# Discriminating smartness among mammals: relationships between dreaming, life duration, and safety of the environment

Project2 - Applied Regression Analysis

**Abstract**

Among mammals, many different characteristics are observed. In particular, explaining the causes of the differences in the interaction with the environment is a hard task. In this study, we try to make a connection between dreaming and behavior in mammals. We argue that time spent sleeping and dreaming in mammals can affect their habits, namely can lead them to be safer. A data analysis on a sample of 62 species is done, investigating also the variations in life span, sleep and dreaming time. We show that both sides of the causal implication are supported: higher exposure to the environment brings about lower sleep time; and lower sleep time brings about higher exposure to the dangers. Strong attention is payed to the dreaming habits of mammals. We show also that mammals with heavier brains live longer, and that not necessarily mammals more exposed to dangers live less. Some issues of causal inference are discussed and used in the determination of the conclusions.

## 1 Introduction

There are around 5000 species of mammals; with a wide range of features, shapes, sizes and behaviors.

What makes mammals similar are their four-chambered hearts, separate sexes with the sex of an embryo being determined by the presence of a XY or XX chromosomes, internal fertilization and highly developed brain.

Nevertheless, among the thousands species of mammals, extreme differences can be observed: some mammals are predators, some other are preys. Some spend all their life in the darkness of a cave, like the bats; and others never leave the water, like the dolphins. Great differences are observed also in the sleeping habits.

Though one would expect mammals to be at least close in the evolutionary scale, great differences can be observed also in this (see [10]). I.e., some mammals appeared on the earth much time before others, and are much less developed.

Particularly interesting questions arise when trying to address these differences, and at this time biologists don‘t yet agree on the answers. When it comes to the differences in the evolutionary scale, many positions can be found. Some researchers agree that only few factors, like the ability to communicate, have brought certain mammals (and in particular, the Man) to a higher level of evolution; while other researchers claim that there are lots of factors affecting the evolution, some of which are constitutional.

Furthermore, the idea of “smartness” is a highly disputed concept, going back also to Darwin‘s studies and to the objections of a few of his fellow biologists. In this report, we won‘t get into that disputation. We will simply assume that a mammal is smarter if leads a life less exposed to dangers.

Almost all mammals experience two phases during the sleep, the paradoxical wave and the slow wave. There are great differences in the length of each phase and the frequency the animal switches from one to another. For instance, the Man tend to interval the slow wave with many paradoxical phases, each of longer length than the previous (usually till a longest of 40 minutes); while other mammals have only one uninterrupted paradoxical wave phase.

Empirical evidence shows that the dreaming phase is a phase of deep sleep, also associated with lethargy and maximum exposure to the environment. Many studies argue that dreaming makes a high quality sleep, in which the rest is maximal. On the opposite side, the non-dreaming phase is associated with absence of peripheral and secondary activities. There is experimental evidence that signals over a certain threshold are recognized by the animals when they are not dreaming. Hence we can consider this phase of sleep to be less deep and less relaxed. Intuition suggest that if an animal is exposed to dangers, will tend to sleep and dream less. We will try to prove this intuition, and also to reverse the idea.

In fact, some studies suggest[1] that the sleep behavior, and in particular the dream behavior, is linked to animals‘ wellness and their ability to protect themselves from the environment. In that way, dreams can be indirectly linked to the survival ability.

The present work is a statistical data analysis, and focuses on the relationships among time spent dreaming and sleeping in mammals, life duration, and smartness. We will also study the way life duration is affected by constitutional correlates. In the end, an attempt will be made to provide a discriminant of "smartness" differentiations between mammals.

The big issue behind this questions is the hypothesis that mammals living better (longer and deeper sleep, longer gestation time) tend also to live longer and to be smarter.

Section 2 presents the data set, Sections 3 and 4 perform simple analyses on the response variables; Section 5 will present multiple models and a cluster analysis. Finally, Section 6 performs a discussion on the analysis, and Section 7 summarizes some conclusions and gives some suggestions for further analysis.

[1] We will discuss this studies later, using the data to show the ideas.

## 2 Description of the Data

This data set is taken from [1], and consists of observations for 62 species of mammals. Ten variables were recorded, among which there are body and brain weight of the mammal, maximum life span and gestation time, total sleep time, dreaming and non dreaming time; and three indexes indicating the overall danger exposure, the danger exposure during the sleep, and the predation likelihood.

See technical appendix A, page 24, for the head of the data set.

Four variables are classified as constitutional: body weight (in Kg), brain weight (in gr), maximum life span (in years) and gestation period (in days).

Other three of them are classified as danger measures: the predation index ranks the species from the least likely to be prayed upon to the most likely (index with values from 1 to 5), the sleep exposure index ranks the den where the animals sleep from the most protected to the most exposed to external dangers (index with values from 1 to 5) and the overall danger exposure has got the same levels (values from 1 to 5, from the least to the most).

The sleep times are expressed in hours per day. Total sleep is a sum of dreaming and non dreaming time; so we don't need the three together [2]. For our purposes the dreaming time is more interesting than the slow wave, so we will try to use dream and total sleep times.

There were missing values for many observations. We completed some of them with the help of the Encyclopedia and some web sites. The information gathered refers mainly to the gestation times:

```
# Gestation times #
Desert Hedgehog       = 40 days
Genet                 = 66 days
Giant Armadillo       = 200 days
Star nosed mole       = 42 days

# Maximum life span #
Artic Ground Squirrel =  2 years
Mole Rat = 20 years
```

### 2.1 Constitutional variables

Please, see technical appendix C for more details on the transformations done on the constitutional variables.

Body weight has got a mean of 198.8 Kg and a great variability (900Kg as standard deviation). The are two big outliers, the African (6654Kg) and the Asian (2547Kg) elephant. A strong skewness was observed, and a log-transformation done to solve this problems.

The same problems occur with brain weight, (mean= $283g$, standard deviation= $930g$) with the same outliers. Again, a log transformation solves the problem.

Gestation time has a mean of 138 days and a standard deviation of 143 days. Again, the two elephants are outliers, with more than 600 days of gestation. This variable shows a long right tail, so, again, a log-transformation will be useful to work with a symmetric variable.

[2] We would get a singular $X$ matrix if we included all the three in a model!

| | Body | Brain | Tot. Sleep | Dreaming | Non-dreaming | Gestation | Life Span |
|---|---|---|---|---|---|---|---|
| Body | 1 | .935 | -.307 | -0.109 | -0.376 | .683 | .303 |
| Brain | | 1 | -.358 | -0.105 | -0.369 | .778 | .508 |
| Tot. Sleep | | | 1 | 0.727 | 0.963 | -.58 | -.415 |
| Dreaming | | | | 1 | 0.514 | -0.380 | -0.296 |
| Non-dreaming | | | | | 1 | -0.542 | -0.384 |
| Gestation | | | | | | 1 | .636 |

Table 1: Correlation structure

| | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Sleep Exposure** | 27 | 13 | 4 | 5 | 13 |
| **Predation** | 14 | 15 | 12 | 7 | 14 |
| **Overall Danger** | 19 | 14 | 10 | 10 | 9 |

Table 2: Summary for danger indexes

Life span has got a mean of 19.57 years, with a standard deviation of 18.2. The maximum is achieved by the man (100 years), which is an outlier. Again, we will need to work with a log-transformation, especially because this is one of the response variables.

Table 1 shows the correlation structure between the quantitative variables in the data set. It will be used as reference also in the further analysis. There is a gigantic correlation between body and brain weight (.935), high correlations between body weight and gestation time; and between brain weight and gestation time. It is reasonable that bigger mammals need more time before birth. The correlation between body or brain weight and total sleep is negative; so heavier mammals are likely to sleep less than other mammals. We expect to say more about this in the following analysis. There is high positive correlation also between gestation time and maximum life span; and negative correlation between life span and total sleep.

## 2.2 Danger Variables

Table 2 shows the distribution of the 62 mammals in the study among the 5 levels of the danger indexes. We can see that great part of the mammals are “safe” (index value less than or equal to 2).

## 2.3 Sleep Times

The total sleep time has a mean of 10.53 hours per day, with a standard deviation of 4.61. There are no particularly strong outliers, and Figure 1 shows very little evidence of non-normality. The observed minimum is 2.6 hours; and the maximum 19.9 (almost all of the day!). This is the Little brown bat, which as we could expect has got all 1s in the danger indexes. The big brown bat is similar, with same index values and 19.7 hours of sleep per day.

Dreaming time has a very low mean of 1.972 hours, with a standard deviation of 1.442. The maximum is achieved by the Water opossum (6.6), and the minimum is achieved by the Echidna,

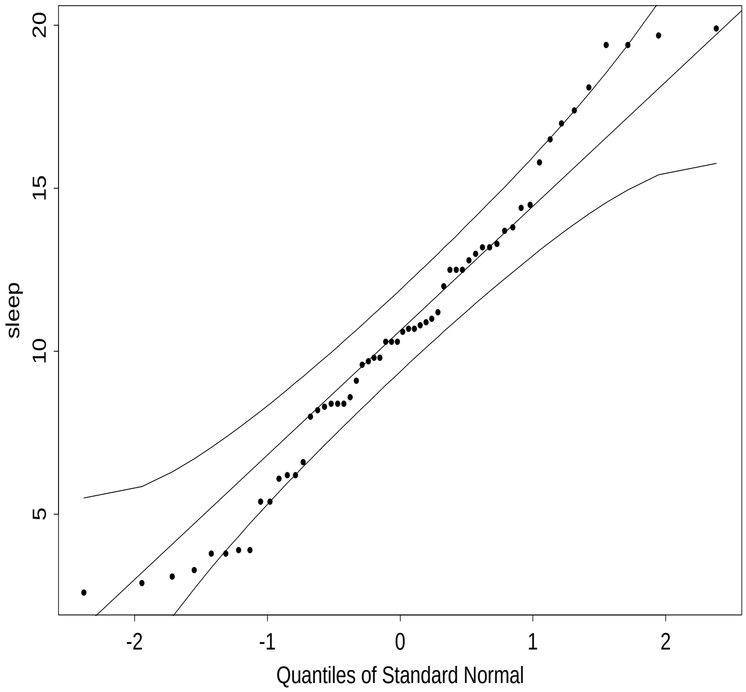


Figure 1: Normality of Total sleep time

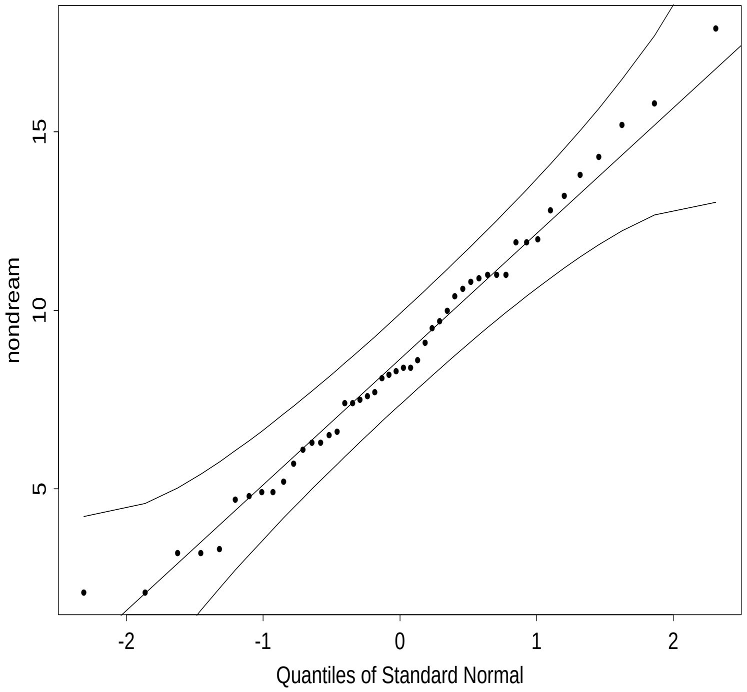


Figure 2: Normality of Non dreaming time

which has no paradoxical wave. The variable is slightly skewed, with a long right tail and at least three outliers (Water opossum with 6.6, Giant Armadillo with 6.1 and N.American Opossum with 5.6). Man dreams 1.9 hours per night, over a total sleep of 8 hours.

The slow wave is longer, with a mean of 8.67 hours per night and a standard deviation of 3.67. From Figure 2, we can see that there is no evidence of non normality for this variable.

A new variable was created, namely the dreaming ratio (hours dreaming divided by total sleep). On average, these mammals dream 18.6% of their sleeping time, with very low standard deviation (.09). The maximum dreaming ratio is achieved by the Asian elephant, with 46%. Man dreams 24% of the time. More details on this variable are given in Technical Appendix C. The low value of the standard deviation (together with other analysis performed) show that, though the times tend to be different, the dreaming ratio is not so different (almost all of the observations are between 10% and 35%)

# 3 Simple Analyses: Life Span

A scatterplot matrix is shown in technical appendix D, together with more details on each of the following analyses.

## 3.1 Body Weight on Span

A few issues arise in examining this relationship, which are described in detail in technical appendix D. The final R-squared is 62%, which is really high for a simple regression. The parameter estimate (on log(body)) is 0.2545 with a $p$ value almost 0. So heavier mammals tend to live longer.

## 3.2 Brain on Span

Again, a few issues arise in examining this relationship, which are described in detail in technical appendix D. This time the final R squared is .77, the parameter estimate (on log(brain)) is .36 with a $p$ value almost 0. For this and other reasons, we can conclude that brain weight is more important than body weight on life duration; which suggests the need for further analysis on this relationship.

