A Linear Model For Predicting Sleep Time in Mammals

Madison Kellar and Nick Koprowicz

December 9, 2015

Introduction

This project required analysis of a real data set. We chose "Sleep in Mammals: Ecological and Constitutional Correlates" [1]. Contained in this data set are observations of 62 mammals on 10 variables. Variables are detailed in the table below.

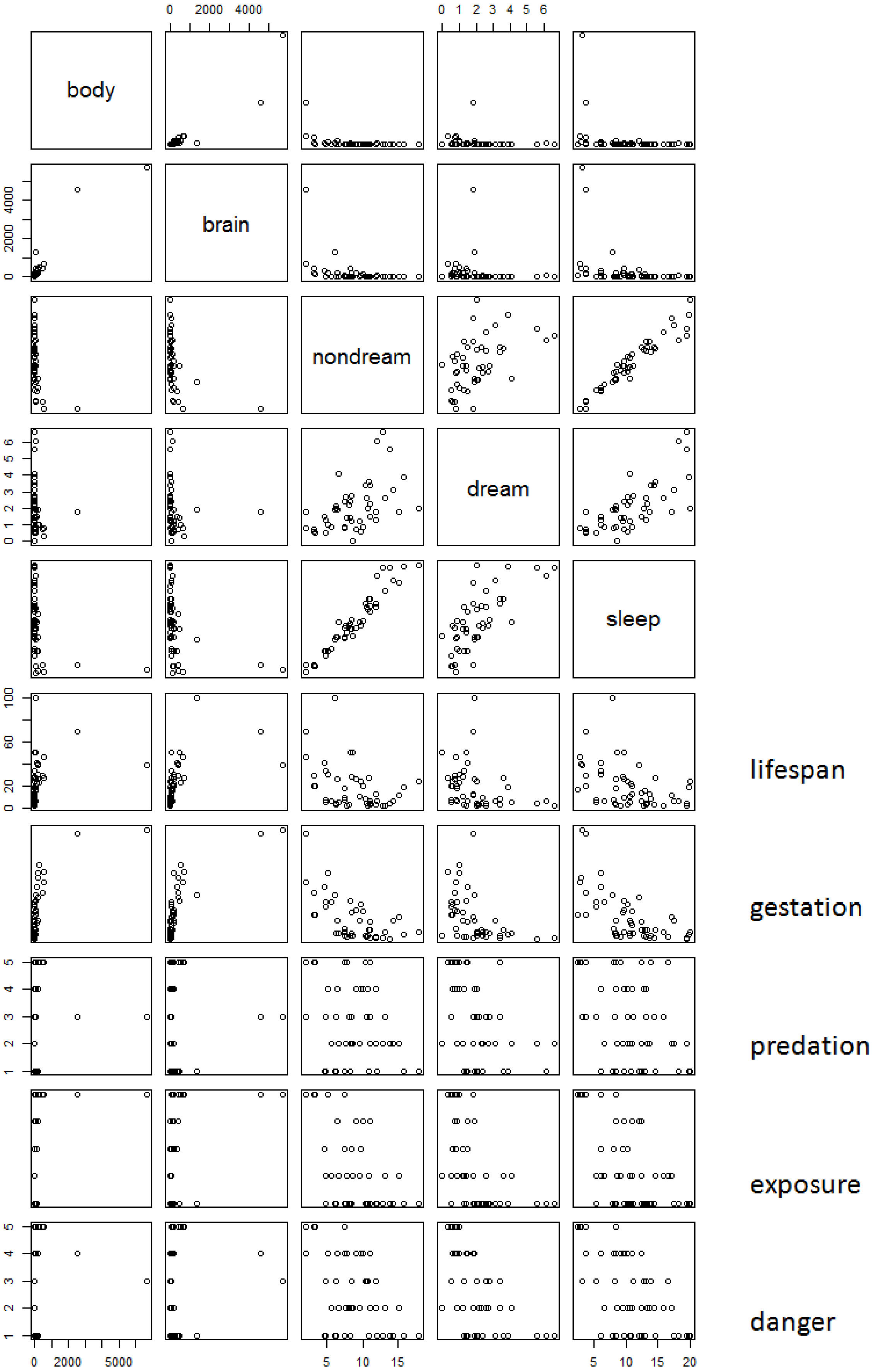
Variable	Measure
body	body weight in kg
brain	brain weight in g
nondream	slow wave ("nondreaming") sleep (hrs/day)
dream	paradoxical ("dreaming") sleep (hrs/day)
sleep	total sleep (hrs/day) [sum of slow wave and paradoxical sleep]
lifespan	maximum lifespan (years)
gestation	gestation time (days)
predation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

