# Department of Statistics

## COURSE STATS 330

Assignment 2, 2006

Instructions: Hand in your completed assignment to the Student Resource Centre by 4pm on Thursday 17th August.

**Note: You must show all R code used in your answers. Failure to do this will cost you marks.**

## Question 1 [30 marks]

The file **`mammalsleep.txt`** contains data on 61 species of mammal. The variables in the data set are

| | |
|---|---|
| **`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 dream and nondream) |
| **`lifespan:`** | maximum life span (years) |
| **`gestation:`** | gestation time (days) |
| **`predation:`** | predation index (1-5) 1 = minimum (least likely to be preyed upon) to 5 = maximum (most likely to be preyed upon) |
| **`exposure:`** | sleep exposure index (1-5) 1 = least exposed (e.g. animal sleeps in a well-protected den) 5 = most exposed |
| **`danger:`** | overall danger index (1-5) (based on the above two indices and other information) 1 = least danger (from other animals) 5 = most danger (from other animals) |

.

The data were gathered in a study to examine what factors are related to the sleep patterns of mammals. See the article "Sleep in Mammals: Ecological and Constitutional Correlates" Allison, T. and Cicchetti, D. (1976), Science, November 12, vol. 194, pp. 732-734. You can download a pdf copy of this article from the Library webpages.

*a.* Load the data into R and create a data frame. Make a pairs plot. Correct any obvious typos in the data. *[3 marks]*

b. The three variables **`body`**, **`brain`** and **`lifespan`** have values that differ by orders of magnitude. Do you think the relationships between these variables and the variable **`sleep`** are linear? If we consider logged versions of these variables, would this difficulty go away? Would another transformation be suitable? *[3 marks]*

*c.* There are also many missing values in this data. When fitting regressions, the **`lm`** function drops out any cases that have missing values in any of the variables involved in the regression. However, this causes difficulties when we come to compare models, as we do in part 4 of this question. What should we do? See tutorial sheet 2 for some hints. Take the appropriate corrective action. *[3 marks]*

d. Fit a model to the data, using sleep as the response and taking into account suitable transformations of the variables body, brain and lifespan. Comment on the quality of the fit. *[4 marks]*

e. In their article, Allison and Cicchetti fitted a model with log **`body`** and **`danger`** as explanatory variables, using **`nondream`** as response . Do these variables adequately explain the variation in the variable **`sleep`**? (in other words, is the model with log **`body`** and **`danger`** as explanatory variables an adequate model?) *[3 marks]*

*f.* It would seem that both the size of the animal, and the danger/risk aspect effect sleep. Does adding other variables improve the model? Try the effect of variables log **`brain`** and **`predation`** only – I don't expect a full-scale model selection exercise. *[4 marks]*

g. Subject you chosen model to the usual diagnostic checks. Are there any high-leverage observations or outliers? You should take your answer into Question 2 into account when answering this. *[6 marks]*

h. The "Okapi" species did not have the variable **`sleep`** recorded in this data set. Use your model to make a prediction of what the value of **`sleep`** is for this animal. *[4 marks]*

**Okapi**

## Question 2 [10 marks]

How can we tell if the largest residual in a regression is truly an outlier, or if it is just due to chance? After all, one of the residuals must be the biggest! One way to answer this is the following. If the model assumptions are satisfied (i.e. no outliers), we can think of the studentised residuals as being (at least approximately) random samples from a standard normal N(0,1) distribution. Suppose there are *n* residuals in a regression. We could simulate a sample of size *n* from a standard normal, and select the largest residual. We could repeat this say 10,000 times and examine the distribution of these 10,000 "largest" residuals. We could then see if the largest residual in our regression was extreme in comparison to this distribution. If it is, the observation is probably a genuine outlier.

Carry out this program for the final model you fitted in Question 1. What do you conclude? *[6 marks for R code, and display of results, 4 marks for comments]*

**Hint:** We can interpret the "largest residual" as the one having largest absolute value. We can calculate this for a normal sample of size *n* using the R statement

```
maxres = max(abs(rnorm(n)))
```

You can collect the results of doing this 10,000 times in a vector with 10,000 elements by using a loop. Display the results as we did for the exchange rate data as we did in Lecture 2. You may find consulting the R documentation for the functions **`rnorm, for, hist`** and **`density`** useful.