## 3.3 Sleep times on span

There is not a strong relationship between total sleep and life span, though animals that sleep more tend to live less than the others. The R squared is only 20% and the sleep parameter is around -.1, though being strongly significant. Even lighter relationship with dream and non-dream sleep was observed, with R squares around 15%.

## 3.4 Gestation on Span

There is an amazing strong linear relationship between gestation time and life span, with a parameter estimate of .57 (on the log transformations) and an R squared of more than 73%.

## 3.5 Span and Danger Indexes

Technical Appendix D.6 shows that the danger indexes are not intuitively related to life span, and so we will use them only to explain the variability in sleep times.

# 4 Simple Analyses: Sleep Times

## 4.1 Body on Sleep times

Though the Multiple R-squared is low (28%), there is a decreasing relationship between total sleep and body weight. So, even if the body weight is not a good predictor for the sleeping time, we expect bigger animals to sleep less than the others. This may be due to the bigger proportion of smaller mammals going to lethargy.

More details, together with influential and residual analysis, are given in technical appendix D.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 0.211 | 0.161 | 0.131 | 0.125 | 0.211 |

Table 3: Dreaming ratio and sleep exposure: means

### 4.2 Brain on sleep times

A faster decreasing and stronger relationship is observed between brain weight and sleep times. Definitely brain weight is more helpful than body weight in explaining our response variables. The parameter estimate on total sleep is -1, -0.87 on non dreaming time and -0.19 on dreaming time. This is good: one of our hypotheses is that dreaming states somewhat auto regulated (we will show this later with a more refined model); so it shouldn't be much affected by any other variable. As usual, more details are given in the technical appendix.

### 4.3 Life span on Sleep Times

Again, we observe a strong decreasing relationship between life span and sleep times, and again, dreaming time has got the lower negative parameter estimate than sleep.

### 4.4 ANOVAs between Sleep and Danger Indexes

Figure 3 shows the boxplot of sleeping times for the danger indexes. As we can see the decreasing relationship between sleep times and danger indexes is significant (as shown in technical appendix D), consequently we can conclude that animals more in danger tend to sleep less.

In addition, it is interesting to notice how the dreaming time gets lower more smoothly than the total sleep time. In fact, sleep is almost unaffected by predation index, while strong differences are seen only between the first four levels and the fifth of sleep exposure and danger index. The same is for the non dreaming time, while the dreaming time declines smoothly as the dangers arise. This particular behavior is explained by the same boxplots, drawn on the dreaming ratio; which are given in Figures 4, 5, 6.

The dreaming ratio, I remind, is the number of hours dreamed divided by the total sleep time. It is almost unaffected by predation and danger indexes. At a contrary, it presents a particularly interesting relationship with sleep exposure. In fact, the dreaming ratio gets lower as the sleep exposure gets higher, but it doesn't go below a threshold, so that, when the total sleep gets really low (index=5), the ratio is higher.

This is a particularly interesting phenomenon, which opens the path to further research: it may suggest that the body "needs" to dream. When the hours slept are very low, great part is spent dreaming. I think that simply focusing such a research on the man would guarantee interesting results. Table 3 shows something confirming this hypothesis: the average dreaming ratio when the sleep exposure is maximal is the same as the dreaming ratio when the sleep exposure is minimal.

In conclusion, mammals tend to have a less deep sleep when in danger. When the danger becomes too strong, they tend to sleep less; while a certain amount of dreams comes up anyway.

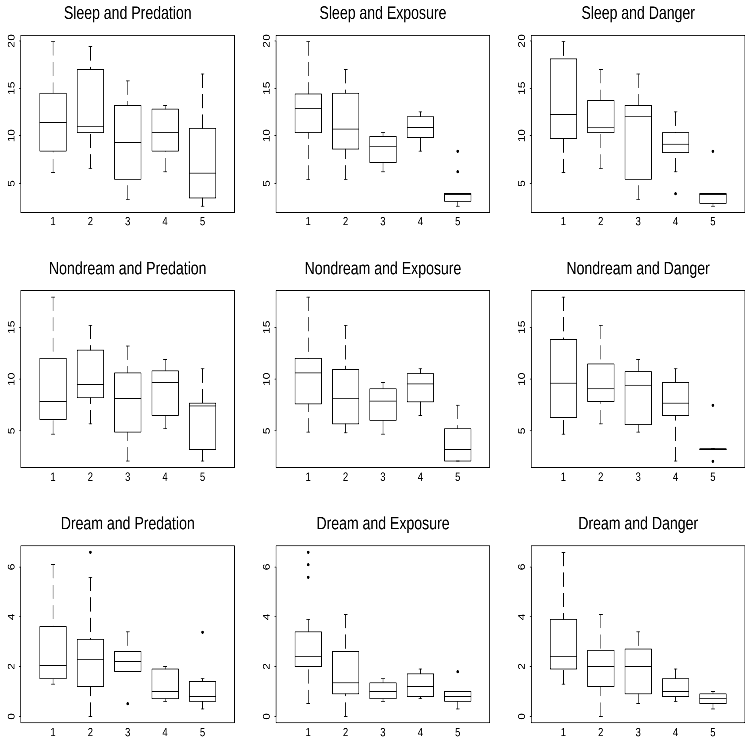


Figure 3: Sleep and dangers

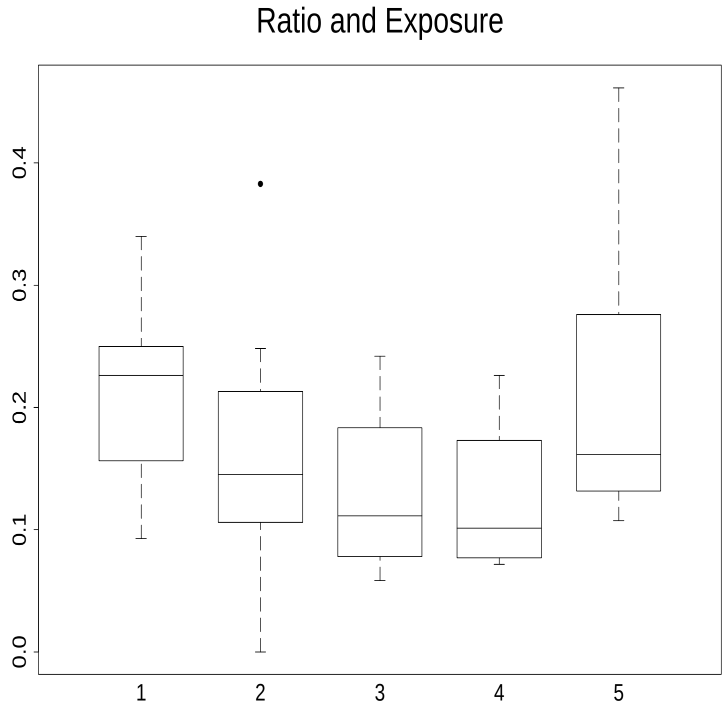


Figure 4: Ratio and Exposure

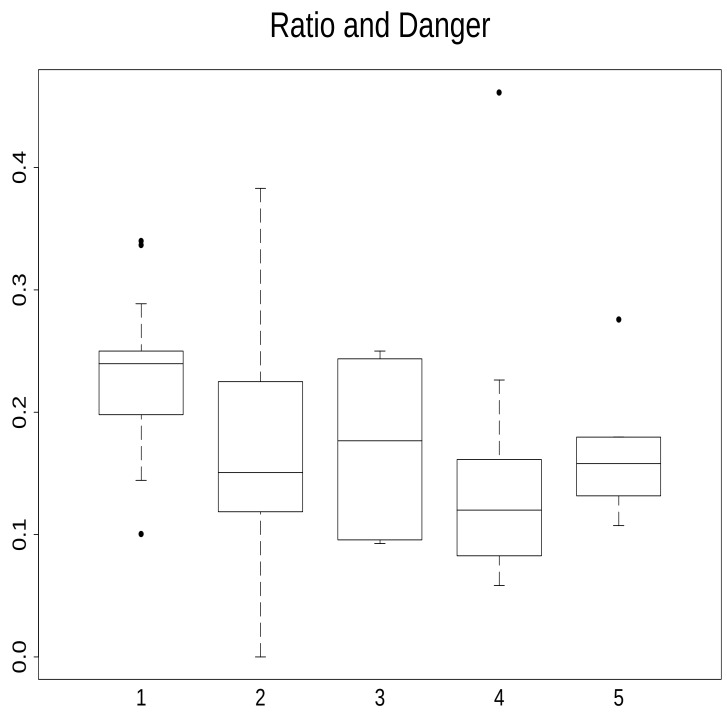


Figure 5: Ratio and Danger

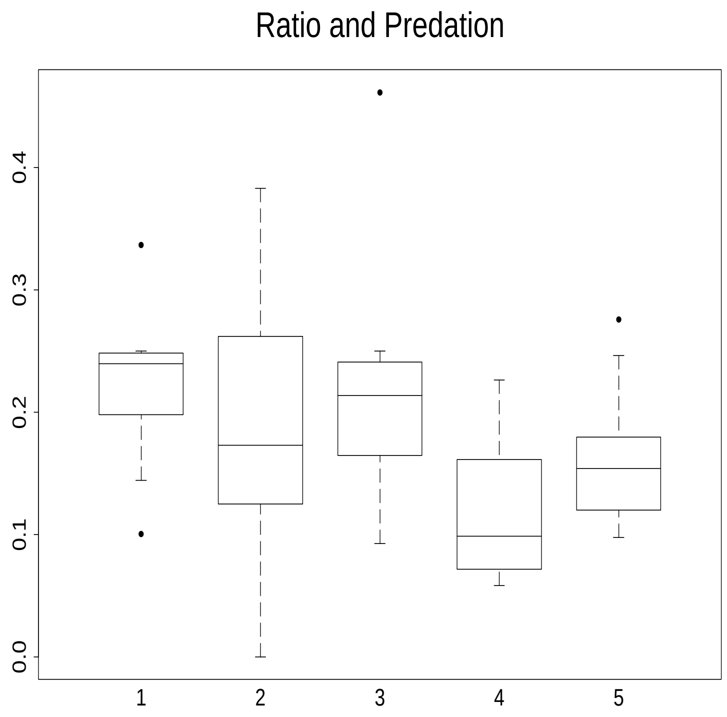


Figure 6: Ratio and Predation

# 5 Multiple Models

Both the multiple regression models turn out to be of little importance. More accurate models are developed in the sections 5.3 and 5.4.

## 5.1 Sleep Model

A few issues about variable selection for this model are discussed in technical appendix E. The final model, determined using a stepwise selection method, is:

```
sleep = 4.14 + 2.30 dream + 1.04 log(span) -0.84 log(body),
```

in which the dreaming time hides the effects of the other variables. The R squared is 0.72. A less good model, but with more interesting consequences, is the model in which `dream` is dropped from the set of possible variables. In this case, the R squared goes down to 56%; and body is not any more in the final model. There are only the danger index and brain, which confirms what we had noticed (brain is more important than body) and is a validation of the overall danger index (and so, of the three indexes).

## 5.2 Life Span Model

The final model[3], which has got an R squared of about .84%; has got dreaming time, predation index and brain weight as explanatory variables. See technical appendix E for more details.

## 5.3 Smart Mammals model: a new approach to the data

Three polytomic logistic regression models were fitted[4], one for each index. The explanatory variables were sleep time, brain weight, gestation time and body weight. Though it is not possible to use $p$-values (the sample size is not enough to safely assume asymptotic distribution properties on the $t$ statistic); the difference between the variables is evident.

For instance, in the model with overall danger as response, sleep ($t$ value equal to -4.2) is important; while the other variables (especially gestation and body) are not; with $t$ values between -1 and 1. The model is summarized by:

```
F_danger_i=\beta_0i - .365 sleep -.474 log(brain)
-.03 log(gestation) +.24 log(body), i=1,...,4;
```

where $F_{danger_i}$ stands for the $i$-th cumulative logit, and $\beta_{0i}$ is the $i$-th intercept[5].

The same happens using the other two indexes as response variables. In all cases, sleep has got a negative parameter. So mammals who sleep more are less exposed to dangers. In this case we can give an opposite conclusion to what we did earlier: instead of saying that mammals less exposed to dangers can sleep more, we can claim that mammals who sleep more tend to be less exposed to dangers (so, they‘re smarter). More details and numerical summaries are given in technical appendix E.3. A brief discussion on causal inference and the reasons and limitations of this conclusion will be given in Section 6.

[3] Again, determined using a stepwise.

[4] See technical appendix E.3 for a brief summary on the theory of the polytomic logistic regression models used in this analysis.

[5] As I said, see the technical appendix for a definition of this quantities

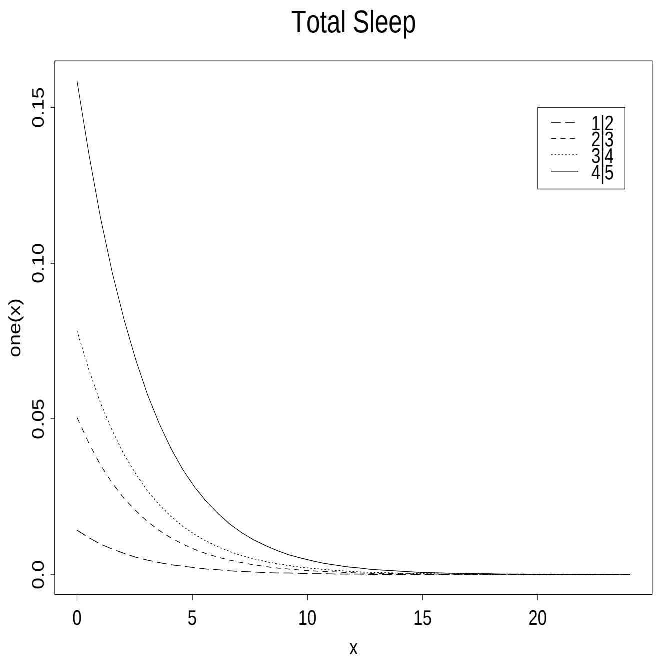


Figure 7: Cumulative logits for Sleep on Sleep Exposure

Including `dream` in the model we see that, because of collinearity, sleep's $t$ value goes to 0 and dream takes an high $|t|$ value. So we may want to conclude that dream is more important than sleep in the smartness of an animal. This is confirmed by the simple polytomic logistic regressions. The model with sleep exposure as response variable was the most interesting and the most representative, so it was taken as a representative example. In the simple models, total sleep has got a parameter estimate of -.37 (standard deviation=.08). Dream, a parameter estimate of -1.35 (standard deviation=.37).