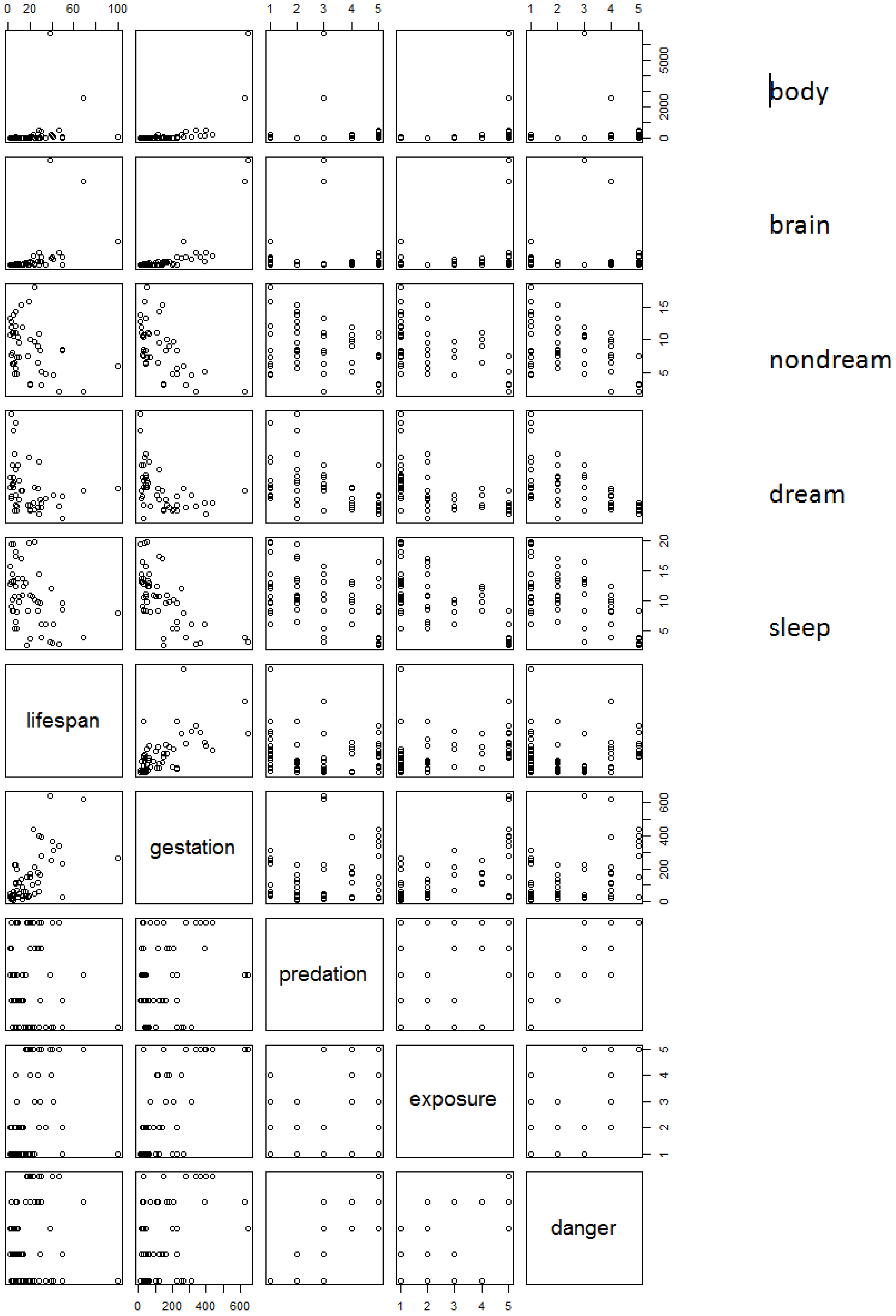
From these observations, we want to investigate modeling the response variable, sleep, as a function of some or all of the other variables. In order to do this, we construct scatterplots of each combination of variables to examine the relationships between them. We consider transformations and interactions as possible variables in the model as well. We then use several methods to decide which variables to include in the linear model.

