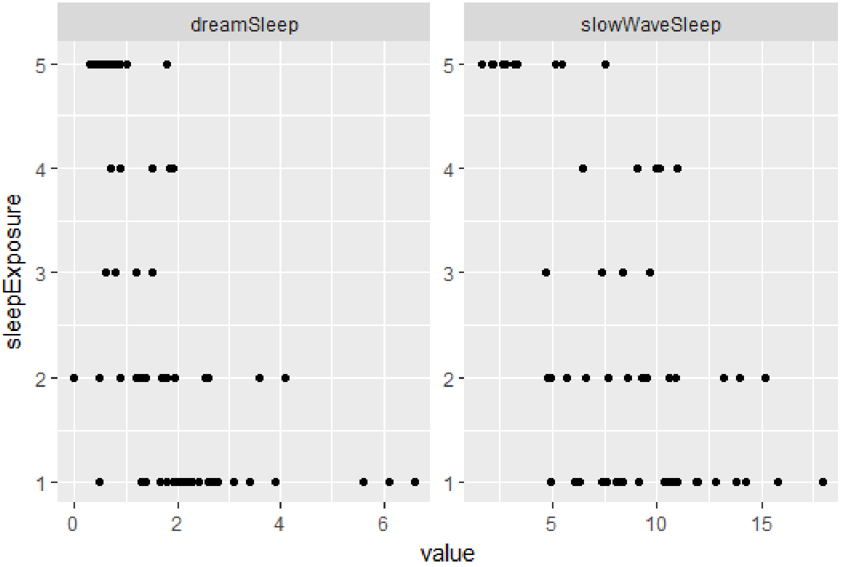

Figure 2: Relationships between sleep exposure and different types of sleep.

The number of possible predictors has now been reduced from 9 to 5 variables, with the remaining ones being brain weight, life span, gestation time, sleep exposure and danger.

3.2 Finding a model

Although the number of variables has been nearly halved, it is possible that not all are relevant for the prediction. A first attempt at finding a model is done by going through all possible variable combinations, using a minimum of 1 and a maximum of 5 predictors. The MSE is calculated through leave-one-out cross validation (LOOCV) of a linear model

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i x_i,$$

where y is the response, sleep, and x_i the predictors. β_0 is the bias and β_i are the predictor coefficients. If $\beta_i = 0$, the response does not depend on predictor x_i. Of course, this oversimplified model is not expected fully capture the essence of such a complex topic as mammal's sleep, but the inflexibility of it will prevent overfitting the small data set. We also note that three of our predictors are discrete. Assuming that the value scales are linear (e.g. a mammal with sleepExposure=4 is twice asexposed as a mammal assigned the value 2), however, they will not pose a problem for the linear regression. This since they will be treated the same way as continuous variables. There are $2^5 - 5 = 27$ possible combinations of zero and non-zero valued coefficients, all of which a LOOCV is performed on.

Figure 3 shows how the MSE depends on the different linear models. The models are not labelled on the x-axis, but the 5 models with lowest MSE, i.e. the best models, for a random training sample can be seen in table 1. From looking at their MSEs, we can see that the models appear to be equally good, as they all have an error around 10. These high errors can be explained by the small data set and oversimplified models, but they still hold valuable information in regards to which predictors are significant.

A linear regression with these five models shows that gestation time and danger level both have very small p-values, while the others have p-values$\gg 0.05$ in most of these regressions. This indicates that gestation and danger are significant predictors of sleep, while the others are not. Furthermore, model 2 in table 1 is the only one whose predictors are all significant. Thus, the model

$$\mathit{sleep} = \beta_0 + (\beta_1 \times \mathit{gestation}) + (\beta_2 \times \mathit{danger}), \tag{1}$$

with non-zero valued β_1 and β_2, is a reasonable choice.

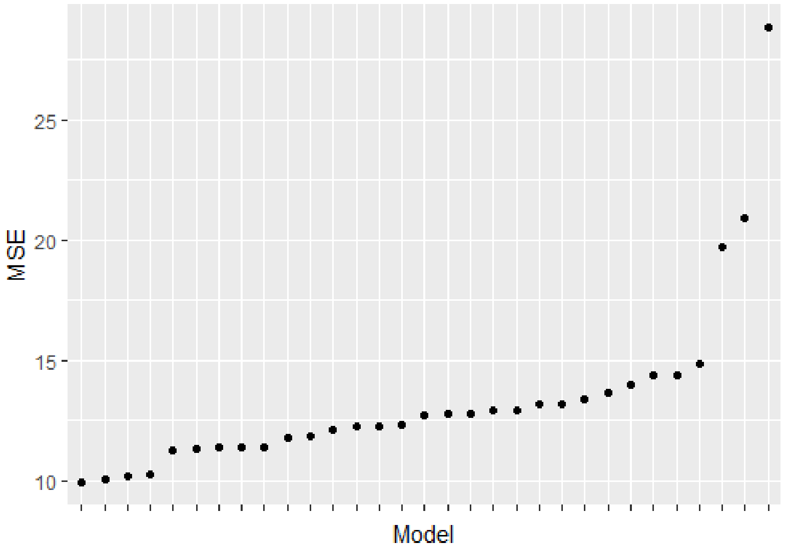

Figure 3: Mean square error of different multiple linear regression models, in ascending order.