Figure 7 shows the four fitted models on the cumulative logits for sleep exposure in dependence of total sleep. We see that the probability of any cumulative logit goes to 0 as sleep hours raise, but the curves are higher in correspondence of low total sleep. That is, a low total sleep brings about an high probability of sleep exposure; and the higher the level considered, the higher the probability.

Figure 8 shows the four fitted cumulative logits for sleep exposure in dependence of dreaming time. We can see the same features of the previous graph. Moreover, as we said, the probability of high exposure is higher than the probability determined by total sleep (which is the important conclusion).

In other words, using the interpretation scheme given in [2], there are $e^{-0.366} = .7$ odds that a high sleeping mammal would be at or above a level of sleep exposure (say $\geq 3$), that is there are 1.43 odds that a mammal who sleeps much is in low sleep exposure.

And there are $e^{-1.358} = .26$ odds that a mammal who dreams much is at or above a level of sleep exposure; that is there are 3.84 odds that, say, a hard dreamer mammal is in category 1 or 2 more than in the others. Once more, we can conclude that dreaming is more discriminating than sleeping in the smartness classification.

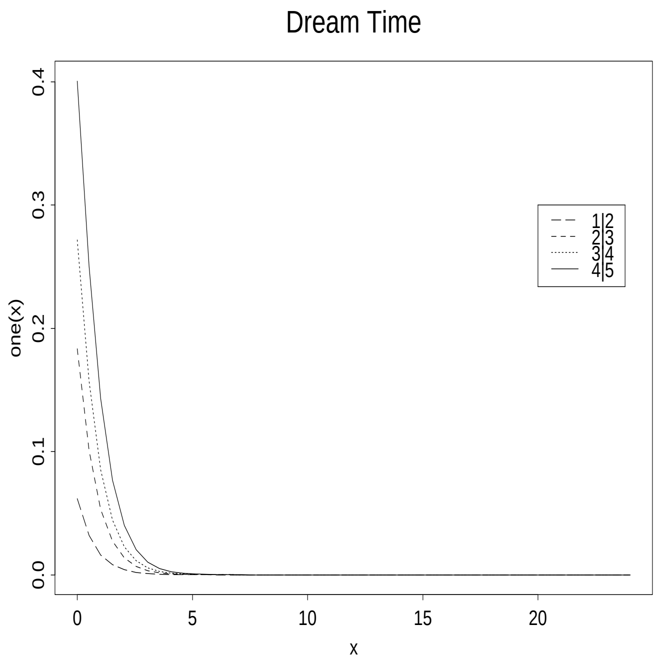


Figure 8: Cumulative logits for Dream on Sleep Exposure

More details and a theory description of the model are given in technical appendix E.3.

### 5.4 Similarities between the mammals

A first hierarchical cluster analysis, with distance defined by the three indexes gave the clustering in Figure 9. We note that the elephants are in the same cluster, but the Asian Elephant is more similar to the Kangaroo than to the African Elephant (which is exposed to more dangers). Man belongs to a populated group, which can be thought of as a "control", less exposed, group[6]. The Gorilla stands alone, because of its particular situation (predation and danger are 1, while sleep exposure is 4). Figure 10 shows the same cluster analysis with values for the sleep exposure instead of the names of the species. This is a sort of validation of the clustering: there are very similar values within each group and different values between. This kind of labeling provides also a validation of our interpretation: we can see that more or less all of the mammals in the left groups are highly exposed during the sleep, and the mammals in the right groups are lowly exposed. We tried many other "labeling" for this clustering, trying for instance to see if this grouping went together with body weight or other features of the data not directly included in the distance matrix. Interestingly enough, total sleep time goes together with the groupings; but not as well as dreaming times, which is shown in Figure 11. No mammal in the left groups dreams more than 1.8 hours; the group with man inside shows also heavy dreamers (i.e., the bats and the opossum, with more than 6 hours).

Another idea is to include sleep times in the distance matrix, thus generating the hierarchical clustering in Figure 12. This results in finer but less distinguishable grouping (too many groups with too few mammals in each), and is more reasonable: the opossums, bats, armadillos, elephants

[6]As we noted before, there are more "safe" mammals than there are exposed ones.

