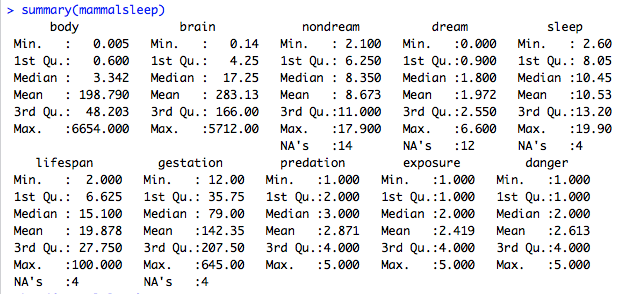
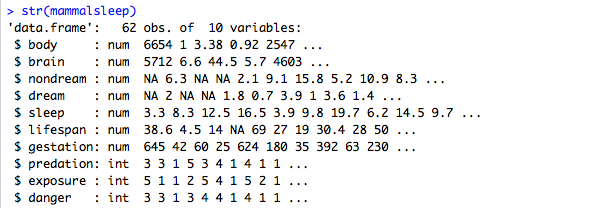
Sleep and Predation in Mammalsleep dataset

The following table explains the meaning and measurement units of each variable.

|  |  |
| --- | --- |
| body | body weight in kg |
| brain | brain weight in g |
| nondream | slow wave ("nondreaming") sleep (SWS) (hrs/day) |
| dream | paradoxical ("dreaming") sleep (PS) (hrs/day) |
| sleep | total sleep (hrs/day) (sum of slow wave and paradoxical sleep) |
| lifespan | maximum life span (years) |
| gestation | gestation time (days) |
| predation | predation index (1-5) 1 = minimum (least likely to be preyed upon) to 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) |

First of all, I’d like to look at the summary of the data.



head(mammalsleep)

plot(mammalsleep)

We need to remove datapoints with NAs in them before further analysis. To do so, I use na.omit function.

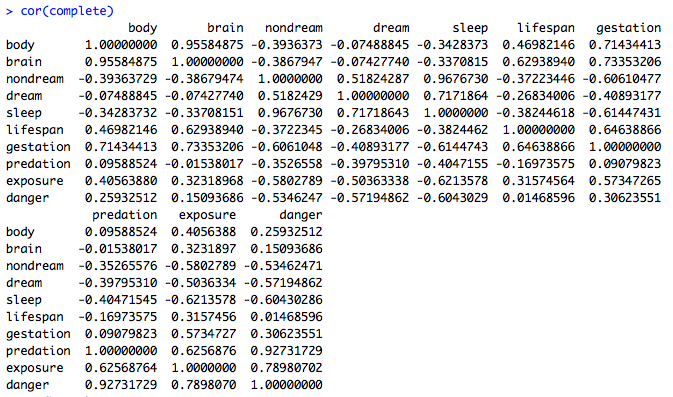
complete = na.omit(mammalsleep)

The “complete” is a subset of “mammalsleep” that only contains rows with no missing values. “complete” has 42 observations.

We see that sleep variable is the sum of SWS(non-dream) and PS(dream). Since our analysis focuses on the relationship between sleep and predation, we want to use only one of “sleep”,”dream”,”nondream” in our model at a time to avoid collinearity.

In my following regression analysis, I decide to choose “sleep”/”dream”/ “nondream” as dependent variables and treat “predation” as an explanatory variable.

Figure



Figure



First, let me check the correlation matrix of the 10 variables. Figure shown above.

“predation” is highly correlated with “danger” variable ; and also semi-highly correlated with “exposure”. So we say these three variables are collinear. Collinearity among explanatory variables is a problem because it masks the importance of each variable Xi in determining Y(dependent variable), so we don’t know how much value the added variable truly contributes to the effect, if we perform stepwise regression by forward selection.

On the flip side, by removing one of correlated explanatory variables, say Xi, from the model, the variation in Y caused by Xi will be inappropriately attributed to the remaining correlated variable Xj (j<>i). Xj’s coefficient will be biased. This phenomena is called the omitted variable bias.