Table 1: Models yielding the five smallest MSE by LOOCV, as seen in fig. 3.

Model no.	Predictors	Estimated test MSE
1	*life + gestation + danger*	9.90
2	*gestation + danger*	10.01
3	*life + gestation + sleepExposure + danger*	10.15
4	*gestation + sleepExposure + danger*	10.26
5	*life + sleepExposure + danger*	11.26

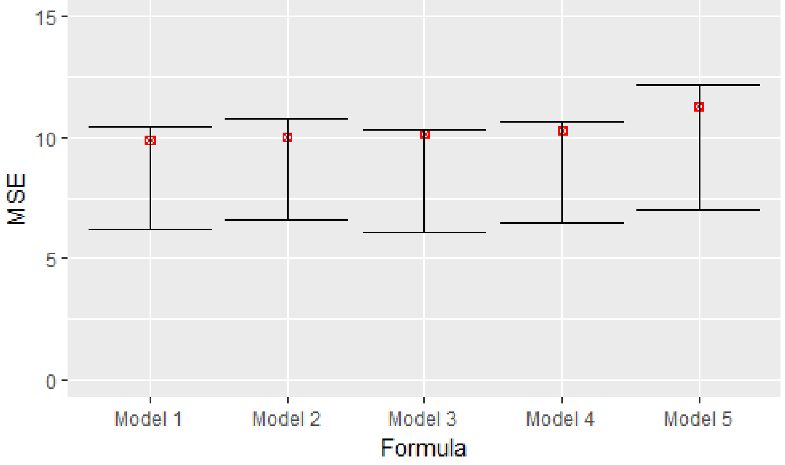

Figure 4: Predicted test MSEs for the different models in table 1. The red squares represent the LOOCV-estimated MSEs, and the error boxes show their 95 % confidence intervals.