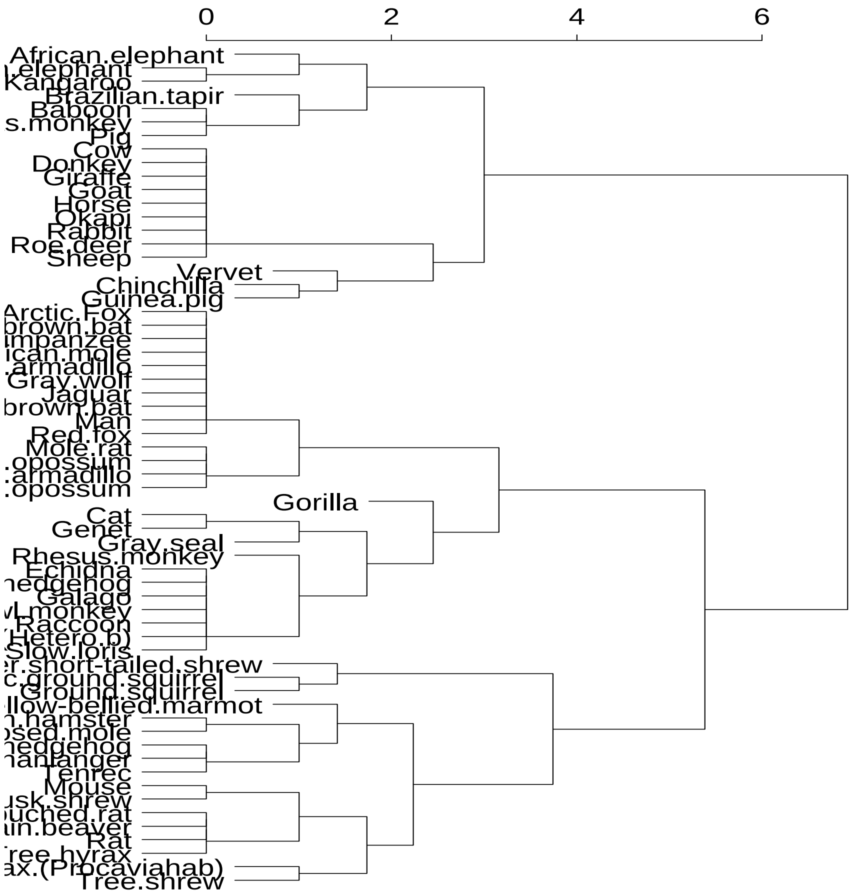


Figure 9: Clustering. Distance=danger indexes

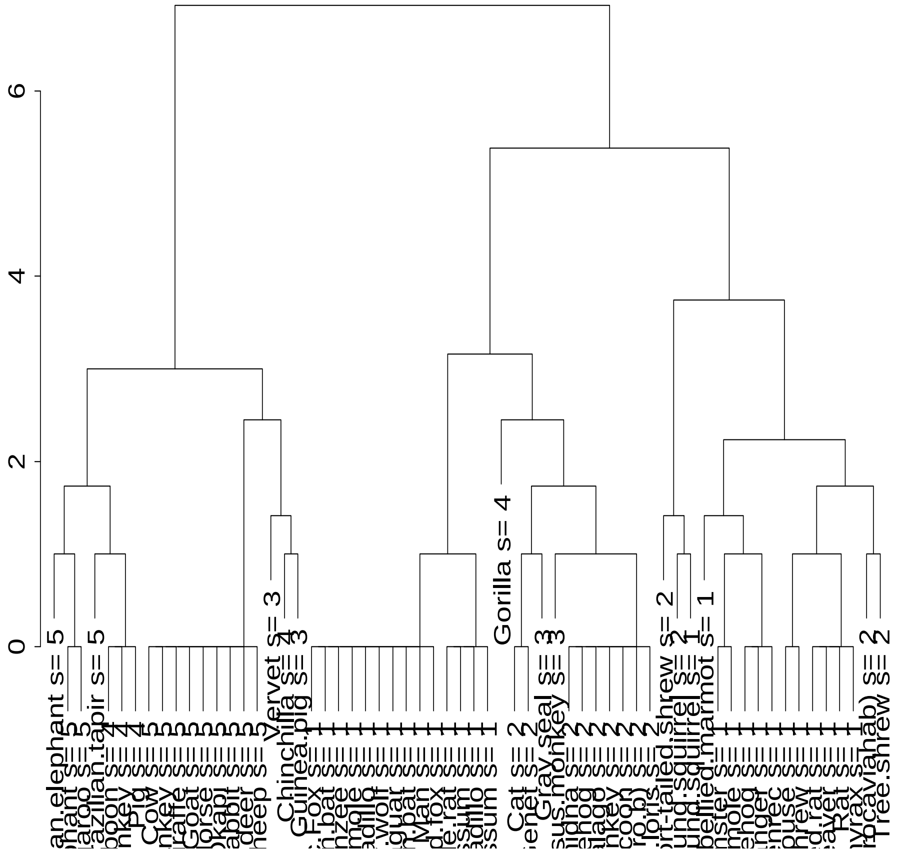


Figure 10: Clustering. Distance=danger indexes. Label=sleep exposure

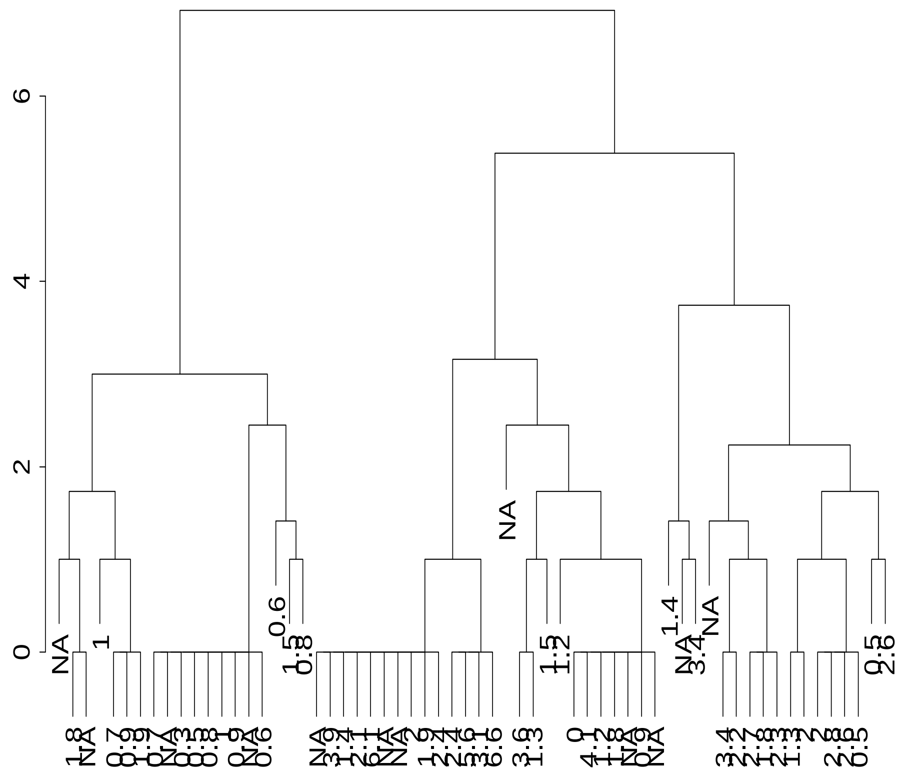

Figure 11: Clustering. Distance=danger indexes. Label=dream

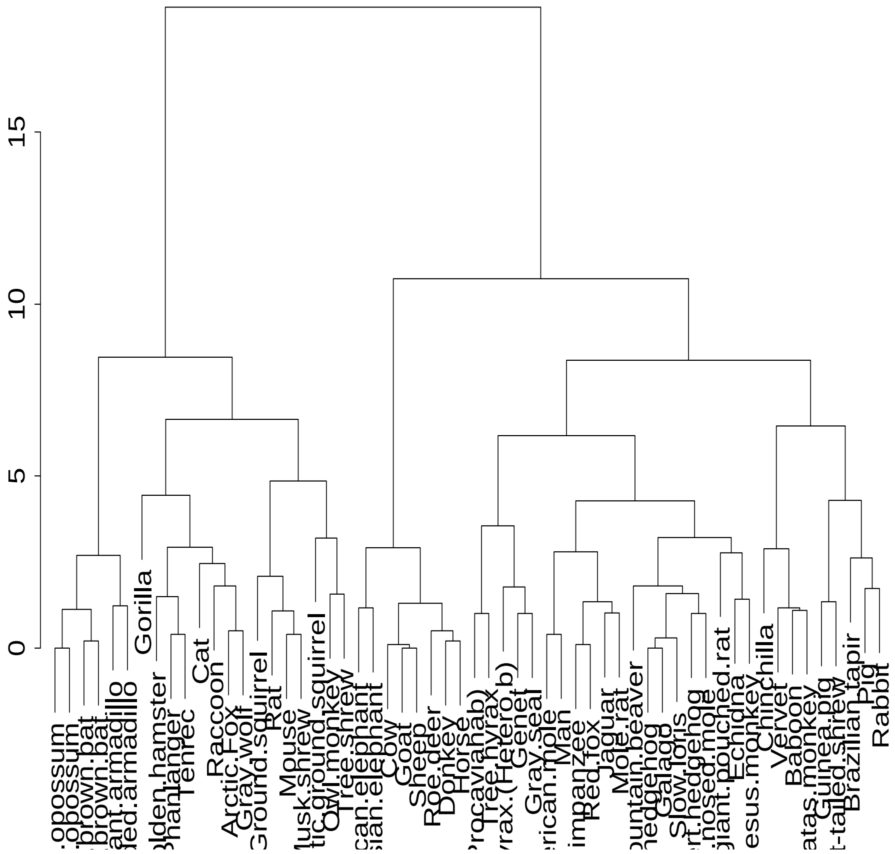


Figure 12: Clustering. Distance=danger indexes and sleep time

are set together. Bats and opossums are similar, as we had noticed (they spend more than 19 hours sleeping).The Gorilla stays, again, alone. The Squirrel, Rat, Mouse and Musk Shrew are put together on two levels, maybe because they have high predation and danger indexes (all between 3 and 5), low sleep exposure (all 1s) and a common consistent amount of hours slept per day (about 13). On the second level, the Man is close to the Chimpanzee and to predators like the Jaguar and the Red Fox (maybe because of their low danger indexes). On the other hand, the Chimpanzee is surprisingly distant from the monkeys, which again suggests the need of biological reasoning and further analysis. On the extreme right side, we can see Pig and Rabbit together: they sleep on average 8 hours, and are strongly exposed to dangers. On a high level, two groups are observed. We can claim that the group on the right is the group of high sleep animals (all more than 12 hours per night), while the left one is the group who sleep less than 12 hours per night. No other particularly interesting labeling was observed.

## 6 Discussion

There are several things that are worth to be pointed out.

First of all, there are some limits to the possibilities of generalizing our results. The data come from an article dating back to 1976, and refers only to informations available to the authors. This is not a random sample of size 62 among the 5000 species of mammals; and even if it were, we would have been very careful generalizing the results to the entire mammals population. As we pointed out in the introduction, the strong differences between the species warn us against general conclusions. Moreover, in almost no case we have been able to include in the model all of the 62; thus using even a smaller sample size. Anyway, what we are going to say holds for the analyzed species, which should be interesting enough.

I would also like, before going ahead, to discuss some issues of causal inference, which turn out to be really important in this analysis. In fact, during the analysis, we switched response and explanatory variables (namely, danger indexes and sleep times). This is a very powerful tool, but we should be very careful in the interpretation of the conclusions. As [11] addresses, the language of statistics and probability doesn't explicitly include the possibility of giving causal conclusions.

> "Correlation does not imply causation"[7]

that is, the presence of two correlated phenomena doesn't tell us which is the cause and which is the effect. As [14] states, it is a duty of the researcher to suggest the causal implications of a relationship between variables, and prove it with more than one scientific tool. For this reason, we tried to validate our hypotheses using different methods (logistic regression, cluster analysis, simple EDA by means of boxplots), achieving and repeating the same conclusions. We can thus consider our hypotheses valid, though

> "it is rare that firm conclusions about causality can be drawn from one study"[8].

Consequently, other studies are needed to confirm what we suggested. On the other hand, some researchers agree that, if the model is identifiable (roughly speaking, if the variables in the model can be observed[9]), causal interpretations are possible:

> "Under what conditions can we give causal interpretation to identified structural coefficients? [...] Always!"[10]

The causal conclusions of this analysis should be perfectly reasonable, but definitely need confirmation by means of further research.

The delicate part is that both danger exposure brings about shorter sleep and shorter sleep brings about danger exposure. Anyway, this is not so strange. The underlying intuition is that, if a mammal lives a scaring experience and it doesn't feel safe, then it will sleep less for a relatively short period of time. All of us experienced this kind of problems once in life. In everyday life, danger exposure brings about shorter sleeping.

---

[7] [15]

[8] [16]

[9] A formal explanation would lead us to discussion of the concept of "counterfactuals", which for length reasons we skip.

[10] [11]

On the other hand, some psychological studies suggest that dreams in animals make them learn ways to protect from the environment and live better. That is, dreams are like a "gym" of life. The polytomic regression models could be used to indirectly assess this idea, proving that dreaming causes a lower exposure to dangers. So, in the long run, a mammal sleeping and dreaming longer would be able to guarantee itself a lower exposure to dangers. Note that, since the variables recorded in the data set deal with general behaviors of each species, it sounds more natural to prove and support this second conclusion, instead of the first.

## 7 Conclusions

A central position in this study is occupied by the dreaming time[11], which turns out to be the most stimulating in this analysis. We showed that it is more important than the total sleep time in the determination of a mammal‘s smartness, and also in the prediction of life duration. Moreover, we verified the hypothesis that dreams are needed by the body, and preferred to a longer sleep. A bigger exposure to dangers brings about a drastic cut of sleep time, while dreaming time declines more smoothly. This behavior may also be due to the high restful power of dreaming.

We saw that dreaming influences the other variables, while it is not much affected by them (low variability in the dreaming ratio).

In my opinion, future research should start by better defining the sleep variables. Taking into account how many times a mammal switches from one phase of sleep to another may yield interesting results.

In addition, a better tuning of the danger indexes is recommended. Though a general reliability of this indexes was assumed in this study, we have had occasions to suspect about their validity; for instance when studying their relationship with life duration[12].

Another issue that arise in this study regards the effects of brain weight. Few studies take into account the brain weight of the animals, and we showed that it is more important than the body weight. Maybe further analyses will need to take more into account this constitutional characteristic.

As a last recommendation, we would suggest including a measure of evolution in the data set (i.e., order position in a reference evolutionary scale, approximate hundreds of years since first appearance on the earth, etc.). This should help confirm the hypothesis that smartness is related to the evolutionary position.

In conclusion, in this study we argued that:

- Heavier mammals, and especially mammals with heavier brains, live more than the others.
- Animals with longer gestation times live more than the others, though gestation time is not important in the determination of the smartness.
- More generally, we came to the counterintuitive conclusion that smart mammals don‘t tend to live longer (i.e., there is no evident relationship between life span and the danger indexes footnoteAs discussed in technical appendix D.6; and life span is not “significant” in the logistic models).
- Dreams are “dull”, that is, not much affected by constitutional correlates.
- Mammals more exposed to dangers tend to sleep less and dream less, though a certain amount of dreams comes anyway.
- Mammals who sleep and dream more are smarter: they tend to lead safer lives.

[11] And its relationship with sleep exposure

[12] And when observing strange indexes for certain mammals, i.e., high danger exposure for the Pig.

## References


[1] T. Allison and D. Cicchetti, *Sleep in mammals: ecological and constitutional correlates*, Science, vol. 194 (Nov., 1976), pp. 732-734

[2] Chap. T. Le, *Applied Categorical Data Analysis*, Wiley Series in Prob. and Stat., 1998

[3] J.F. Hair Jr., R.E. Anderson, R.L. Tatham, W.C. Black *Multivariate Data Analysis, third edition*, MacMillan, 1992

[4] R. Coppi, *Lezioni di Statistica Multivariata*, Dipartimento di Statistica, Probabilita' e Statistiche Applicate, Serie C - Didattica, 1998

[5] P. McGullag, J.A. Nelder, *Generalized Linear Models, 2nd Edition*, London, Chapman&Hall, 1989

[6] R.A. Johnson, D.W. Wichern, *Applied multivariate statistical analysis*, Prentice-Hall, 1982

[7] C. Helberg, *Pitfalls of Data Analysis (or How to Avoid Lies and Damned Lies*, 1995: `http://www.execpc.com/~helberg/pitfalls`

[8] Online Zoologists: `http://www.rtis.com/nat/user/elsberry/`

[9] Animal Diversity Web: `http://animaldiversity.ummz.umich.edu/`

[10] The evolutionary timeline: `http://www.talkorigins.org/origins/geo_timeline.html`

[11] J. Pearl, *The New Challenge: From a Century of Statistics to an Age of Causation* `ftp://ftp.cs.ucla.edu/pub/stat_ser/R249.ps`

[12] D.R. Cox, *Causality: Some Statistical Aspects*, Journal of the Royal Statistical Society, Series A (Statistics in Society), Volume 155, Issue 2 (1992), pp. 291-301

[13] R. Stone, *The Assumptions on Which Causal Inference Rest*, Journal of the Royal Statistical Society, Series B (Methodological), Volume 55, Issue 2 (1993), pp. 455-466

[14] P.W. Holland, *Statistics and Causal Inference*, J.A.S.A., Volume 81, Issue 396 (Dec., 1986), pp. 945-960

[15] S. Kotz, N.L. Johnson, editors. *Encyclopedia of Statistical Sciences* New York, 1982. Wiley and Sons, Inc.

[16] D.R. Cox, N. Wermuth. *Multivariate Dependencies - Models, Analysis and Interpretation* Chapman & Hall, London, 1996

# A Head of the dataset

```
Data from which conclusions were  drawn  in  the  article  "Sleep  in
Mammals: Ecological and Constitutional Correlates" by Allison, T.  and
Cicchetti, D. (1976), _Science_, November 12, vol. 194,  pp.  732-734.
Includes brain and body  weight,  life  span,  gestation  time,  time
sleeping, and predation and danger indices for 62 mammals.

Variables below (from left to right) for Mammals Data Set:

species of animal

body weight in kg

brain weight in g

slow wave ("nondreaming") sleep (hrs/day)

paradoxical ("dreaming") sleep (hrs/day)

total sleep (hrs/day)  (sum of slow wave and paradoxical sleep)

maximum life span (years)


gestation time (days)

predation index (1-5)
                1 = minimum (least likely to be preyed upon)
                5 = maximum (most likely to be preyed upon)

sleep exposure index (1-5)
                1 = least exposed (e.g. animal sleeps in a
                    well-protected den)
                5 = most exposed

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)

Note: Missing values denoted by -999.0



African elephant         6654.000 5712.000  -999.0  -999.0
3.3    38.6   645.0       3       5       3
```

```
African giant pouched rat   1.000    6.600     6.3     2.0
8.3     4.5    42.0       3       1       3
Arctic Fox                  3.385   44.500  -999.0  -999.0
12.5    14.0    60.0       1       1       1
Arctic ground squirrel       .920    5.700  -999.0  -999.0
16.5  -999.0    25.0       5       2       3
Asian elephant           2547.000 4603.000     2.1     1.8
3.9    69.0   624.0       3       5       4
Baboon                     10.550  179.500     9.1      .7
9.8    27.0   180.0       4       4       4
Big brown bat                .023     .300    15.8     3.9
19.7    19.0    35.0       1       1       1
Brazilian tapir           160.000  169.000     5.2     1.0
6.2    30.4   392.0       4       5       4
Cat                         3.300   25.600    10.9     3.6
14.5    28.0    63.0       1       2       1
Chimpanzee                 52.160  440.000     8.3     1.4
9.7    50.0   230.0       1       1       1
Chinchilla                   .425    6.400    11.0     1.5
12.5     7.0   112.0       5       4       4
```

## B NA Fill in

Since there were so many missing values in such a small dataset, we tried in many ways to fill in what possible. Using the Encyclopedia and Internet sites we found some values, and we wanted to determine others regressing the variable on the other ones, and predicting the missing values.

> "This has some drawbacks, but it is a reasonably sensible thing to do as long as the missing values are not in the response variable in a regression you want to perform"[13].

But, unfortunately, the one variables with NAs once that the gestation time for the Artic Ground Squirrel has been found are life span and sleep times, which are response variables.

As a validation, we fitted two multiple models: one using the original data set, and the other using the new data set. The idea is that, if the model is not too different (especially in the Multiple R squared), then we found reasonable values.