Model Development

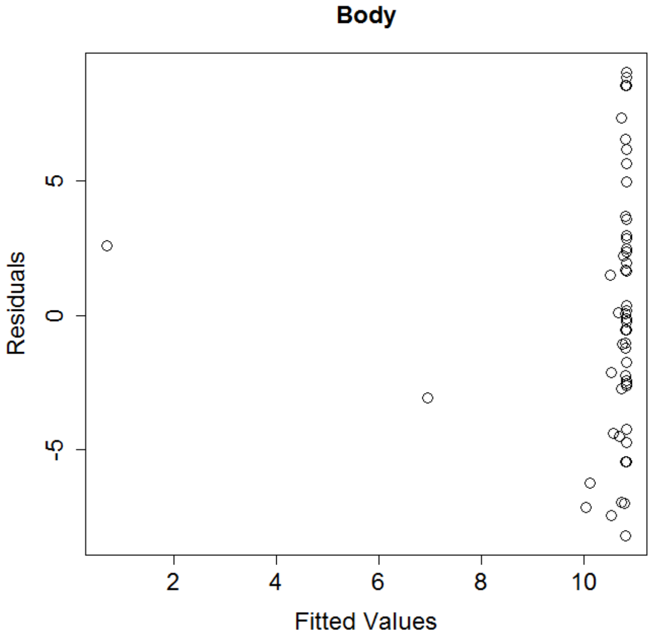
Scatterplots

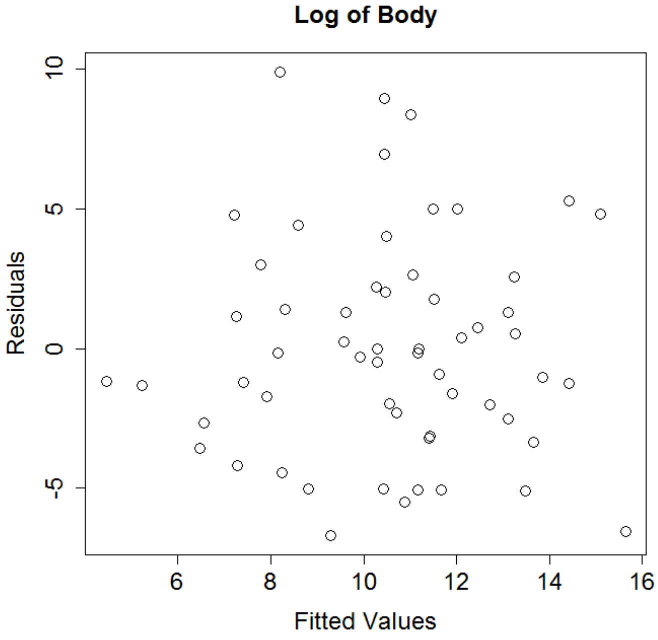
We begin by looking at scatterplots of each pair of variables. For variables that do not appear to have a linear relationship with sleep, we use the scatterplots as a guide to select transformations. Recognize that the sum of the variables dream and nondream is equivalent to sleep. Therefore, we cannot include both variables in our model, as that would be using the response variable to predict itself. We choose to include nondream, rather than dream, since it appears to have a stronger linear relationship with sleep.


body
brain
nondream
dream
sleep
lifespan
gestation
predation
exposure
danger
0 2000 4000
0 1 2 3 4 5 6
0 2000 4000
0 1 2 3 4 5 6
0 20 60 100
1 2 3 4 5
1 2 3 4 5
0 2000 5000
5 10 15
5 10 15 20


0
20
60
100
1
2
3
4
5
1
2
3
4
5
0
2000
5000
body
brain
5
10
15
nondream
dream
5
10
15
20
sleep
lifespan
gestation
0
200
400
600
predation
exposure
1
2
3
4
5
danger
0
200
400
600
1
2
3
4
5

From these scatterplots, nondream is the only variable that appears to have a strong linear relationship with sleep. Therefore, we look at several transformations, including powers and logarithms of variables. It is not clear from the scatterplots that any of the variables have an exponential relation to sleep, but we decide to check this transformation anyway. We try square roots and integer powers. None of the scatterplots indicate that any powers of variables would be useful in predicting sleep, so we do not consider these when constructing the model. However, we find a correlation between the logarithm of some variables with sleep. For example, the scatterplot of sleep and log(body) appears to show a negative correlation, which, while not strong, qualifies the variable to be considered in construction of a linear model. These variables produce homoscedastic residual plots, further qualifying them to be considered. The residual plots from the variable body and its logarithm are below. We show this variable because it is the only logarithm of a variable that appears in the final model. It is clear that the homoscedastic plot on the right is likely to be better in predicting sleep than the left plot. For the same reasons, in addition to body, we consider the logarithms for brain, lifespan and gestation during model selection.





