UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Examination in:	STK4030/9030 — Modern data analysis.
Day of examination:	Tuesday, December 5, 2006.
Examination hours:	15.30 – 18.30.
This examination set consists of 2 pages.	
Appendices:	None.
Permitted aids:	Approved calculator.

Make sure that your copy of the examination set is complete before you start solving the problems.

Problem 1.

In a linear regression model we have

$$E(y|X) = X\beta\,,$$

where X is an $\mathrm{N} \times (p+1)$ data matrix.

a) Find least squares estimator $\hat{\beta}$ of β.

b) Show that $\hat{y} = X\hat{\beta}$ can be written as $\mathrm{H}y$, where H is a projection matrix, that is, a matrix satisfying: $\mathrm{H}^T = \mathrm{H}$ and $\mathrm{H}^2 = \mathrm{H}$.

c) Specialize to the case $p = 1$, where the first column of X consists of 1's. Show that this gives the usual simple regression model, and show also that the estimates become the usual for this situation.

Problem 2.

a) What is meant by a cubic spline?

b) In a cubic spline with K knots, how many basis functions are needed? Give reasons for your answer.

(Continued on page 2.)

Problem 3.

Describe K-fold cross validation. In particular, tell how it is used to estimate prediction error.

Problem 4.

A treebased method aims at estimating a function

$$f(x) = \sum_{m=1}^{M} c_M \; I(x \in R_m).$$

a) Tell roughly how the regions R_m are determined.

b) How are the constants c_M estimated? Show that when two regions are collapsed together, the new estimate $\hat{c}_m$ will be between the two old ones.

END