This is the model fitted on the old data set:

```
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.0389  0.6116     3.3340  0.0028
       body -0.0018  0.0011    -1.6610  0.1097
      brain  0.0012  0.0006     1.9751  0.0599
      dream -0.1216  0.1487    -0.8177  0.4216
      sleep  0.0571  0.0441     1.2932  0.2082
```

[13] Brian Junker, email sent to me on 11/10/2001

```
  gestation  0.0027  0.0020     1.3366  0.1939
      pred1 -0.5415  0.2926    -1.8507  0.0766
      pred2 -0.5552  0.2244    -2.4741  0.0208
      pred3 -0.2930  0.1857    -1.5779  0.1277
      pred4 -0.2244  0.1359    -1.6519  0.1116
    slexpo1  0.0611  0.1796     0.3401  0.7367
    slexpo2  0.2341  0.1591     1.4714  0.1542
    slexpo3  0.2041  0.1471     1.3868  0.1783
    slexpo4  0.1118  0.1843     0.6065  0.5499
    danger1  0.3166  0.3482     0.9094  0.3722
    danger2  0.2901  0.2610     1.1115  0.2774
    danger3  0.1505  0.2322     0.6484  0.5229
    danger4  0.2565  0.2574     0.9965  0.3289

Residual standard error: 0.6565 on 24 degrees of freedom
Multiple R-Squared: 0.7559
F-statistic: 4.372 on 17 and 24 degrees of freedom, the p-value is 0.0005368
```

And this is the model fitted on the new one:

```
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.7971  0.4537     6.1656  0.0000
       body -0.0014  0.0011    -1.3575  0.1855
      brain  0.0011  0.0006     1.9345  0.0632
      dream -0.2482  0.1207    -2.0557  0.0492
      sleep  0.0222  0.0400     0.5545  0.5836
  gestation  0.0009  0.0016     0.5251  0.6036
      pred1 -0.2118  0.2185    -0.9692  0.3407
      pred2 -0.3612  0.1810    -1.9958  0.0558
      pred3 -0.1594  0.1677    -0.9507  0.3499
      pred4 -0.1421  0.1260    -1.1276  0.2691
    slexpo1  0.1050  0.1630     0.6439  0.5249
    slexpo2  0.2429  0.1573     1.5446  0.1337
    slexpo3  0.2481  0.1445     1.7164  0.0971
    slexpo4  0.1800  0.1772     1.0163  0.3182
    danger1 -0.1411  0.2560    -0.5512  0.5859
    danger2  0.0302  0.2045     0.1475  0.8838
    danger3 -0.0446  0.2047    -0.2181  0.8289
    danger4  0.0474  0.2291     0.2069  0.8376

Residual standard error: 0.6587 on 28 degrees of freedom
Multiple R-Squared: 0.7332
F-statistic: 4.526 on 17 and 28 degrees of freedom, the p-value is 0.0002133
```

We can see that the R squared, the residuals standard error and the coefficients are almost unchanged.

# C EDA

## C.1 Constitutional Variables

We can see from the stem and leaf that body weight is strongly skewed, with 27 species weighting less than 2Kgs and 10 high outliers, the highest two being really far from the others. This two mammals are the African (6654Kg) and the Asian (2547Kg) elephant.

```
Decimal point is 1 place to the right of the colon

    0 : 00000000000000011111111112222333333444447
    1 : 015
    2 : 8
    3 : 56
    4 :
    5 : 25
    6 : 02
    7 :
    8 : 5
    9 :
   10 : 0

High:  160.0  187.1  192.0  207.0  250.0  465.0  521.0  529.0 2547.0 6654.0
```

Figure 13 shows that a log-transformation is really useful to eliminate this skewness problem.
Figure 14, Figure 15, Figure 16 show the same transformation for the other constitutional variables.

## C.2 Dreaming Ratio

This is the stem and leaf for the dreaming ratio:

```
N = 48   Median = 0.1755466
Quartiles = 0.1160748, 0.2434167

Decimal point is 1 place to the left of the colon

   0 : z
   0 : 67899
   1 : 000111223344
   1 : 56666788
   2 : 00113344444
   2 : 5555689
   3 : 44
   3 : 8
   4 :
   4 : 6
```

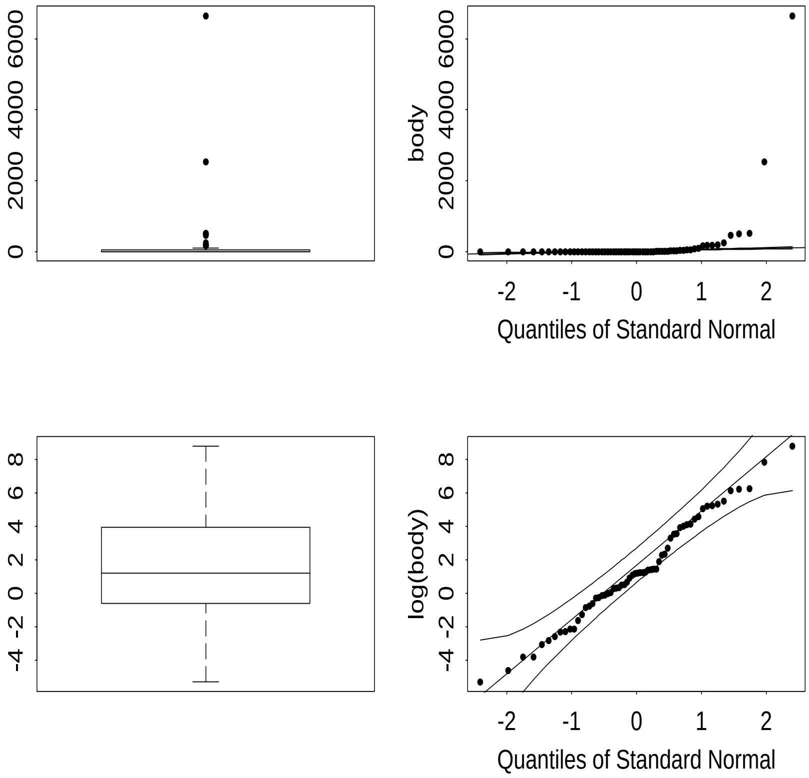


Figure 13: Body Transformation. Up: untransformed. Down: log-transformation

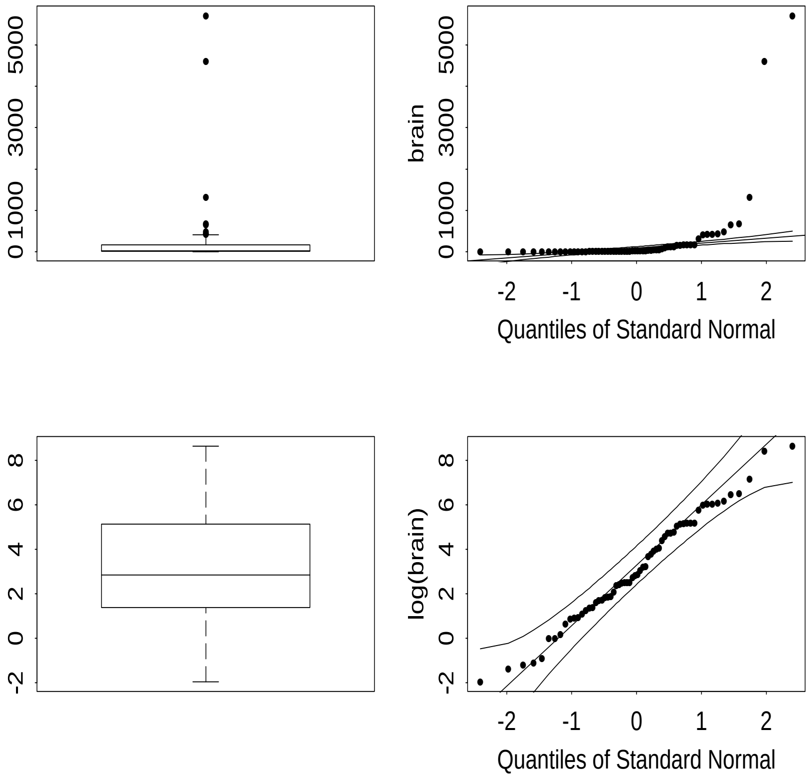


Figure 14: Brain Transformation. Up: untransformed. Down: log-transformation

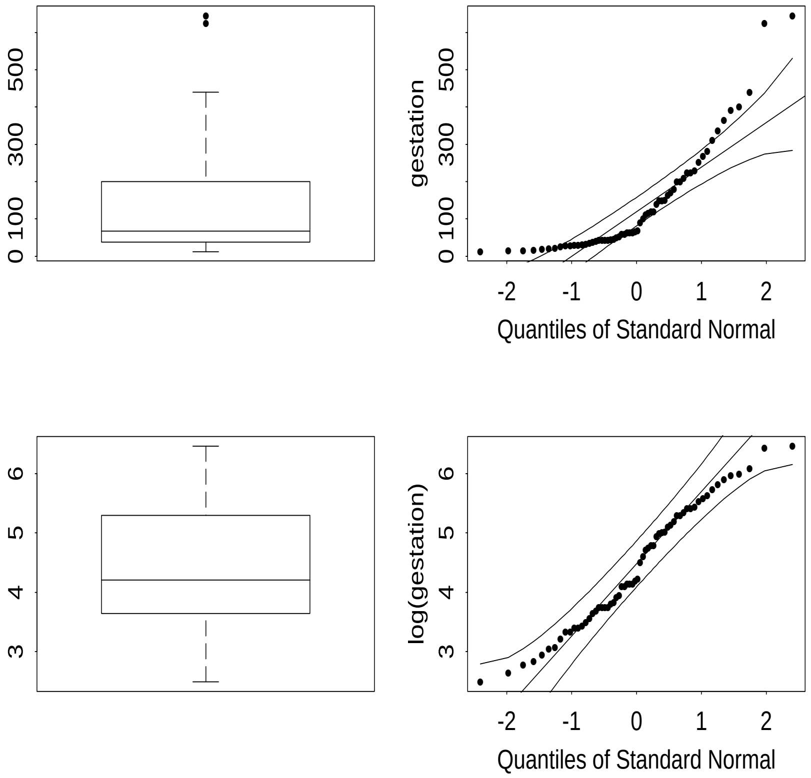


Figure 15: Gestation Transformation. Up: untransformed. Down: log-transformation

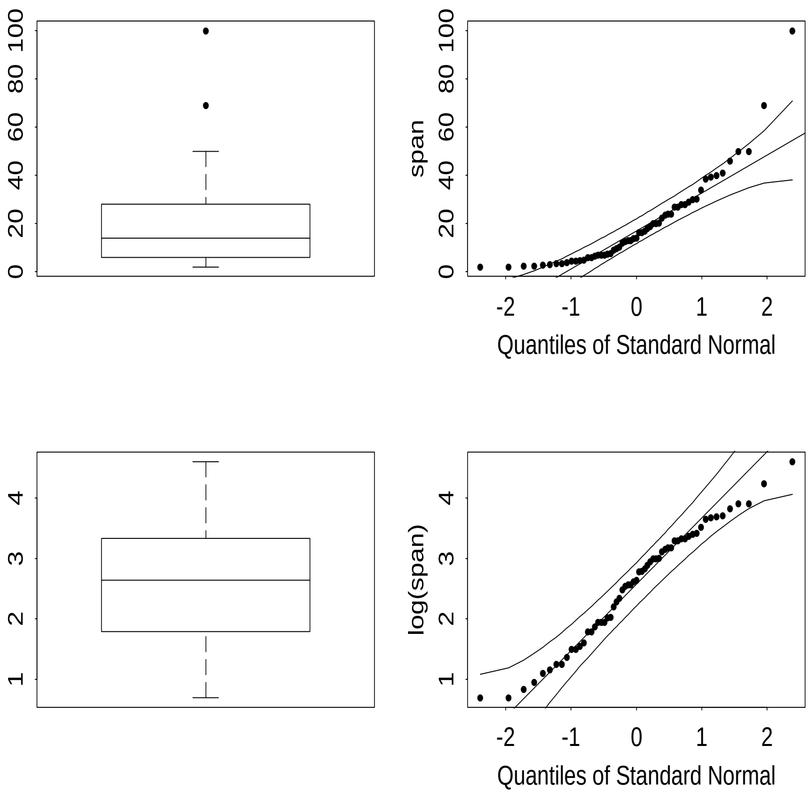


Figure 16: Life Span Transformation. Up: untransformed. Down: log-transformation

## D Simple analyses

Figure 17 shows a scatterplot matrix between the quantitative variables in this data set. Strong relationships are observed, though very few are linear; and some outliers (which are the elephants, the Echidna, and the bats, as usual) are observed. The transformed variables come up with good linear relationships.

### D.1 Body weight on life span

There is a strongly significant relationship between body weight and life span, though the multiple R-squared is only 50%.

```
Call: lm(formula = log(span[span != "NA"]) ~ log(body[span != "NA"]))
Residuals:
    Min      1Q Median     3Q   Max
 -1.539 -0.3989 -0.118 0.3534 1.919

Coefficients:
                         Value Std. Error t value Pr(>|t|)
            (Intercept)  2.2498  0.0997    22.5568  0.0000
log(body[span != "NA"])  0.2151  0.0289     7.4341  0.0000

Residual standard error: 0.6954 on 57 degrees of freedom
Multiple R-Squared: 0.4923
F-statistic: 55.27 on 1 and 57 degrees of freedom, the p-value is 6.012e-10

Correlation of Coefficients:
                        (Intercept)
log(body[span != "NA"]) -0.4196
```

But, from Figure 18 we can see that observation 32 (Little brown bat), maybe because of its light body (.01 Kg) and high span (24) is so influential to hide the real features of the relationship. When the Little brown bat is dropped from the dataset, the big brown bat becomes heavily influential (which is reasonable). Without bats in the data set, no observation is particularly influential; the parameter estimates are almost identical but the Multiple R squared raises to 62%. So there are other features of this particular animals that makes them different from the others. We can expect the bats to be influential also in other analyses.

```
Call: lm(formula = log(span[ - c(7, 32)]) ~ log(body[ - c(7, 32)]))
Residuals:
    Min      1Q    Median     3Q   Max
 -1.411 -0.2864 -0.002064 0.3173 1.508

Coefficients:
                        Value Std. Error t value Pr(>|t|)
           (Intercept)  2.1249  0.0925    22.9725  0.0000
log(body[ - c(7, 32)])  0.2545  0.0271     9.4017  0.0000
```