Variable Selection

In order to construct our model, we use three methods of model selection as guides: stepwise regression, best subsets regression, and the Akaike Information Criterion (AIC). Stepwise regression ultimately led us to our model; the other methods backed up our decision. Note that we use these methods only as useful tools, not as absolute rules.

First of all, we see which, if any, of the indexed variables might be useful in predicting sleep. We fit a linear model to a wide range of variables to start a stepwise selection process. Within the model, we include all three indexed factors: predation, exposure, and danger. Predation and exposure do not appear to have any effect in the model, based on p-values, but danger seems like it could be useful to include in the model. We decide to exclude predation and exposure as factors as we move forward with the selection process.

We proceed with backward selection steps - removing variables from our model whose p-values are significantly high. Following the exceptions to the principle of parsimony, if an interaction is included in a model, we include each individual variable as well. This influences our process as we do not always remove the variable with the highest p-value if the removal will violate this criteria. As we eliminate variables from the model, the p-values for the danger index continue to decrease, reinforcing our decision to include it in the model as a factor.

After six steps of backward step-selection, we arrive at the following model:

$$\text{sleep} = \beta_0 + \beta_1 \text{ nondream} + \beta_2 \text{ lifespan} + \beta_3 \text{ gestation} + \beta_4 \text{ lifespan*gestation} + \text{factor(danger1)} \\ + \beta_5 \text{ factor(danger2)} + \beta_6 \text{ factor(danger3)} + \beta_7 \text{ factor(danger4)} + \beta_8 \text{ factor(danger5)}$$

This model has the following estimates and p-values.

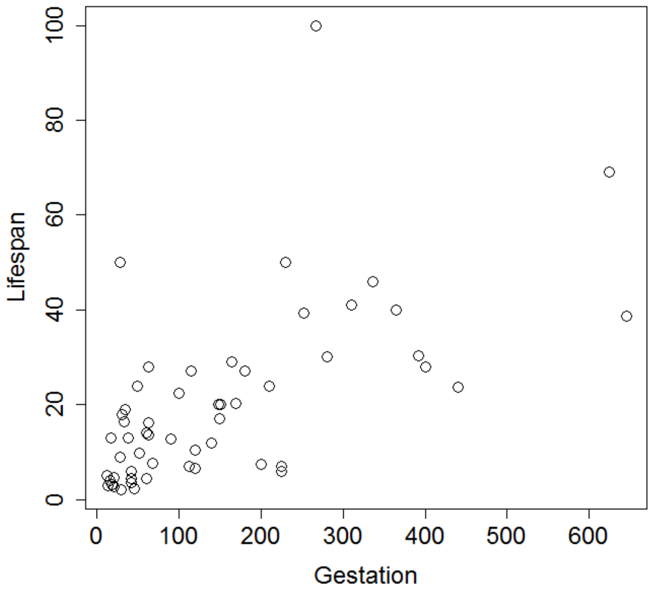
Coefficients	Estimate	P-value
Intercept	3.591e+00	1.81e-05
nondream	1.115e+00	< 2e-16
lifespan	-6.934e-02	4.89e-06
gestation	-1.137e-02	5.06e-05
logbody	2.681e-01	0.001807
lifespan*gestation	2.272e-04	9.30e-06
factor(danger)2	-1.049e+00	0.007271
factor(danger)3	-1.325e+00	0.006213
factor(danger)4	-1.623e+00	0.000273
factor(danger)5	-1.813e+00	0.001766

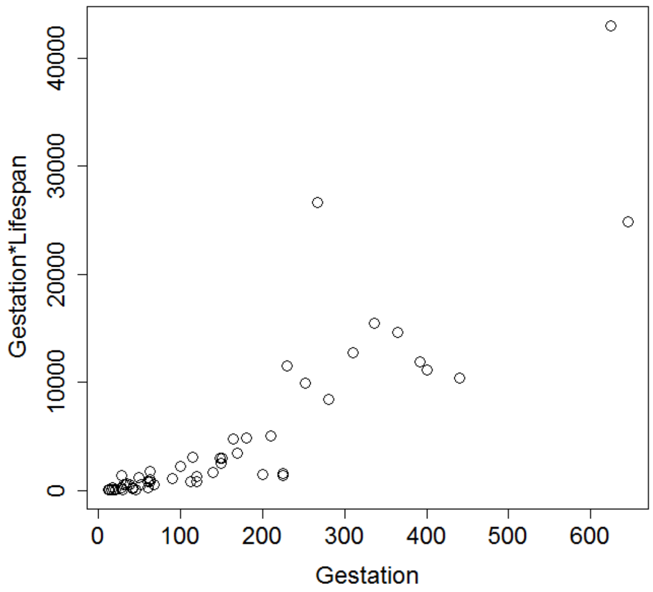
The small p-values for each of these variables indicate the model fits well. We then move on to other methods to either confirm that the model we arrived at is good, or suggest ways to improve it. We run a best-subsets regression, again considering all three possible factors. We see that the best 5-variable model contains the same variables as our 6-variable model, missing only danger as a factor. None of the other factors, predation or exposure, were chosen frequently for any subset. So we decide that our model will have danger as a factor and no others, as we did before, and do the best-subsets regression again excluding the factors. We see that, per Mallow's CP, the best model is the 5-variable model, which has all the same variables as the model we arrived at through stepwise regression, without the danger index.