**Data for Assignment 2** (also in file mammalsleep.txt)

| | body | brain | nondream | dream | sleep | lifespan | gestation | predation | exposure | danger |
|---|---|---|---|---|---|---|---|---|---|---|
| African.elephant | 6654.000 | 5712.00 | NA | NA | 3.3 | 38.6 | 645.0 | 3 | 5 | 3 |
| African.giant.pouched.rat | 1.000 | 6.60 | 6.3 | 2.0 | 8.3 | 4.5 | 42.0 | 3 | 1 | 3 |
| Arctic.Fox | 3.385 | 44.50 | NA | NA | 12.5 | 14.0 | 60.0 | 1 | 1 | 1 |
| Arctic.ground.squirrel | 0.920 | 5.70 | NA | NA | 16.5 | NA | 25.0 | 5 | 2 | 3 |
| Asian.elephant | 2547.000 | 4603.00 | 2.1 | 1.8 | 3.9 | 69.0 | 624.0 | 3 | 5 | 4 |
| Baboon | 10.550 | 179.50 | 9.1 | 0.7 | 9.8 | 27.0 | 180.0 | 4 | 4 | 4 |
| Big.brown.bat | 0.023 | 0.30 | 15.8 | 3.9 | 19.7 | 19.0 | 35.0 | 1 | 1 | 1 |
| Brazilian.tapir | 160.000 | 169.00 | 5.2 | 1.0 | 6.2 | 30.4 | 392.0 | 4 | 5 | 4 |
| Cat | 3.300 | 25.60 | 10.9 | 3.6 | 14.5 | 28.0 | 63.0 | 1 | 2 | 1 |
| Chimpanzee | 52.160 | 440.00 | 8.3 | 1.4 | 9.7 | 50.0 | 230.0 | 1 | 1 | 1 |
| Chinchilla | 0.425 | 6.40 | 11.0 | 1.5 | 12.5 | 7.0 | 112.0 | 5 | 4 | 4 |
| Cow | 465.000 | 423.00 | 3.2 | 0.7 | 3.9 | 30.0 | 281.0 | 5 | 5 | 5 |
| Desert.hedgehog | 0.550 | 2.40 | 7.6 | 2.7 | 10.3 | NA | NA | 2 | 1 | 2 |
| Donkey | 187.100 | 419.00 | NA | NA | 3.1 | 40.0 | 365.0 | 5 | 5 | 5 |
| Eastern.American.mole | 0.075 | 1.20 | 6.3 | 2.1 | 8.4 | 3.5 | 42.0 | 1 | 1 | 1 |
| Echidna | 3.000 | 25.00 | 8.6 | 0.0 | 8.6 | 50.0 | 28.0 | 2 | 2 | 2 |
| European.hedgehog | 0.785 | 3.50 | 6.6 | 4.1 | 10.7 | 6.0 | 42.0 | 2 | 2 | 2 |
| Galago | 0.200 | 5.00 | 9.5 | 1.2 | 10.7 | 10.4 | 120.0 | 2 | 2 | 2 |
| Genet | 1.410 | 17.50 | 4.8 | 1.3 | 6.1 | 34.0 | NA | 1 | 2 | 1 |
| Giant.armadillo | 60.000 | 81.00 | 12.0 | 6.1 | 18.1 | 7.0 | NA | 1 | 1 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Giraffe | 529.000 | 680.00 | NA | 0.3 | NA | 28.0 | 400.0 | 5 | 5 | 5 |
| Goat | 27.660 | 115.00 | 3.3 | 0.5 | 3.8 | 20.0 | 148.0 | 5 | 5 | 5 |
| Golden.hamster | 0.120 | 1.00 | 11.0 | 3.4 | 14.4 | 3.9 | 16.0 | 3 | 1 | 2 |
| Gorilla | 207.000 | 406.00 | NA | NA | 12.0 | 39.3 | 252.0 | 1 | 4 | 1 |
| Gray.seal | 85.000 | 325.00 | 4.7 | 1.5 | 6.2 | 41.0 | 310.0 | 1 | 3 | 1 |
| Gray.wolf | 36.330 | 119.50 | NA | NA | 13.0 | 16.2 | 63.0 | 1 | 1 | 1 |
| Ground.squirrel | 0.101 | 4.00 | 10.4 | 3.4 | 13.8 | 9.0 | 28.0 | 5 | 1 | 3 |
| Guinea.pig | 1.040 | 5.50 | 7.4 | 0.8 | 8.2 | 7.6 | 68.0 | 5 | 3 | 4 |
| Horse | 521.000 | 655.00 | 2.1 | 0.8 | 2.9 | 46.0 | 336.0 | 5 | 5 | 5 |
