

## STK4900/9900: Statistical methods and applications

### Compulsory assignment spring 2014

This is the compulsory assignment in STK4900/9900 for the spring semester 2014. The written solution to the assignment must be handed in no later than

*Thursday 13th of March 2014, at 2:30 pm.*

in Niels Henrik Abels hus, 7th floor (by the reception). If you do not live nearby the University of Oslo you may send the solution by regular mail to: Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, 0316 Oslo. The solution must be stamped Thursday 13, March. Please mark the envelope with STK4900/9900.

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 measurement was 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 co-variates **LENGTH** and **CHEST**.

- Report the main features of the variables **WEIGHT**, **LENGTH** and **CHEST** by numerical summaries and plots.
- Fit a model of the form:

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

Also fit a model of the form:

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

Which of the two models seems to give the best fit? Can the well-known formula “Volume = area  $\times$  height” help to explain this?

- Use plots of the residuals to examine the fit of the second model in question b. Comment on what the plots tell you.
- The residual for observation 52 is quite large, so this observation seems to be an “outlier”. Fit the model

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

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

In the remaining parts of the assignment, we will concentrate on modeling the weight of adult bears, i.e. bears that are older than 12 months.

- e) Fit the second model in question b, but now only for bears older than twelve months. Check 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. To this end we fit the covariates in the following order:  $\log(\text{CHEST})$ ,  $\log(\text{LENGTH})$ ,  $\log(\text{HEADWTH})$ ,  $\log(\text{AGE})$ , **MONTH**,  $\log(\text{HEADLEN})$ ,  $\log(\text{NECK})$ , 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 largest 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.