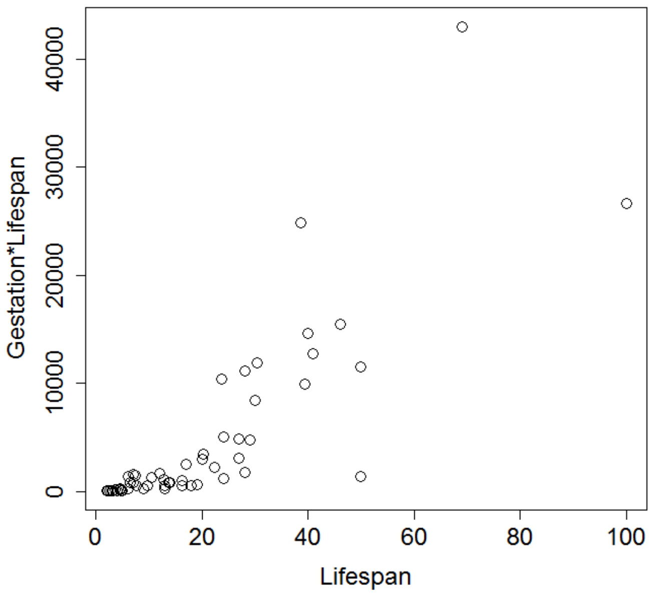
As another measure, we try the AIC method with the same variables as our other methods, but excluding the factors. The AIC recommends a model with only an intercept and nondream as a variable. While nondream alone is a good predictor of sleep, we have seen that several other variables help to predict the model better, while retaining low p-values. We decide to move forward with the model we had arrived at earlier.

Alongside our model techniques, we note that when using model selection techniques with a large number of variables, a model might be found that should not be. The more candidate variables there are, the greater the likelihood of finding correlations just by chance. Now, we did not use too many variables when checking the model with best subsets and AIC, but just in case, we select a model using one more method. This time, we start with one variable, nondream, and add variables into the model until it is as good or close as the model we have. Doing this, we ended up with the same model. This is the model we will choose.

Before moving on to analyzing our model, we check for collinearity between variables. Of particular concern is that our model includes gestation, lifespan, and gestation*lifespan, and there might be collinearity between them. We make scatterplots of each pair of these variables, and none of them show a strong linear relationship.