| Jaguar | 100.000 | 157.00 | NA | NA | 10.8 | 22.4 | 100.0 | 1 | 1 | 1 |
| Kangaroo | 35.000 | 56.00 | NA | NA | NA | 16.3 | 33.0 | 3 | 5 | 4 |
| Lesser.short-tailed.shrew | 0.005 | 0.14 | 7.7 | 1.4 | 9.1 | 2.6 | 21.5 | 5 | 2 | 4 |
| Little.brown.bat | 0.010 | 0.25 | 17.9 | 2.0 | 19.9 | 24.0 | 50.0 | 1 | 1 | 1 |
| Man | 62.000 | 1320.00 | 6.1 | 1.9 | 8.0 | 100.0 | 267.0 | 1 | 1 | 1 |
| Mole.rat | 0.122 | 3.00 | 8.2 | 2.4 | 10.6 | NA | 30.0 | 2 | 1 | 1 |
| Mountain.beaver | 1.350 | 8.10 | 8.4 | 2.8 | 11.2 | NA | 45.0 | 3 | 1 | 3 |
| Mouse | 0.023 | 0.40 | 11.9 | 1.3 | 13.2 | 3.2 | 19.0 | 4 | 1 | 3 |
| Musk.shrew | 0.048 | 0.33 | 10.8 | 2.0 | 12.8 | 2.0 | 30.0 | 4 | 1 | 3 |
| N.American.opossum | 1.700 | 6.30 | 13.8 | 5.6 | 19.4 | 5.0 | 12.0 | 2 | 1 | 1 |
| Nine-banded.armadillo | 3.500 | 10.80 | 14.3 | 3.1 | 17.4 | 6.5 | 120.0 | 2 | 1 | 1 |
| Okapi | 250.000 | 490.00 | NA | 1.0 | NA | 23.6 | 440.0 | 5 | 5 | 5 |
| Owl.monkey | 0.480 | 15.50 | 15.2 | 1.8 | 17.0 | 12.0 | 140.0 | 2 | 2 | 2 |
| Patas.monkey | 10.000 | 115.00 | 10.0 | 0.9 | 10.9 | 20.2 | 170.0 | 4 | 4 | 4 |
| Phanlanger | 1.620 | 11.40 | 11.9 | 1.8 | 13.7 | 13.0 | 17.0 | 2 | 1 | 2 |
| Pig | 192.000 | 180.00 | 6.5 | 1.9 | 8.4 | 27.0 | 115.0 | 4 | 4 | 4 |
| Rabbit | 2.500 | 12.10 | 7.5 | 0.9 | 8.4 | 18.0 | 31.0 | 5 | 5 | 5 |
| Raccoon | 4.288 | 39.20 | NA | NA | 12.5 | 13.7 | 63.0 | 2 | 2 | 2 |
| Rat | 0.280 | 1.90 | 10.6 | 2.6 | 13.2 | 4.7 | 21.0 | 3 | 1 | 3 |
| Red.fox | 4.235 | 50.40 | 7.4 | 2.4 | 9.8 | 9.8 | 52.0 | 1 | 1 | 1 |
| Rhesus.monkey | 6.800 | 179.00 | 8.4 | 1.2 | 9.6 | 29.0 | 164.0 | 2 | 3 | 2 |
| Rock.hyrax.Hetero | 0.750 | 12.30 | 5.7 | 0.9 | 6.6 | 7.0 | 225.0 | 2 | 2 | 2 |
| Rock.hyrax.Procavia | 3.600 | 21.00 | 4.9 | 0.5 | 5.4 | 6.0 | 225.0 | 3 | 2 | 3 |
| Roe.deer | 14.830 | 98.20 | NA | NA | 2.6 | 17.0 | 150.0 | 5 | 5 | 5 |
| Sheep | 55.500 | 175.00 | 3.2 | 0.6 | 3.8 | 20.0 | 151.0 | 5 | 5 | 5 |
| Slow.loris | 1.400 | 12.50 | NA | NA | 11.0 | 12.7 | 90.0 | 2 | 2 | 2 |
| Star.nosed.mole | 0.060 | 1.00 | 8.1 | 2.2 | 10.3 | 3.5 | NA | 3 | 1 | 2 |
| Tenrec | 0.900 | 2.60 | 11.0 | 2.3 | 13.3 | 4.5 | 60.0 | 2 | 1 | 2 |
| Tree.hyrax | 2.000 | 12.30 | 4.9 | 0.5 | 5.4 | 7.5 | 200.0 | 3 | 1 | 3 |
| Tree.shrew | 0.104 | 2.50 | 13.2 | 2.6 | 15.8 | 2.3 | 46.0 | 3 | 2 | 2 |
| Vervet | 4.190 | 58.00 | 9.7 | 0.6 | 10.3 | 24.0 | 210.0 | 4 | 3 | 4 |
| Water.opossum | 3.500 | 3.90 | 12.8 | 6.6 | 19.4 | 3.0 | 14.0 | 2 | 1 | 1 |
| Yellow-bellied.marmot | 4.050 | 17.00 | NA | NA | NA | 13.0 | 38.0 | 3 | 1 | 1 |