STK4900/9900: Statistical methods and applications
Compulsory assignment spring 2007

This is the compulsory assignment in STK4900/9900 the spring semester 2007. The written solution to the assignment must be handed in no later than

1 pm Friday March 30th

either by regular mail or by e-mail to Ørnulf Borgan, Matematisk institutt, Universitetet i Oslo, P.B. 1053 Blindern, 0316 Oslo (e-mail: `borgan@math.uio.no`).

You may use the software package of your choice. But whether you use R or not, you must be able to answer all questions. (We recommend that you use R.)

You are allowed to collaborate and discuss the problems with other students, but each student has to formulate her or his own answers. You should give the names of the students you collaborate with, so that it is possible to compare the written solutions.

The written solution may be divided into two parts. In the main part you answer the questions and present the numerical results and plots that are necessary for your arguments. (It is not sufficient to present the numerical results and the plots. You should also discuss what you can learn from them!) In an appendix you may include computer printouts and other "technical" material that do not fit nicely into the main part.

Remember to write your full name, postal address and e-mail address on the written solution!

Weight of bears

On the course web-page you find the data set `bears.dat` which contains measurements on 54 wild bears. For each bear the measurements are:

`AGE`:	Age in months
`MONTH`:	The month in which the measurements were taken (1 = January, 2 = February, etc.)
`SEX`:	1 = male, 2 = female
`HEADLEN`:	head length measured in inches
`HEADWTH`:	width of the head measured in inches
`NECK`:	distance around neck measured in inches
`LENGTH`:	length of body measured in inches
`CHEST`:	distance around chest measured in inches
`WEIGHT`:	weight measured in pounds

When obtaining measurements from anesthetized wild bears, it is easy to find values such as body length and chest size. The weight of a bear is more difficult to obtain, since then the bear must be lifted. The question is therefore whether one may predict the weight of a bear well enough from the other measurements.

We start out in questions a-f by only considering the response `WEIGHT` and the covariates `LENGTH` and `CHEST`.

a) Report the main features of the variables `WEIGHT`, `LENGTH` and `CHEST` by numerical summaries and plots.

b) Fit a model of the form:

$$\texttt{WEIGHT} = \beta_0 + \beta_1 \,\texttt{LENGTH} + \beta_2 \,\texttt{CHEST} + \varepsilon$$

Also fit a model of the form:

$$\log(\texttt{WEIGHT}) = \beta_0 + \beta_1 \log(\texttt{LENGTH}) + \beta_2 \log(\texttt{CHEST}) + \varepsilon$$

Which of the two models seems to give the best fit? Can the well-known formula "Volume = area × height" help to explain this?

c) Use plots of the residuals to examine the fit of the second model in question b. Comment on what the plots tell you.

d) Observation 52 seems to be an outlier. Fit the model

$$\log(\texttt{WEIGHT}) = \beta_0 + \beta_1 \log(\texttt{LENGTH}) + \beta_2 \log(\texttt{CHEST}) + \varepsilon$$

without this observation. Compare with the results in question b.

e) A more sensible strategy than leaving out one observation, may be to renounce on getting a prediction rule for all bears. Instead one may concentrate on adult bears. Fit the second model in question b, but now only for bears older than twelve months. Comment on how well the model fits and give an interpretation of the fitted model.

f) What weight would you predict for a grown-up bear being 65 inches tall and measuring 40 inches around the chest?

We then consider models for adult bears using all covariates.

g) Fit a model including all covariates for the bears older than twelve months. Log-transform all variables, except `MONTH` and `SEX` that should be treated as factors. What is the square of the multiple correlation coefficient, R^2, for this model? How does it compare to R^2 for the model in question e? Comment!

h) Find the cross-validated R^2 for the model in question g. Discuss why the cross-validated R^2 is better suited for model selection than the ordinary R^2.

i) In order to find a "best possible" model for predicting the weight of an adult bear from the other covariates, we fit a sequence of regression models and compute the cross-validated R^2 for each of the models. We start with the model with no covariates and add successively the covariates in the order: log(`CHEST`), log(`LENGTH`), log(`HEADWTH`), log(`AGE`), `MONTH`, log(`HEADLEN`), and `SEX` (corresponding to forward selection of the covariates according to their significance). Compute the cross-validated R^2 for these models. Which model gets the smallest cross-validated R^2?

j) Summarize what you have found, and conclude what may be a reasonable rule for predicting the weight of an adult bear.