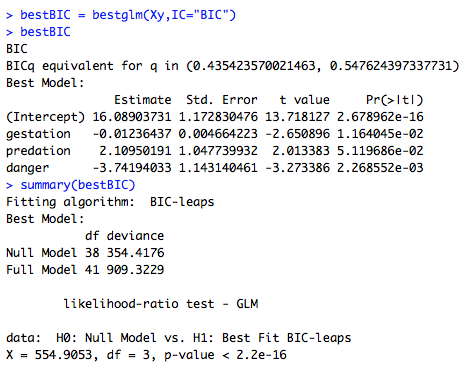
So it seems it’s inappropriate to either use all or removing some correlated variables. I have a dilemma here.

Instead of manually selecting variables, I resort to bestglm function and use BIC as selecting criterion. I use “sleep” as dependent variable, and include “body”, “brain”, “lifespan”, “gestation”, “predation”, ”exposure” and “danger” as possible predictors for the bestglm function to choose from.

Figure



Figure



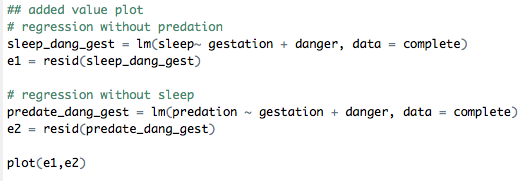
As seen from above, the function automatically picked gestation, predation and danger as predictors.

The fitted model is sleep = -0.0123 gestation + 2.1095 predation - 3.742 danger + 16.089

Gestation and danger are negatively associated with sleep, while predation is positively associated. Note that the earlier correlation matrix shows -0.405 correlation between predation and sleep, but here the model shows they are positively correlated. This seeming anomaly is because correlation measures relationships only between pairs of variables, but the regression fit can account for several variables simultaneously. To illustrate it more visually, I am making an added variable plot.

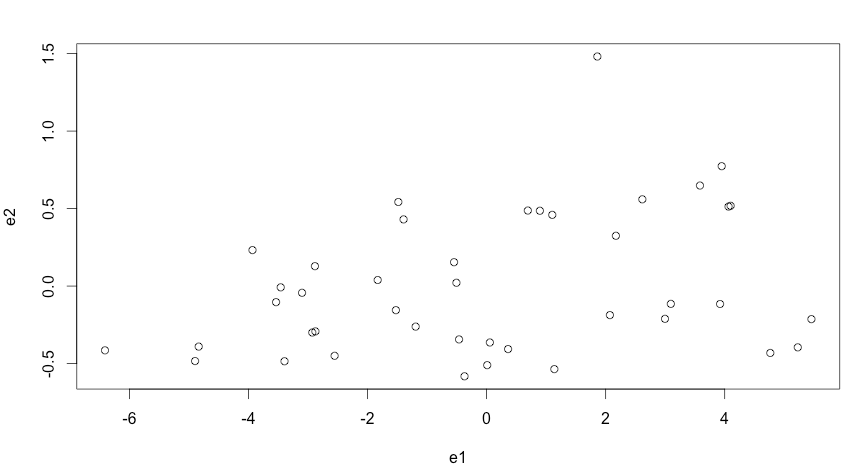
First I fit a regression model of sleep on gestation and danger, excluding predation. I get its residuals1. Then I fit a regression model of predation on gestation and danger, excluding sleep, and I get its residuals e2 By plotting the two residuals, I can see the relationship between sleep and predation, with other variables controlled.

Figure



The plot looks like below. We can clearly see a generally positive relationship.