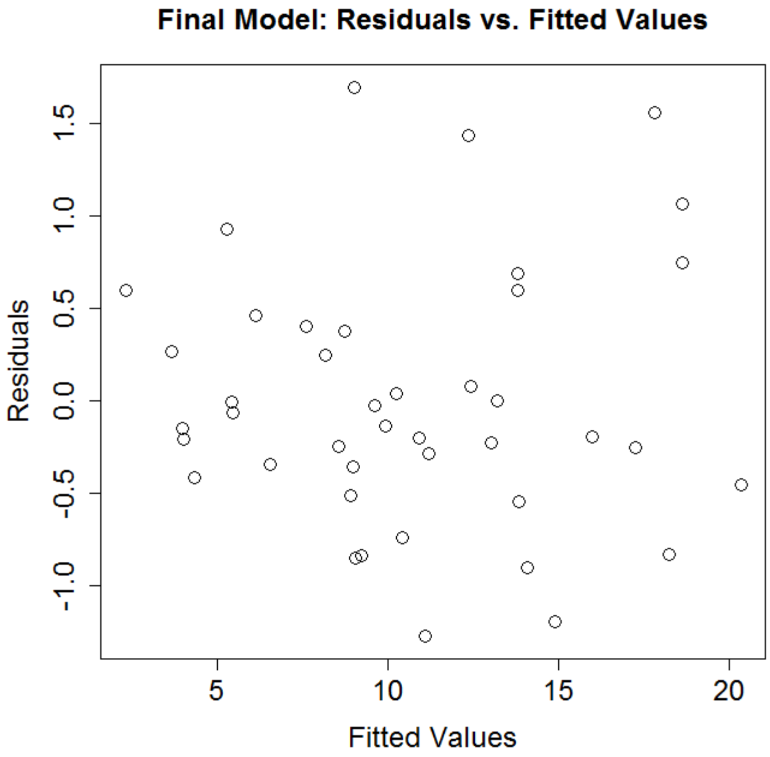




Even though it did not appear that there was collinearity, we try building a model without lifespan*gestation at all. We found that the p-values for the other variables increase significantly. Now reassured that there is no collinearity and that the interaction of lifespan*gestation is a necessary component of the model, we move on to analysis.

Analysis

We start by plotting residuals versus fitted values for our selected model. The plot is homoscedastic and shows no significant trend or curve. Therefore, it is likely that the assumptions of a linear model hold. We can proceed with analyzing our model under these assumptions.



First, the scatterplot does not appear to have any outliers. Nest, we identify any high-leverage points. To do this, we compute the diagonals of the hat matrix, **H**. We find that two mammals have high leverage: the Asian elephant and Man. After that, we identify any influential points using Cook's distance. There is only one variable with a Cook's distance above a threshold of 0.5, that is the Asian elephant with Cook's distance = 0.6084338. This is not surprising, as the Asian elephant is the largest mammal in the data set. We fit the model without this influential point, to see if there are significant changes. We include the estimates and p-values for the model with and without the influential point in the table below, in order to compare.

	Without Influential Point		With Influential Point	
Coefficients	Estimate	P-value	Estimate	P-value
(Intercept)	3.593e+00	1.75e-05	3.591e+00	1.81e-05
nondream	1.123e+00	< 2e-16	1.115e+00	< 2e-16
lifespan	-8.290e-02	2.62e-05	-6.934e-02	4.89e-06
gestation	-1.235e-02	3.28e-05	-1.137e-02	5.06e-05
logbody	2.742e-01	0.001437	2.681e-01	0.001807
lifespangestation	2.998e-04	0.000299	2.272e-04	9.30e-06
factor(danger)2	-9.443e-01	0.016807	-1.049e+00	0.007271
factor(danger)3	-1.275e+00	0.008132	-1.325e+00	0.006213
factor(danger)4	-1.516e+00	0.000725	-1.623e+00	0.000273
factor(danger)5	-1.764e+00	0.002231	-1.813e+00	0.001766

The estimates have not changed significantly by removing the Asian elephant from the data set. Consequently, the least-squares line of our model will not change substantially and this point is not considered influential.

Summary

In conclusion, sleep response time in mammals can be predicted from the following explanatory variables: lifespan, gestation, the interaction of the two, nondream, the logarithm of body, and the danger factor. These variables seem logical when we consider their physical properties. Mammals that have a larger body mass typically have longer gestation times, and so possibly need more sleep. Gestation times seem to interact with lifespan, so more sleep per day could mean a longer life. Since nondream and dream sum to sleep, it is likely that one would be appropriate for predicting sleep. Finally, mammals that in more danger from other animals may need to stay awake and alert in order to survive. We note that model selection is an art, not a science, so there are likely several models that predict sleep about equally well. However, based on the analysis, our model predicts sleep well.

Bibliography

[1] Sleep in Mammals: Ecological and Constitutional Correlates by Allison, T. and Cicchetti, D. (1976), Science, November 12, vol. 194, pp. 732-734.