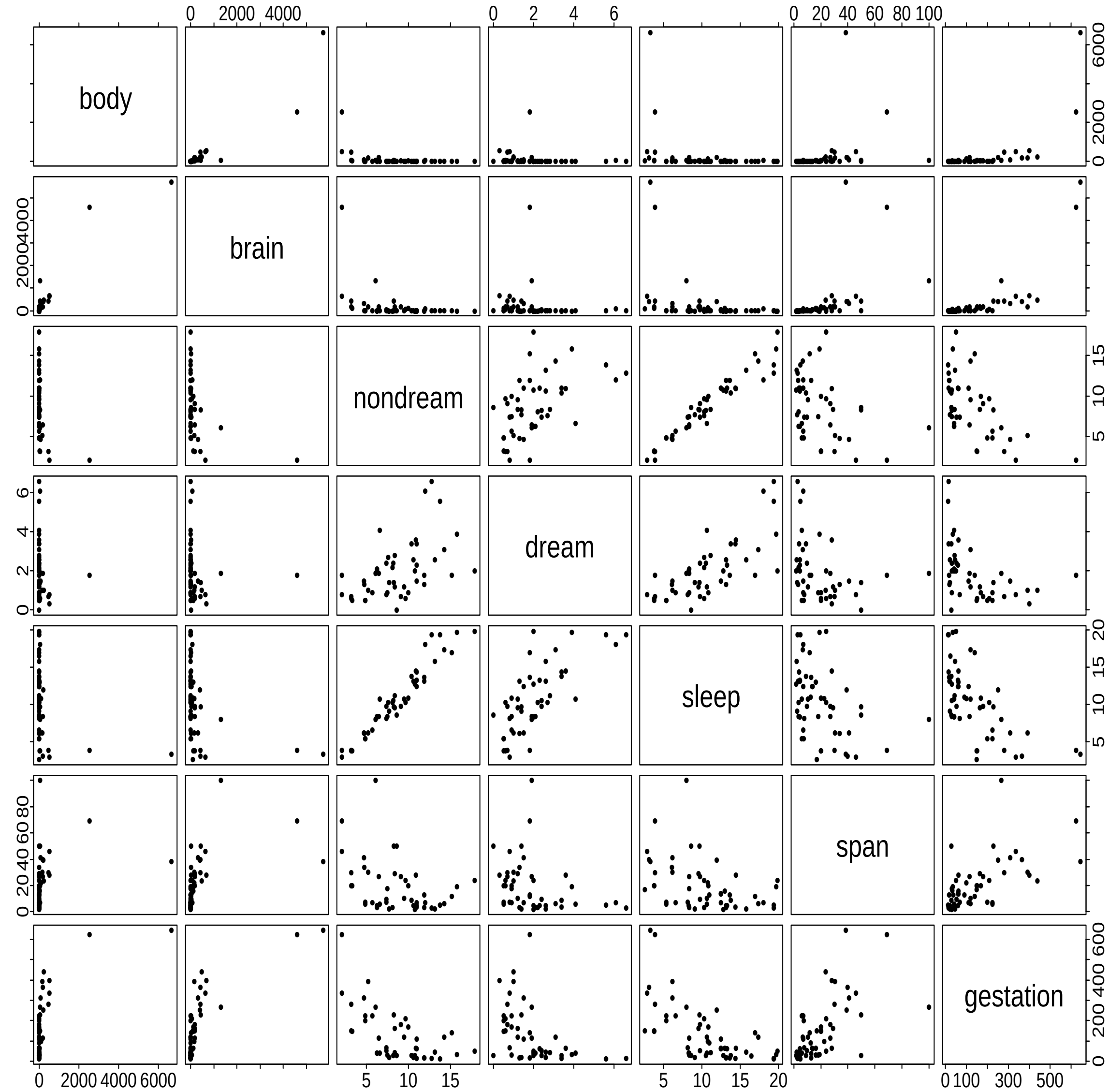


Figure 17: Scatterplot Matrix

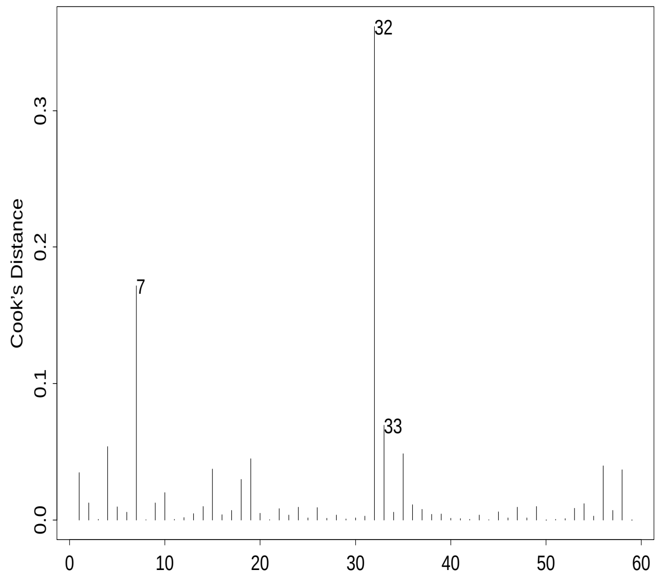


Figure 18: Cook distances

```
Residual standard error: 0.6122 on 55 degrees of freedom
Multiple R-Squared: 0.6164
F-statistic: 88.39 on 1 and 55 degrees of freedom, the p-value is 4.851e-13

Correlation of Coefficients:
                       (Intercept)
log(body[ - c(7, 32)]) -0.4811
```

### D.2 Brain on Span

Figure 19 shows a plot of the brain weight variable against the residuals of the regression between body weight and life span. No trend is observed ($cor(brain, body) > .9!$). The same strong influence problem given by the bats is observed in this simple regression; while this time the final R squared is .77 (so we will want to include brain instead of body in the multiple model).

```
Call: lm(formula = log(span[ - c(32, 7)]) ~ log(brain[ - c(32, 7)]))
Residuals:
   Min      1Q   Median     3Q   Max
 -1.25 -0.2142 -0.02734 0.2241 1.431

Coefficients:
                          Value Std. Error t value Pr(>|t|)
```

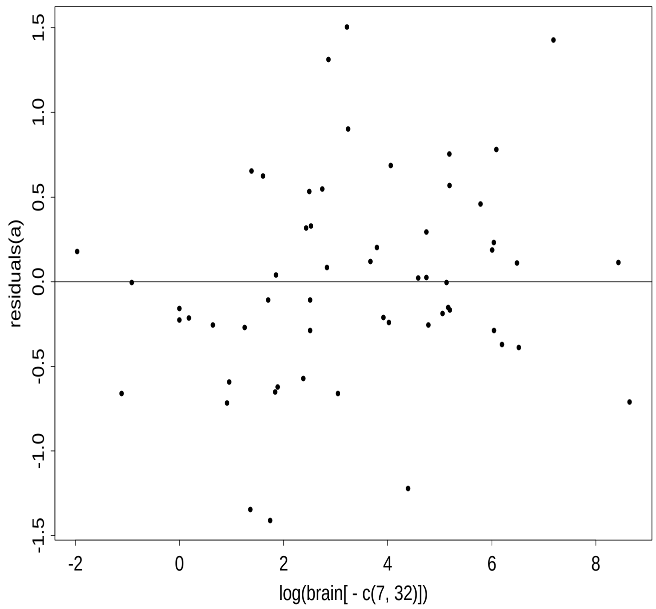


Figure 19: Brain on residuals of log(span) ˜log(body)

```
            (Intercept)  1.3105  0.1112    11.7834  0.0000
log(brain[ - c(32, 7)])  0.3637  0.0270    13.4665  0.0000

Residual standard error: 0.4769 on 55 degrees of freedom
Multiple R-Squared: 0.7673
F-statistic: 181.3 on 1 and 55 degrees of freedom, the p-value is 0

Correlation of Coefficients:
                        (Intercept)
log(brain[ - c(32, 7)]) -0.8231
```

### D.3 Sleep on span

There is not a strong relationship between total sleep and life span, though animals that sleep more tend to live less than the others. Figure 20 shows the scatterplot with the fitted line. The relationship is mild because no strong trend, especially linear, is observed in the scatterplot.

```
Call: lm(formula = log(span) ~ sleep)
Residuals:
    Min     1Q Median     3Q   Max
 -1.712 -0.727 0.1354 0.5725 1.833

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   3.5333   0.2959    11.9398   0.0000
```

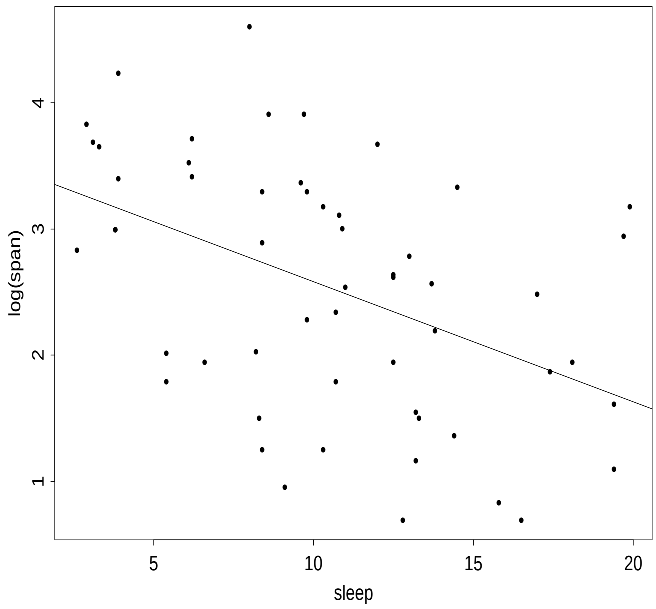


Figure 20: Sleep on log(span)

```
      sleep  -0.0952   0.0257    -3.7055   0.0005

Residual standard error: 0.8932 on 53 degrees of freedom
Multiple R-Squared: 0.2058
F-statistic: 13.73 on 1 and 53 degrees of freedom, the p-value is 0.0005049

Correlation of Coefficients:
      (Intercept)
sleep -0.9134
```

### D.4 Dream and nondream on Span

Even lighter relationship with dream and non-dream sleep was observed:

```
Call: lm(formula = log(span) ~ dream)
Residuals:
    Min      1Q Median     3Q   Max
 -1.788 -0.7752 0.1038 0.6634 2.097

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   3.0290   0.2245    13.4901   0.0000
      dream  -0.2741   0.0927    -2.9556   0.0050
```

```
Residual standard error: 0.9297 on 45 degrees of freedom
Multiple R-Squared: 0.1626
F-statistic: 8.736 on 1 and 45 degrees of freedom, the p-value is 0.004953

Correlation of Coefficients:
      (Intercept)
dream -0.797


Call: lm(formula = log(span) ~ nondream)
Residuals:
    Min      1Q   Median     3Q   Max
 -1.611 -0.8263 -0.04668 0.7126 1.882

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  3.3222  0.3606     9.2128  0.0000
   nondream -0.0982  0.0380    -2.5814  0.0133

Residual standard error: 0.9547 on 43 degrees of freedom
Multiple R-Squared: 0.1342
F-statistic: 6.664 on 1 and 43 degrees of freedom, the p-value is 0.01333

Correlation of Coefficients:
         (Intercept)
nondream -0.9188
```

### D.5 Gestation on Span

We can expect the gestation time to be important on life span, and in fact it achieves an R squared of 40%. The intercept is not significant:

```
Call: lm(formula = log(span) ~ log(gest))
Residuals:
    Min      1Q Median     3Q   Max
 -1.344 -0.4463 0.0468 0.4649 2.033

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -0.1154  0.4266    -0.2705  0.7877
  log(gest)  0.5986  0.0929     6.4417  0.0000

Residual standard error: 0.7424 on 57 degrees of freedom
Multiple R-Squared: 0.4213
F-statistic: 41.5 on 1 and 57 degrees of freedom, the p-value is 2.693e-08

Correlation of Coefficients:
```

```
          (Intercept)
log(gest) -0.974
```

That's the model without the intercept:

```
Call: lm(formula = log(span) ~ log(gest) - 1)
Residuals:
    Min      1Q  Median     3Q   Max
 -1.365 -0.4719 0.05715 0.4709 1.999

Coefficients:
            Value Std. Error t value Pr(>|t|)
log(gest)  0.5741  0.0209    27.4855  0.0000

Residual standard error: 0.7365 on 58 degrees of freedom
```

Anyway, we observe a strong positive relationship. So higher gestation times lead to longer lives.

### D.6 ANOVAs between Span and Danger Indexes

We can expect animals more exposed to dangers to be also likely to live shorter lives. Instead, Figures 21, 22 and 23 show that this is not true, and that in some cases animals exposed to dangers live more than the others (i.e., animals with an overall danger index of 5 live on average 27 years, while animals with an overall danger index of 3 live only 8.6 years). This implies that any significant relationship between life span and danger indexes would be due to the sample size, and misleading for us. For this reasons, we decided not to consider the danger indexes in the regression analysis of life duration. It is nevertheless interesting to conclude that life span is not linked to danger exposure. The expected life duration of a mammal is influenced only by constitutional characteristics.

The high outlier with index equal to 1 is, obviously, the Man.