Figure



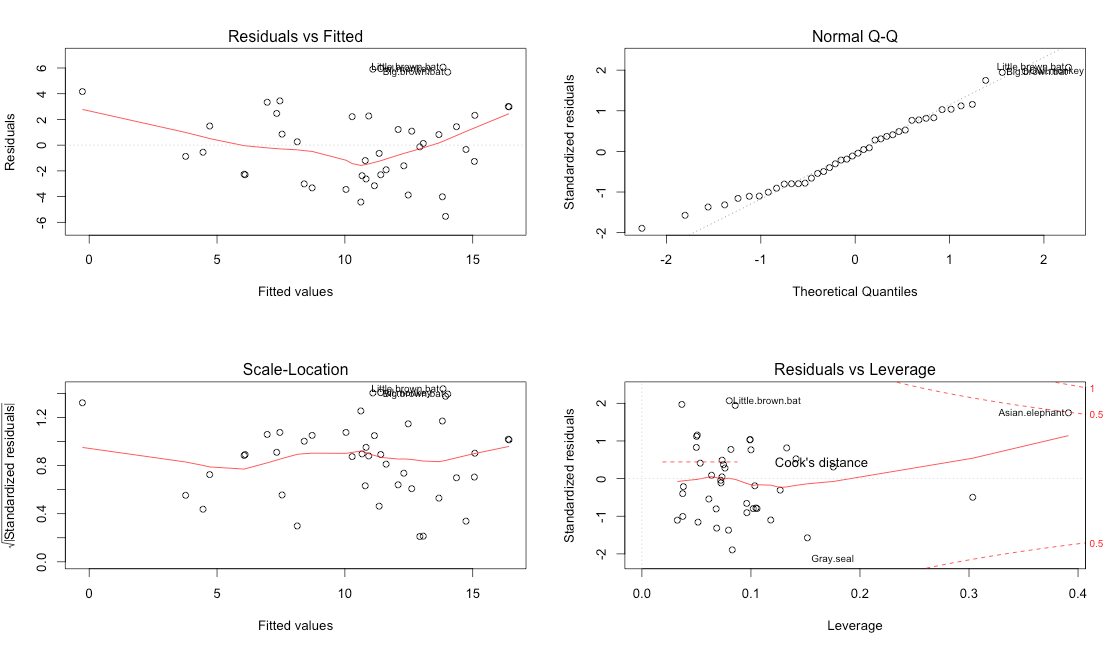
Now that we have a model, I want to check the validity of our model assumptions.

best = lm(sleep~ gestation + predation + danger, data= complete)

par(mfrow = c(2,2))

plot(best)

Figure



This result is less than satisfactory. The residual v.s. fitted plot has data crowed on the right hand side and show a v shaped pattern. This plot along with scale location plot and QQ plot identify “little brown bat”, “big brown bat”, “owl monkey” as outliers. Cooks’ distance plot identifies “Asian elephant” as a high leverage point. The QQ plot is not linear, especially in the lower quantile portion.

Let me remove “little brown bat”, “big brown bat”, “owl monkey” and “Asian elephant” to see what the fit looks like.

I perform the following code.

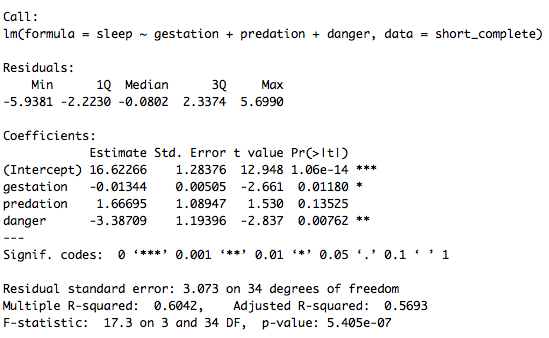
short\_complete = complete[-c(5,7,33,42),]

short\_best = lm(sleep~ gestation + predation + danger, data = short\_complete)

summary(short\_best)

The result is shown in the figure below.

Figure



The absolute value of t value became smaller for most explanatory variables after removing the outliers and high leverage points. The R squared is also smaller. Both facts suggest that this truncated model is not as good as the earlier one containing all data points.

Further discussion

The bestglm function with BIC selecting criteria selects “gestation”, “predation” and “danger” to explain “sleep”. All three variables are statistically significant, as shown in Figure 4. Although “sleep” and “predation” have negative correlation coefficient as seen from the correlation matrix. The regression model indicates a positive relation between them. That is, even though we observe higher chance to be preyed upon is associated less hours of sleep among the species, this negative relation is caused by other factors influencing “sleep” simultaneously. Our model takes this into account and suggests the negative relation comes from “danger” and “gestation”. When we keep “danger” and “gestation” constant, “predation” exhibits positive relation with “sleep”, demonstrated by the added value plot. This result suggests that the more danger a species faces, the less sleep it gets. It’s easy to understand because the species in danger tends to sacrifice sleep time to watch out the surroundings. The longer gestation time, the less sleep can also be explained by the “danger” variable. During gestation, females are more vulnerable, exposed to more danger than their normal condition. So they tend to sleep less to keep safe. With the same level of danger and gestation time, higher chance to be preyed upon can be a result of more sleep time. But this causal relation is beyond the scope of this report.