Code

```
IMPORT DATA
setwd("Z:/")
data = read.table("mammalsleep.dat.txt")

body <- data[,1]
brain <- data[,2]
nondream <- data[,3]
dream <- data[,4]
sleep <- data[,5]
lifespan <- data[,6]
gestation <- data[,7]
predation <- data[,8]
exposure <- data[,9]
danger <- data[,10]

#####################################################################
TRY TRANSFORMATIONS
NONE OF THE TRANSFORMATIONS WITH POWERS SEEMED TO MAKE ANYTHING BETTER
NOTE: WE EXCLUDED THE INDEXED VARIABLES

body2 <- body^2
brain2 <- brain^2
nondream2 <- nondream^2
dream2 <- dream^2
sleep2 <- sleep^2
lifespan2 <- lifespan^2
gestation2 <- gestation^2

plot(sleep ~ body^2) #NO CHANGE
plot(sleep ~ brain2) #WORSE
plot(sleep ~ nondream2) #SAME
plot(sleep ~ dream2) #NO CHANGE
plot(sleep ~ lifespan2) #WORSE
plot(sleep ~ gestation2) #WORSE

sqrtbody <- sqrt(body)
sqrtbrain <- sqrt(brain)
sqrtnondream <- sqrt(nondream)
sqrtdream <- sqrt(dream)
sqrtsleep <- sqrt(sleep)
sqrtlifespan <- sqrt(lifespan)
sqrtgestation <- sqrt(gestation)

plot(sleep ~ sqrtbody)
plot(sleep ~ sqrtbrain)
plot(sleep ~ sqrtnondream)
plot(sleep ~ sqrtdream)
plot(sleep ~ sqrtlifespan)
plot(sleep ~ sqrtgestation)

###############################################################################
CONSIDER LOG TRANSFORMATIONS
AND INTERACTIONS

logbody <- log(body)
logbrain <- log(brain)
lognondream <- log(nondream)
logdream <- log(dream)
logsleep <- log(sleep)
loglifespan <- log(lifespan)
loggestation <- log(gestation)

bodybrain = body*brain
bodynondream = body*nondream
brainnondream = body*nondream
lifespangestation = lifespan*gestation

###########################################################################
STEPWISE SELECTION CONSIDERING EXPOSURE,
PREDATION, AND DANGER AS FACTORS

model = lm(sleep ~ body+brain+nondream+lifespan+gestation+logbody+logbrain+
lognondream+loggestation+bodybrain+brainnondream+lifespangestation+
factor(predation)+factor(exposure)+factor(danger))

summary(model)

#### 1. PREDATION AND EXPOSURE ARE SMALL DIFFERENCE.  TAKE THEM OUT

model = lm(sleep ~ body+brain+nondream+lifespan+gestation+logbody+logbrain+lognondream+
loggestation+bodybrain+brainnondream+lifespangestation+factor(danger))

summary(model)

#### 2. DANGER SEEMS TO BE RELEVANT.  START REMOVING THINGS WITH HIGHER P-VALUES
#### REMOVE BODY, BRAIN-NONDREAM

model = lm(sleep ~ brain+nondream+lifespan+gestation+logbody+logbrain+lognondream+
loggestation+bodybrain+lifespangestation+factor(danger))

summary(model)

#### 3. TAKE OUT BRAIN AND LOGNONDREAM

model = lm(sleep ~ nondream+lifespan+gestation+logbody+logbrain+loggestation+bodybrain+
lifespangestation+factor(danger))

summary(model)
```

```
#### 4. TAKE OUT LOGBRAIN

model = lm(sleep ~ nondream+lifespan+gestation+logbody+loggestation+bodybrain+lifespangestation+
factor(danger))

summary(model)

#### 5. TAKE OUT BODY-BRAIN, SINCE NEITHER OF THESE IS IN THE MODEL INDIVIDUALLY

model = lm(sleep ~ nondream+lifespan+gestation+logbody+loggestation+lifespangestation+factor(dang

summary(model)

#### 6. TAKE OUT LOG GESTATION.  IT DOESN'T HAVE THE HIGHEST P VALUE -
#GESTATION DOES - BUT LIFESPAN-GESTATION HAS A LOW P-VALUE

model = lm(sleep ~ nondream+lifespan+gestation+logbody+lifespangestation+factor(danger))

summary(model)

!!!!!!!!!!!!THE MODEL ABOVE IS A GOOD FINAL MODEL!!!!!!!!!!!!!!!!
#############################################################################

QUESTION: IS THERE COLLINEARITY BETWEEN LIFESPAN AND GESTATION? OR LIFESPAN AND
LIFESPAN-GESTATION OR GESTATION AND LIFESPAN-GESTATION?

CHECK WITH SCATTERPLOTS

plot(lifespan, gestation)                       NOT A STRONG LINEAR RELATIONSHIP
plot(gestation, lifespangestation) NOT A STRONG LINEAR RELATIONSHIP
plot(lifespan, lifespangestation) NOT A STRONG LINEAR RELATIONSHIP

CONCLUSION: NO, THERE DOES NOT APPEAR TO BE COLLINEARITY,
# BUT LETS TRY TAKING LIFESPAN-GESTATION OUT ANYWAY

model = lm(sleep ~ nondream+lifespan+gestation+logbody+factor(danger))

summary(model) NO GOOD.
############################################################################

NOW WE HAVE A PROPOSAL FOR A MODEL.  LET'S TRY
BEST SUBSETS REGRESSION CONSIDERING EXPOSURE, PREDATION, AND DANGER AS FACTORS

install.packages("leaps")

library(leaps)

best = regsubsets(sleep ~ body+brain+nondream+lifespan+gestation+logbody+logbrain+
lognondream+loggestation+bodybrain+brainnondream+lifespangestation+factor(predation)+
factor(exposure)+factor(danger), data = data)

best.sum = summary(best)

best.sum$which

cbind(best.sum$rsq, best.sum$adjr2, best.sum$cp)
```