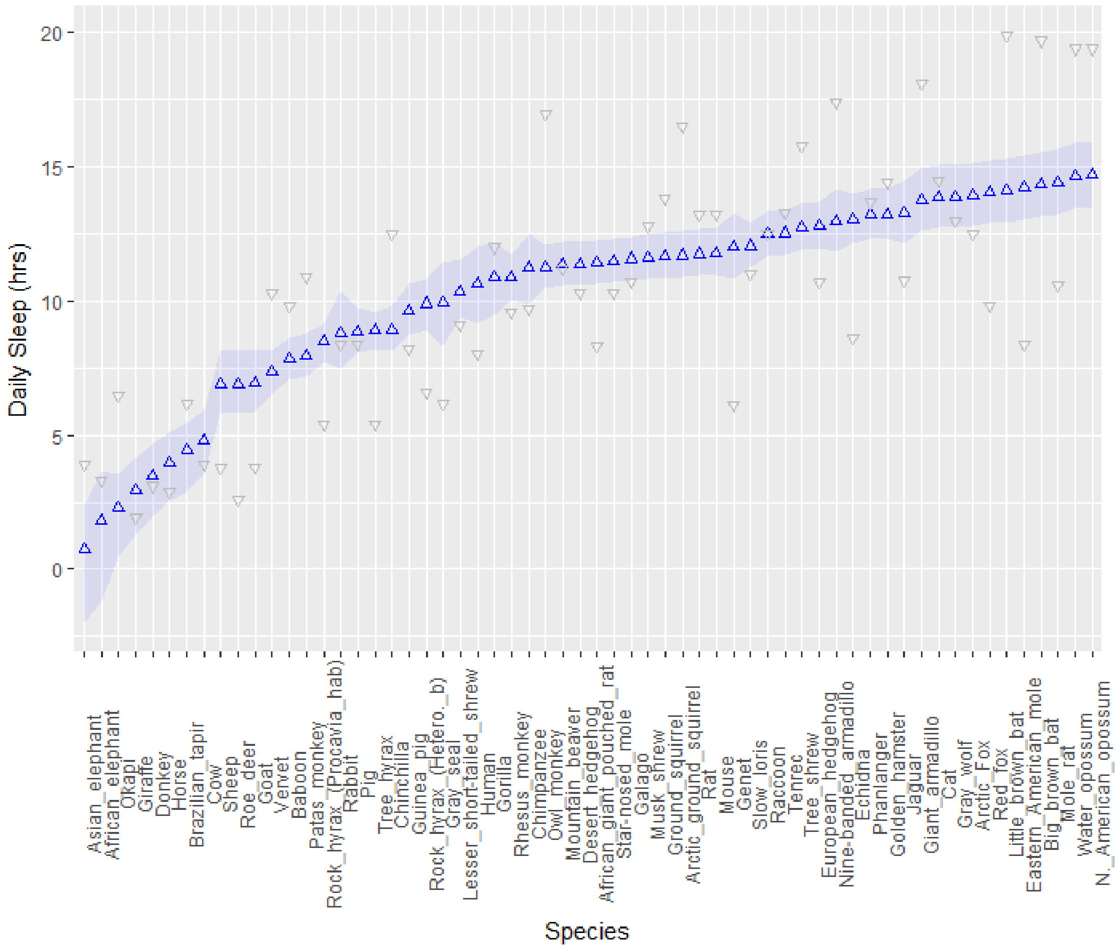

Figure 5: Daily sleep in hours, predicted by 10,000 bootstrap samples. The shaded area shows the 95% CIs of each prediction, and the gray triangles the actual sleep of each mammal, given by the original data set.

4 Quantification of uncertainty

Since the data points available might not be accurate representations of the daily sleeping requirements of all mammals on the planet, bootstrap is used to quantify this uncertainty. This is done by using 100,000 bootstrap samples to estimate the 95% confidence intervals of the predicted MSE for the models in table 1. Figure 4 shows the LOOCV-predicted MSEs with their bootstrapped confidence interval. It can be observed that the predicted MSEs lie in the upper range of the confidence interval. We also find that the confidence intervals range over approximately 4 error points each, showing a big variation in predicted test MSE. However, the CIs still seem to follow the general fitness of the models and does not hint at any disqualification of our choice of model (model 2 in table 1 and fig. 4, defined in eq. (1)).

As a final evaluation of the model accuracy, the sleep of each of the species in the data set is predicted using 10,000 boostrap samples and our chosen model, eq. (1). The results are shown in fig. 5. The confidence intervals from these bootstrap predictions are also included, as is the real sleep data. It is apparent that most real sleep data points lie outside the specified CIs, but they still seem to follow the overall trend somewhat. The spread of real sleep data points is also smaller for the shorter sleeping mammals than for the long sleepers. This might be an indicator that these points have high leverages, and thus might have influenced the regression more.

5 Conclusion

Our findings showed that the regression model in eq. (1) yielded a large predicted test error. Even if in the lower range of the wide confidence interval, it is still too huge to fully accept the current regression model. Thus, the task of finding a well-fitting model has just begun. The choice of model is one of the key factors of a good regression, but there are also others that can further be tweaked. For instance, more data exploration could be useful to find and treat high leverage points and outliers. Another suggested field to look into is adding categorical variables to the data set. Examples of such variables are diets and habitats.

One of the assumptions of multiple linear regression is no or low collinearity. Although highly correlated variables have been removed, it is possible for the variables to exhibit multiple correlation. Further exploration in this topic would benefit the regression.

The other linear regression assumptions are an underlying linear relationship, multivariate normality and homoscedasticity [1]. Exploring these further might indicate the appropriateness of choosing a linear regression model instead of a nonlinear one.

We still conclude that a simple linear regression model is a good start, especially when using a small data set. For the linear model, the total sleeping hours seems to depend mainly on the gestation time and how dangerous a mammal's life is. Other predictors might prove significant if a more flexible model is fitted, including e.g. polynomial and logarithmic terms. These, however, risk to overfit the data set. More robust analysis will have to wait for a much larger data set to appear.

Overall, from the data set available, it is apparent that the sleeping requirements of mammals are heavily determined by gestation time and danger. To simplify, mammals that are at high risk of attack sleep less in order to avoid danger. While we can confidently conclude that mammals that have longer gestation periods sleep less, the reason for this sleep behaviour cannot be determined from the information available in this data set.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani *An Introduction to Statistical Learning*. Springer 2017.

[2] Andrew Gelman, Jennifer Hill *Data Analysis Using Regression and Multilevel/Hierachical Models*. Cambridge University Press, 2006.

A Individual contributions

All group members have been active in all parts of the projects, and discussions and conclusions have been reached together. However, the main responsibility of different parts of the project has been split within the group. Eliza was responsible for the regression part. This includes finding and removing highly correlated variables, finding a fitting model, performing leave-one-out cross validation and bootstrap estimations. Thus, sections 3 and 4 have been authored by her, along with a big part of the conclusion. Sarah and Alexander authored the Introduction, the Data exploration and the data preparation in sections 1 and 2. The methods explained in section 2.1 was carried out by Alexander.