### D.7 Body on Sleep times

```
Call: lm(formula = sleep[sleep != "NA"] ~ log(body[sleep != "NA"]))
Residuals:
    Min     1Q  Median   3Q  Max
 -6.699 -2.626 -0.2441 2.17 9.91

Coefficients:
                            Value Std. Error  t value Pr(>|t|)
             (Intercept)  11.4377   0.5510    20.7588   0.0000
log(body[sleep != "NA"])  -0.7931   0.1683    -4.7120   0.0000

Residual standard error: 3.933 on 56 degrees of freedom
Multiple R-Squared: 0.2839
F-statistic: 22.2 on 1 and 56 degrees of freedom, the p-value is 1.664e-05
```

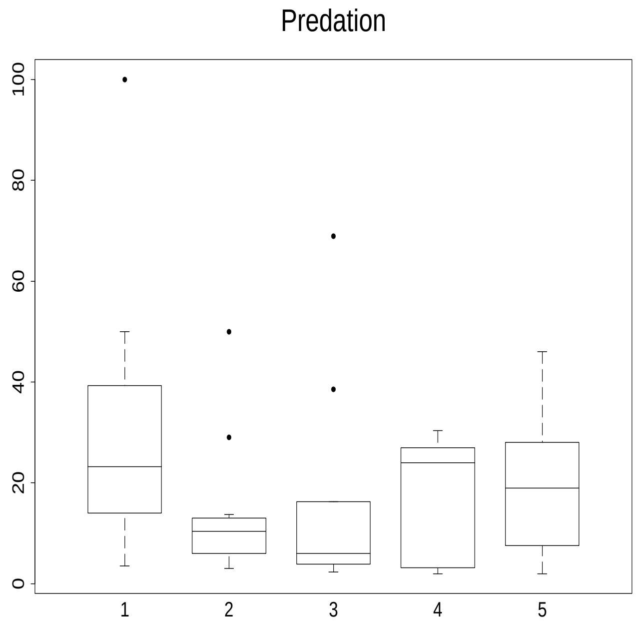


Figure 21: Span on Predation index

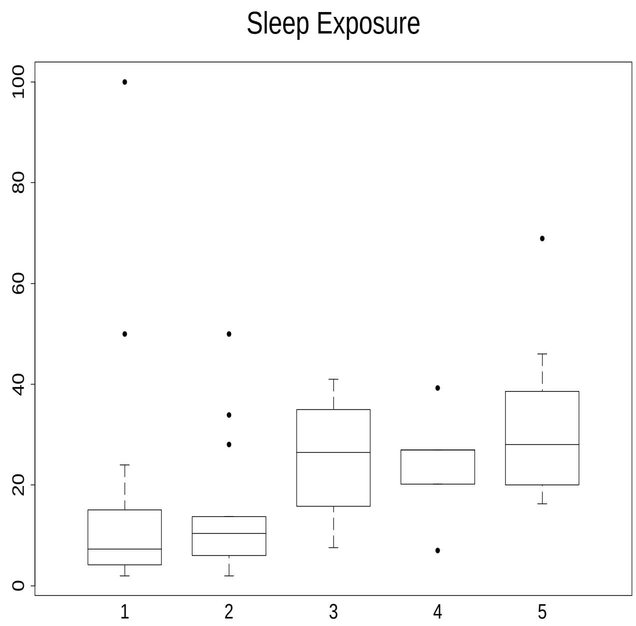


Figure 22: Span on Sleep Exposure index

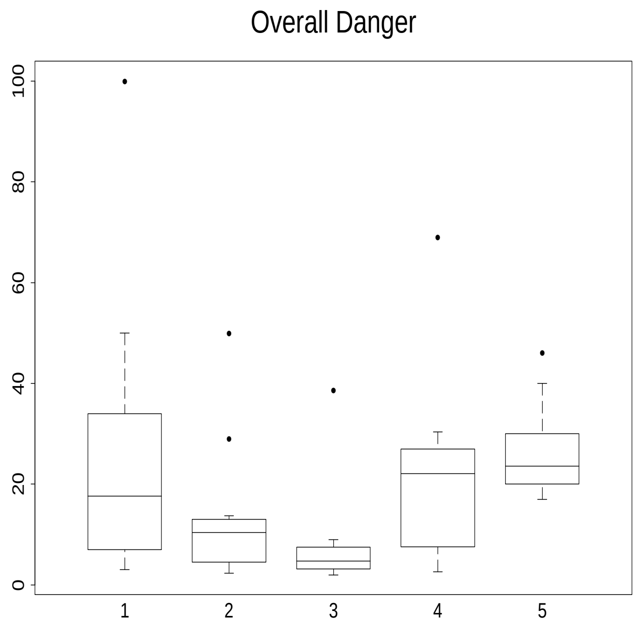


Figure 23: Span on Overall Danger index

```
Correlation of Coefficients:
                          (Intercept)
log(body[sleep != "NA"]) -0.3486
```

When we regress the dream time on body weight, we get a non significant slope; but the Giant Armadillo has got a residual of 4.47 and a high Cook D. Even without this outlier in the analysis, no significant effect is detected.

The non dreaming time has got instead a strongly significant negative slope (-.72) and a higher Multiple R-Squared (0.3412).

### D.8 Brain on sleep times

```
Call: lm(formula = sleep[sleep != "NA"] ~ log(brain[sleep != "NA"]))
Residuals:
    Min      1Q  Median    3Q   Max
 -6.703 -3.241 0.04978 2.516 9.021

Coefficients:
                              Value Std. Error  t value Pr(>|t|)
              (Intercept)  13.7245   0.8046    17.0576   0.0000
log(brain[sleep != "NA"])  -1.0571   0.2076    -5.0924   0.0000

Residual standard error: 3.842 on 56 degrees of freedom
Multiple R-Squared: 0.3165
F-statistic: 25.93 on 1 and 56 degrees of freedom, the p-value is 4.297e-06
```

```
Correlation of Coefficients:
                          (Intercept)
log(brain[sleep != "NA"]) -0.779


Call: lm(formula = nondream[nondream != "NA"] ~ log(brain[nondream != "NA"]))
Residuals:
    Min      1Q   Median    3Q   Max
 -5.011 -2.703 -0.04862 2.604 6.587

Coefficients:
                                Value Std. Error  t value Pr(>|t|)
                 (Intercept)  10.9991   0.6570    16.7416   0.0000
log(brain[nondream != "NA"])  -0.8705   0.1832    -4.7514   0.0000

Residual standard error: 3.035 on 46 degrees of freedom
Multiple R-Squared: 0.3292
F-statistic: 22.58 on 1 and 46 degrees of freedom, the p-value is 2.014e-05

Correlation of Coefficients:
                             (Intercept)
log(brain[nondream != "NA"]) -0.7452
```

There is a significant relationship even with dreaming time:

```
Call: lm(formula = dream[dream != "NA"] ~ log(brain[dream != "NA"]))
Residuals:
    Min      1Q Median     3Q   Max
 -1.894 -0.8859 -0.309 0.5812 4.437

Coefficients:
                            Value Std. Error t value Pr(>|t|)
              (Intercept)  2.5249  0.2958     8.5360  0.0000
log(brain[dream != "NA"]) -0.1961  0.0792    -2.4772  0.0168

Residual standard error: 1.373 on 48 degrees of freedom
Multiple R-Squared: 0.1134
F-statistic: 6.137 on 1 and 48 degrees of freedom, the p-value is 0.01682

Correlation of Coefficients:
                          (Intercept)
log(brain[dream != "NA"]) -0.7546
```

### D.9 Life span on Sleep Times

```
Call: lm(formula = sleep[sleep != "NA" & span != "NA"] ~ log(span[sleep != "NA" &
```

```
        span != "NA"]))
Residuals:
    Min   1Q Median    3Q   Max
 -7.272 -3.2 0.5489 2.161 10.77

Coefficients:
                                           Value Std. Error  t value Pr(>|t|)
                            (Intercept)  15.9967   1.5846    10.0950   0.0000
log(span[sleep != "NA" & span != "NA"])  -2.1619   0.5834    -3.7055   0.0005

Residual standard error: 4.257 on 53 degrees of freedom
Multiple R-Squared: 0.2058
F-statistic: 13.73 on 1 and 53 degrees of freedom, the p-value is 0.0005049

Correlation of Coefficients:
                                        (Intercept)
log(span[sleep != "NA" & span != "NA"]) -0.9321


Call: lm(formula = dream[dream != "NA" & span != "NA"] ~ log(span[dream != "NA" &
        span != "NA"]))
Residuals:
   Min      1Q  Median     3Q   Max
 -1.85 -0.8757 -0.3418 0.4573 3.842

Coefficients:
                                          Value Std. Error t value Pr(>|t|)
                            (Intercept)  3.4126  0.5399     6.3208  0.0000
log(span[dream != "NA" & span != "NA"]) -0.5931  0.2007    -2.9556  0.0050

Residual standard error: 1.368 on 45 degrees of freedom
Multiple R-Squared: 0.1626
F-statistic: 8.736 on 1 and 45 degrees of freedom, the p-value is 0.004953

Correlation of Coefficients:
                                        (Intercept)
log(span[dream != "NA" & span != "NA"]) -0.9292

Call: lm(formula = nondream[nondream != "NA" & span != "NA"] ~ log(span[nondream !=
        "NA" & span != "NA"]))
Residuals:
   Min     1Q Median    3Q   Max
 -4.79 -3.036 0.6157 1.959 10.16

Coefficients:
                                              Value Std. Error  t value
                               (Intercept)  12.0853   1.4101     8.5707
log(span[nondream != "NA" & span != "NA"])  -1.3669   0.5295    -2.5814
```

```
                                           Pr(>|t|)
                              (Intercept)   0.0000
log(span[nondream != "NA" & span != "NA"])   0.0133

Residual standard error: 3.562 on 43 degrees of freedom
Multiple R-Squared: 0.1342
F-statistic: 6.664 on 1 and 43 degrees of freedom, the p-value is 0.01333

Correlation of Coefficients:
                                           (Intercept)
log(span[nondream != "NA" & span != "NA"]) -0.9264
```

### D.10 ANOVAs between Sleep and Danger Indexes

```
                              Df Sum of Sq  Mean Sq  F Value      Pr(F)
as.factor(pred[sleep != "NA"])  4  243.6218 60.90545 3.341444 0.01635581
                     Residuals 53  966.0460 18.22728

                                     Df Sum of Sq  Mean Sq  F Value        Pr(F)
as.factor(slexpo[sleep != "NA"])  4  573.6617 143.4154 11.95117 5.414503e-07
                       Residuals 53  636.0060  12.0001

                                    Df Sum of Sq  Mean Sq  F Value        Pr(F)
as.factor(danger[sleep != "NA"])  4  457.2556 114.3139 8.052284 3.779999e-05
                       Residuals 53  752.4122  14.1965

                              Df Sum of Sq  Mean Sq  F Value      Pr(F)
as.factor(pred[dream != "NA"])  4  22.13698 5.534246 3.119103 0.02390248
                     Residuals 45  79.84382 1.774307

                                Df Sum of Sq  Mean Sq  F Value       Pr(F)
as.factor(slexpo[dream != "NA"])  4  33.58751 8.396878 5.524804 0.001049436
                       Residuals 45  68.39329 1.519851


                                Df Sum of Sq  Mean Sq  F Value        Pr(F)
as.factor(danger[dream != "NA"])  4  36.29911 9.074778 6.217335 0.0004523689
                       Residuals 45  65.68169 1.459593
```

## E Multiple Models

### E.1 Sleep Model

We found significant relationships of all the variables with the sleep times. Since we are interested mainly in the relationship between sleep and danger exposures, the model will be useful especially to detect the relationship between this variables net of the other significant ones. But, since the correlation between the constitutional variables is so high, and the relationship with sleep times not so strong; we can expect some of them to become non significant in the full model. This is the case, since all of the variables become insignificant if we include dream in the model:

```
              Value Std. Error t value Pr(>|t|)
(Intercept)  7.9973  1.8820     4.2494  0.0002
       body  0.0015  0.0061     0.2381  0.8136
      brain -0.0007  0.0038    -0.1866  0.8533
      dream  1.6120  0.5053     3.1903  0.0035
       span -0.0112  0.0522    -0.2144  0.8318
  gestation -0.0078  0.0077    -1.0199  0.3165
      pred1  0.1407  1.0376     0.1356  0.8931
      pred2 -0.2003  0.8632    -0.2321  0.8182
      pred3  0.6075  0.7872     0.7718  0.4467
      pred4  0.1239  0.5942     0.2085  0.8364
    slexpo1 -0.6611  0.7659    -0.8632  0.3954
    slexpo2 -0.2626  0.7424    -0.3537  0.7262
    slexpo3  0.0365  0.6880     0.0530  0.9581
    slexpo4 -0.3022  0.8359    -0.3615  0.7204
    danger1  0.2881  1.2187     0.2364  0.8148
    danger2 -0.7488  0.9675    -0.7739  0.4454
    danger3 -0.2030  0.9759    -0.2080  0.8367
    danger4 -0.5481  1.0762    -0.5093  0.6146

Residual standard error: 3.108 on 28 degrees of freedom
Multiple R-Squared: 0.7256
F-statistic: 4.355 on 17 and 28 degrees of freedom, the p-value is 0.0002949
```

The problem here is that the indexes take away too many degrees of freedom from the residuals. A good idea[14] is using an automatic variable selection method, like the stepwise, which ends with the model:

```
sleep ~ dream + log(span) + log(body).
```

This is a summary of this model:

```
Call: lm(formula = sleep ~ dream + log(span) + log(body), data = slold, na.action =
        na.omit)
Residuals:
    Min     1Q   Median    3Q   Max
 -4.928 -1.815 0.009719 1.538 5.519

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  4.1445  1.6052     2.5819  0.0134
      dream  2.2963  0.2819     8.1460  0.0000
  log(span)  1.0408  0.5399     1.9276  0.0607
  log(body) -0.8410  0.1673    -5.0277  0.0000
```

[14]For avoiding many unsuccessful trials.

```
Residual standard error: 2.552 on 42 degrees of freedom
Multiple R-Squared: 0.7225
F-statistic: 36.45 on 3 and 42 degrees of freedom, the p-value is 9.188e-12
```