WE SEE THAT THE BEST 5-VARIABLE MODEL CONTAINS THE SAME VARIABLES AS OUR 6-VARIABLE MODEL, EXCLUDING DANGER AS FACTOR.

BUT WE'VE ALREADY DECIDED THAT EXPOSURE AND DANGER AREN'T FACTORS AND THAT DANGER IS A FACTOR. LETS RUN BEST SUBSETS WITHOUT THESE AS FACTORS

```
best = regsubsets(sleep ~ body+brain+nondream+lifespan+gestation+logbody+logbrain+lognondream+
loggestation+bodybrain+brainnondream+lifespangestation, data = data)

best.sum = summary(best)

best.sum$which

cbind(best.sum$rsq, best.sum$adjr2, best.sum$cp)
```

THE BEST MODEL IS THE 5-VARIABLE MODEL
(PER MALLOW'S CP - IT'S THE ONE THAT'S CLOSEST TO THE NUMBER OF VARIABLES IN THE MODEL)
AND IT HAS ALL THE VARIABLES OUR MODEL HAS WITHOUT THE DANGER INDEX

AIC

```
fullModel = lm(sleep~body+lifespan+gestation+danger
+nondream+brainbody+bodynondream+brainnondream+predgest+preddanger+gestdangeR
+gestlife+predlife+explife+dangerlife+expgest+exppred+body_log)
modelZero = lm(sleep~1)
step(modelZero, scope = list(lower=NULL, upper=data.frame(cbind(body,brain,
lifespan,gestation,danger,nondream,brainbody,bodynondream, bodylife
nondreamlife,nondreamgest,nondreampred,nondreamdanger,gestlife,body_log)),
direction="both"))
```

FOR THE INDEXED VARIABLES, THE ESTIMATE IN THE 'R' CODE IS HOW MUCH THEY DIFFER FROM THE FIRST LEVEL OF THE INDEX. THEN WE USE VARIABLES FOR EACH LEVEL OF THE FACTOR THAT ARE EITHER 0 OR 1 DEPENDING ON THE LEVEL.

HERE'S WHAT THE EQUATION LOOKS LIKE WRITTEN OUT:

```
sleep = 3.591 + 1.1115*nondream - 6.934*lifespan - 1.137*gestation + .2618*log(body)
+ 0.0002272*lifespangestation+ danger1 -1.049*danger2 - 1.325*danger3
- 1.623*danger4 - 1.813*danger5.
```

So for example if the danger level is 1, then danger1 = 1 and danger2 = ... = danger5 = 0
If the level of danger is 2, then danger2 = 1 and danger1 = danger3 = ... = danger5 = 0.

```
model = lm(sleep ~ nondream+lifespan+gestation+logbody+lifespangestation+factor(danger))

plot(model, cex.lab=1.5, cex.axis=1.5, cex.main=2, cex.sub=1.5, cex=1.5)

# Compute high-leverage points
hatvalues(model)

# Compute Cook's Distance
cooks.distance(model)

#Fit model without influential points
data_rm5 <- data[-c(5),]
body <- data_rm5[,1]
nondream <- data_rm5[,3]
sleep <- data_rm5[,5]
lifespan <- data_rm5[,6]
gestation <- data_rm5[,7]
danger <- data_rm5[,10]
body_log = log(body)
danger_log = log(danger)
gestlife=gestation*lifespan

model_rm5 = lm(sleep ~ nondream + factor(danger) + logbody + lifespangestation +
gestation + lifespan)
Residuals = residuals(model)
fitt = fitted(model)
plot(fitt, Residuals, xlab="Fitted Values",
cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5, cex=1.5)

plot(model_rm5)

Res_rm5 = residuals(modelFinal)
Fit_rm5 = fitted(model_rm5)
plot(Fit_rm5,Res_rm5, cex.lab=1.5, cex.axis=1.5,
cex.main=1.5, cex.sub=1.5, cex=1.5, xlab="Fitted Values", ylab="Residuals")
```