Computer exercise 2: Multiple linear regression

Exercise 2: Sleep in Mammals

A study on sleep in mammals collected, among other things, the body weight, the average time spent sleeping during a day (24 hours), and a danger index for 57 species of mammals. Of interest is to predict the time a species spends sleeping, using the body weight and the danger index.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TotalSleep</td>
<td>total sleep (hrs/day)</td>
</tr>
<tr>
<td>BodyWt</td>
<td>body weight (kg)</td>
</tr>
<tr>
<td>Danger</td>
<td>overall danger index (1–3)</td>
</tr>
<tr>
<td></td>
<td>1 = low danger (from other animals);</td>
</tr>
<tr>
<td></td>
<td>2 = medium danger (from other animals);</td>
</tr>
<tr>
<td></td>
<td>3 = high danger (from other animals)</td>
</tr>
</tbody>
</table>

The data for this exercise is available as a tab-separated txt-file, sleep.txt, on the course web page. Download and save it to your R working directory and then read it into R. Since the danger index is a categorical variable, we should also turn it into a factor in R:

```r
sleep <- read.delim("sleep.txt")
sleep$Danger <- factor(sleep$Danger, levels=c(1,2,3),
                       labels=c("low","medium","high"))
```

(a) Let’s start with some exploratory analysis. Make a frequency table of the Danger variable. How many observations are there in each category?

(b) Calculate the average daily sleep in each of the three Danger categories. Then plot TotalSleep versus Danger (this should become a boxplot) and compare with the calculated averages.

(c) Fit the linear model $\text{TotalSleep}_i = \beta_0 + \beta_1 \text{Danger}_{\text{medium},i} + \beta_2 \text{Danger}_{\text{high},i} + \epsilon_i$. Use the estimates to calculate the predicted mean in each category "by hand" (i.e., not using `predict()`). Compare your result with the category averages you calculated above and make sure you understand how the $\beta$-parameters relate to the category averages.

(d) Use a (global) F-test to test whether we need the Danger variable. Then test if it is necessary to have three groups, or whether we could merge two of them.

(e) Obviously the amount mammals sleep varies with the amount of danger. We now want to see if body weight affects the sleep. Plot total sleep vs body weight. Does a linear relationship seem reasonable? Redraw the plot using the log of the body weight instead. Does this seem more reasonable? Why might this be more reasonable?

(f) First fit a model using log(BodyWt), but not Danger. Then fit a new model by adding Danger as well. Use a (partial) F-test to test whether we need the Danger variable now. Do we need all three groups?
(g) Calculate confidence intervals for the four $\beta$-parameters.

(h) Homo Sapiens is a mammal with body weight 62 kg and low Danger. Predict its expected total sleep together with a 95 % confidence interval.

(i) The Marmot and the Vervet both weigh 4 kg but the Marmot lives in low danger and the Vervet in high danger. How many hours less would the Vervet be expected to sleep, compared to the Marmot? Write down the theoretical value, estimate it and calculate a 95 % confidence interval.

(j) How many hours longer would the small Marmot be expected to sleep, compared to the large Homo Sapiens? Write down the theoretical value, estimate it and calculate a 95 % confidence interval.

(k) How many hours longer (or shorter) would the small, high danger Vervet be expected to sleep, compared to the large, low danger Homo Sapiens? Write down the theoretical value, estimate it and calculate a 95 % confidence interval.

*Hint:* since $V(a\hat{\beta}) = a \text{Var}(\hat{\beta})a'$, you will need the covariance matrix for the $\beta$-estimates.
Useful R-commands

Se also Computer exercise 1!

• Creating a factor variable by replacing the values (1, 2, 3) by the category names (name 1, name 2, name 3):
  ```r
  mydata$x <- factor(mydata$x,
                     levels=c(1,2,3),
                     labels=c("name 1","name 2","name 3"))
  ```
  The `lm`-command uses the first level of a factor variable as reference group. You can change reference group to, e.g., name 2, by
  ```r
  relevel(mydata$x, "name 2")
  ```

• Tabulate a variable: `table(mydata$x)`.

• In `mydata`, calculate the mean of variable `y`, for the separate categories in variable `x`:
  ```r
  aggregate(y~x, data=mydata, FUN=mean)
  ```

• Set the limits of the `y`-axis: `plot(..., ylim=c(0,24))`
  Add your own axis-labels: `plot(..., ylab="Total sleep (hrs/day)").`
  ... and plot title: `plot(..., main="Mammalian sleep").`
  Plot with log-scale on the x-axis: `plot(..., log="x")`.

• Estimate a multiple regression
  ```r
  lm(y~x1+x2, data=mydata)
  ```
  If you want to transform a variable in the model you don’t have to calculate it first:
  ```r
  lm(y~log(x3), data=mydata)
  ```
  Update an already estimated model by adding a new variable:
  ```r
  update(oldmodel, .~.+x4)
  ```
  The dot stands for “whatever was here in the old model”.

• `anova(model1)` creates the (sequential) ANOVA table corresponding to `model1`.
  `anova(model1,model2)` compares `model1` with `model2` via a Partial F-test. For ease of interpretation, start with the smallest model. The comparison between two or more models will only be valid if they are fitted to the same dataset.

• Extract $(X'X)^{-1}$ for a model:
  ```r
  summary(model)$cov.unscaled
  ```

• Matrix algebra:
  Create a (row) matrix $a = (1, 4, 3)$: $a <- matrix(c(1,4,3),nrow=1)$
  Matrix multiplication: use the operator `%*%`, e.g. $A%*%B$
  Element-wise multiplication: use `*`, e.g. $A*B$.
  Transposition: use `t()` e.g. `t(A)` to transpose a matrix $A$. 