In a sense, the strong effect of dream on the total sleep is hiding the effects of the danger indexes. If we drop dream from the set of possible variables and repeat the stepwise, we end up with the following model:

```
Call: lm(formula = sleep ~ danger + log(brain), data = slold, na.action = na.omit)
Residuals:
    Min     1Q  Median    3Q   Max
 -7.345 -1.857 0.01351 2.017 6.085

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  12.4515   0.7564    16.4617   0.0000
    danger1  -1.1962   0.5810    -2.0587   0.0446
    danger2  -0.8786   0.3888    -2.2602   0.0280
    danger3  -0.4103   0.3029    -1.3545   0.1814
    danger4  -0.9703   0.2738    -3.5434   0.0008
 log(brain)  -0.8856   0.1906    -4.6455   0.0000

Residual standard error: 3.198 on 52 degrees of freedom
Multiple R-Squared: 0.5604
F-statistic: 13.26 on 5 and 52 degrees of freedom, the p-value is 2.453e-08
```

### E.2 Life Span Model

Again, a stepwise variable selection is performed, ending with the following model:

```
log(span) ~ dream + sleep + pred + log(brain)
```

This is a summary of the model:

```
Call: lm(formula = log(span) ~ dream + sleep + pred + log(brain), data = slold,
        na.action = na.omit)
Residuals:
     Min     1Q   Median     3Q   Max
 -0.9822 -0.301 -0.05648 0.3666 1.167

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  1.6345  0.3583     4.5617  0.0001
      dream -0.2954  0.0879    -3.3591  0.0018
      sleep  0.0528  0.0331     1.5943  0.1192
      pred1 -0.2860  0.1179    -2.4266  0.0201
      pred2 -0.2502  0.0804    -3.1119  0.0035
      pred3 -0.0757  0.0611    -1.2382  0.2232
```

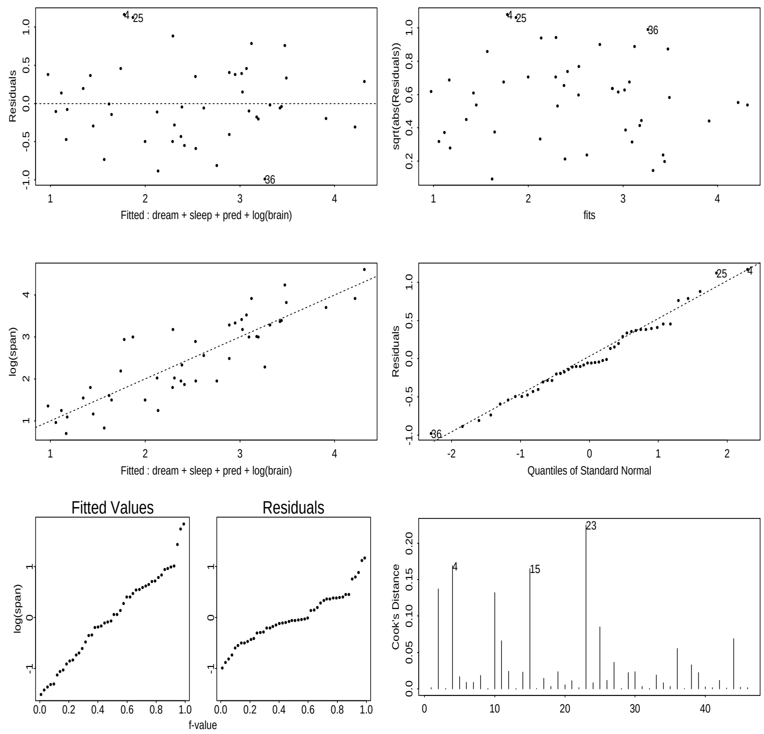

Figure 24: Residual graphs for Span multiple model

```
      pred4 -0.0107  0.0440    -0.2434  0.8090
 log(brain)  0.3058  0.0431     7.0901  0.0000

Residual standard error: 0.5444 on 38 degrees of freedom
Multiple R-Squared: 0.7527
F-statistic: 16.52 on 7 and 38 degrees of freedom, the p-value is 8.965e-10
```

The R squared is really good. Because of the collinearity, we can't interpret directly the parameters. I.e., dream and sleep parameter estimates have different sign, though the same effect on span was detected in the simple regression models.

Figure 24 shows fit plots for this model. Everything is ok, but we see that the bats are strongly influential (high residual and cookd). They however have a leverage of .19 and .31, which are below the $2 * \bar{h} = .35$ threshold.

If we fit the model without the bats we get different estimates, especially for the dream and sleep parameters, and a sensibly better fit (R squared=.84). Note that total sleep is not significant at all. This is interesting, and happens because the bats are outliers in the total sleep variable. Without them, there isn't as much information in sleep on life span.

```
Call: lm(formula = log(span) ~ dream + sleep + pred + log(brain), data = slold[
        row.names(slold) != "Little.brown.bat" & row.names(slold) !=
        "Big.brown.bat",  ], na.action = na.omit)
```

```
Residuals:
     Min      1Q   Median     3Q   Max
 -0.7508 -0.2508 -0.03489 0.2026 1.041

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  1.8279  0.3016     6.0608  0.0000
      dream -0.1790  0.0801    -2.2359  0.0317
      sleep -0.0011  0.0307    -0.0350  0.9722
      pred1 -0.0538  0.1116    -0.4821  0.6326
      pred2 -0.2107  0.0674    -3.1283  0.0035
      pred3 -0.0321  0.0519    -0.6185  0.5401
      pred4 -0.0144  0.0365    -0.3932  0.6965
 log(brain)  0.3322  0.0363     9.1445  0.0000

Residual standard error: 0.452 on 36 degrees of freedom
Multiple R-Squared: 0.8358
F-statistic: 26.18 on 7 and 36 degrees of freedom, the p-value is 2.571e-12
```

### E.3 Smart Mammals model: Logistic Regressions

In this section, we will give a short description of the polytomic logistic regression model, and some summaries of the fitted models.

A polytomic logistic regression model is a model in which the response variable is qualitative, and has got more than two levels. In a sense, this is simply a generalization of the usual logistic model. The link function used is again the logit ($g(x) = \ln(\frac{x}{1-x})$).

The logistic model is:

$$logit(p) = \beta_0 + \sum \beta_i x_i;$$

where $x_i$ is the $i$-th covariate, the $\beta$s are the parameters and $p$ is the expected value of $Y_i$ (i.e., the probability that $Y_i = 1$), the two-level qualitative response variable.

Supposing there are $k \geq 2$ levels of the response variable, the polytomic logistic model becomes:

$$logit(p_j) = \beta_{0j} + \sum \beta_{ij} x_i$$

where this time $p_j$ is $\Pr(Y_i = a_j)$, $a_j$ being the $j$-th level of the response.

The parametrization proposed is the most straightforward, but many different parameterizations can be used [15].In particular, if the response variable is ordered (like in our case), a more informative model is the following: called

$$F_i = \frac{\sum_{j=1}^{i} p_j}{\sum_{j=1}^{k} p_j},$$

the cumulative logit, we could fit:

$$logit(F_i) = \beta_{0i} + \sum \beta_j x_j.$$

[15] See [4, Pages 253-256] for more details.

This is the "proportional odds model"[16].

> "In other words, the proportional odds model assumes that each logit follows a linear model which has a separate intercept parameter but other regression parameters (i.e., the slopes relating the response to each covariate) are constant across all cumulative logits for different levels of the response[17]."

Note that there are $k-1$ cumulative logits (the $k$-th is always one, and so it is useless fitting a model on it). Moreover, it can be shown that the intercepts are bounded by the following constraint:

$$\beta_{0i} \leq \beta_{0(i+1)}, i = 1, \ldots, k-2.$$

This is the model we will fit; so our models will have 4 intercepts and one parameter estimate for each covariate; and will refer to the cumulative logits.

These are the fitted models:

```
Call:
polr(formula = danger ~ sleep + log(brain) + log(gestation) + log(body), data
        = sl, na.action = na.omit)

Coefficients:
                     Value Std. Error     t value
         sleep -0.36531240 0.08666771 -4.21509232
    log(brain) -0.47361208 0.40313876 -1.17481157
log(gestation) -0.03469964 0.45154646 -0.07684623
     log(body)  0.24158509 0.29468163  0.81981726

Intercepts:
      Value Std. Error t value
1|2 -6.2966  2.3487    -2.6809
2|3 -4.9400  2.2928    -2.1546
3|4 -3.9215  2.2756    -1.7233
4|5 -2.5650  2.2493    -1.1404

Residual Deviance: 153.1676
AIC: 169.1676

Call:
polr(formula = slexpo ~ sleep + log(brain) + log(gestation) + log(body), data
        = sl, na.action = na.omit)

Coefficients:
                     Value Std. Error     t value
         sleep -0.24773559 0.08348445 -2.96744596
    log(brain)  0.03015135 0.44662589  0.06750918
log(gestation)  0.71844888 0.51804459  1.38684755
```

[16]See [5] for more details on this model.
[17][2, Pag.147]

```
     log(body)  0.15236088 0.31679481  0.48094500

Intercepts:
      Value Std. Error t value
1|2  0.2472  2.3632     0.1046
2|3  1.7864  2.4116     0.7408
3|4  2.3909  2.4254     0.9858
4|5  3.3610  2.4206     1.3885

Residual Deviance: 121.548
AIC: 137.548

Call:
polr(formula = pred ~ sleep + log(brain) + log(gestation) + log(body), data =
        sl, na.action = na.omit)

Coefficients:
                    Value Std. Error    t value
         sleep -0.2781336 0.07667924 -3.6272346
    log(brain) -0.6155270 0.39171391 -1.5713688
log(gestation) -0.1970252 0.40778564 -0.4831587
     log(body)  0.3246381 0.27646282  1.1742558

Intercepts:
      Value Std. Error t value
1|2 -6.7013  2.1835    -3.0691
2|3 -5.2701  2.1072    -2.5010
3|4 -4.3975  2.0844    -2.1097
4|5 -3.6670  2.0689    -1.7724

Residual Deviance: 166.50
AIC: 182.50
```

Including dream in the model we see that, because of collinearity, sleep's $t$ value goes to 0 and dream takes an high $t$ value. So we may want to conclude that dream is more important than sleep in the smartness of an animal.

```
Call:
polr(formula = slexpo ~ sleep + dream + log(brain) + log(gestation) + log(
        body), data = sl, na.action = na.omit)

Coefficients:
                     Value Std. Error    t value
         sleep -0.02751533  0.1236481 -0.2225293
         dream -1.02107039  0.4520310 -2.2588505
    log(brain) -0.39531189  0.5009308 -0.7891547
log(gestation)  0.47255048  0.5524379  0.8553911
     log(body)  0.65099327  0.3955019  1.6459927
```

```
Intercepts:
      Value Std. Error t value
1|2 -1.2612  2.5144    -0.5016
2|3  0.4045  2.5317     0.1598
3|4  1.2108  2.5398     0.4767
4|5  2.2169  2.5331     0.8751

Residual Deviance: 94.90745
AIC: 112.9074



Call:
polr(formula = pred ~ sleep + dream + log(brain) + log(gestation) + log(body),
        data = sl, na.action = na.omit)

Coefficients:
                      Value Std. Error    t value
         sleep -0.03921101  0.1092825 -0.3588041
         dream -0.98276467  0.3865036 -2.5427052
    log(brain) -0.98590449  0.4439130 -2.2209410
log(gestation) -0.68799693  0.5107880 -1.3469324
     log(body)  0.89871188  0.3563186  2.5222146

Intercepts:
      Value Std. Error t value
1|2 -9.1864  2.6909    -3.4139
2|3 -7.5407  2.5823    -2.9202
3|4 -6.4874  2.5240    -2.5702
4|5 -5.4743  2.4750    -2.2119

Residual Deviance: 133.429
AIC: 151.429

Call:
polr(formula = danger ~ sleep + dream + log(brain) + log(gestation) + log(
        body), data = sl, na.action = na.omit)

Coefficients:
                      Value Std. Error    t value
         sleep -0.07741362  0.1140785 -0.6785998
         dream -1.33421118  0.4395928 -3.0351067
    log(brain) -0.99584498  0.4610082 -2.1601458
log(gestation) -0.54103280  0.5485306 -0.9863311
     log(body)  0.98395341  0.3777959  2.6044573

Intercepts:
```

```
      Value Std. Error t value
1|2 -9.2904  2.8701    -3.2369
2|3 -7.6562  2.7561    -2.7779
3|4 -6.5340  2.7110    -2.4102
4|5 -4.6225  2.6288    -1.7584

Residual Deviance: 116.8024
AIC: 134.8024
```

For a better interpretation of the parameters, let's do two simple models on sleep exposure, which is the most interesting for our purposes:

```
Call:
polr(formula = slexpo ~ sleep, data = sl, na.action = na.omit)

Coefficients:
      Value Std. Error   t value
 -0.3663992 0.07912708 -4.630516

Intercepts:
      Value Std. Error t value
1|2 -4.2350  0.8930    -4.7423
2|3 -2.9342  0.8098    -3.6235
3|4 -2.4663  0.7889    -3.1262
4|5 -1.6695  0.7441    -2.2437

Residual Deviance: 132.292
AIC: 142.292

Call:
polr(formula = slexpo ~ dream, data = sl, na.action = na.omit)

Coefficients:
     Value Std. Error  t value
 -1.358429  0.3724438 -3.64734

Intercepts:
      Value Std. Error t value
1|2 -2.7198  0.7272    -3.7400
2|3 -1.4926  0.6195    -2.4093
3|4 -0.9857  0.6007    -1.6410
4|5 -0.4029  0.6008    -0.6706

Residual Deviance: 116.3317
AIC: 126.3317
```