Exam

Linköpings Universitet, Institutionen för datavetenskap, Statistik

---

Course:	732A33 Advanced Programming in R
Date and time:	2013-05-17, 8-12.
Teacher:	Patrik Waldmann
Material:	The Art of R Programming (Matloff). The book should be free from notes and comments, but can contain underlined lines and indicators for quickly finding chapters/sections etc. Slides from all lectures, the lab reports as well as the three papers on data mining in R can be included in a personal folder.
Grades:	Maximum is 20 points. A=19-20 points. B=17-18 points. C=12-16 points. D=10-11 points. E=8-9 points. F=0-7 points.

Solutions should be written in a Word file, do not include the questions.

The R code should be complete and readable code, possible to run by copying directly into a script.

Comment directly in the code whenever something needs to be explained or discussed.

Follow the instructions carefully.

1 Data and programming structure (4p)

a). Below follows explanations but the code is missing. Your task is to fill in code that performs according to explanations.

Creates a sample matrix named mymat with 50 columns and 100 rows. The entries are filled with values from N(0,1).

Computes the mean and standard deviation of each row in mymat by applying a for loop. Stores the results in a list called myresfor with components mean and sd.

Computes the mean and standard deviation of each row in mymat using the apply function. Stores the results in a list called myresappl with components mean and sd.

Confirms that the values from both calculations are identical for the mean and standard deviation, respectively.

(2p)

b). Write a function that that reads two matrices of the same size (n x n). The function should then perform element-wise multiplication of the rows of matrix A with the corresponding columns of matrix B. For example, the first entry of row 1 of matrix A should be multiplied with the first entry of column 1 of matrix B, the second entry of row 1 of matrix A should be multiplied with the second entry of column 1 of matrix B, and so on. The output should be a matrix of size n x n. Only present the code. (2p)

2 Simulation, input/output and graphics (6p)

a). Write a function that reads an integer value (`myvalue`) from the keyboard, evaluates if the number is evenly divisible by either 3 or 5 (i.e. the result is an integer), and finally prints the results to the screen saying "`myvalue` is evenly divisible by 3", or "`myvalue` is evenly divisible by 5", or "`myvalue` is not evenly divisible by 3 or 5". `myvalue` should be replaced with its number in the output. (3p)

b). Carefully explain the code below:

```
> Sigma<-cov(matrix(rnorm(30),nrow=10))
> A<-t(chol(Sigma))
> x<-A%*%rnorm(3)
```

Use the code to simulate 500 x-vectors based on the same Sigma. Merge the vectors into an 3x500 matrix and name the three rows X,Y and Z and plot in a 3D scatterplot. Present explanation, code and plot. (3p)

3 Strings (3p)

Read the text in file `Instructions.txt` and produce a table that consists of the words in the text in alphabetical order in the first column, and the number of times the word is in the text in the second column. The . and , should be excluded, and uppercase and lowercase letters should be treated as the same. Save the table to a file called `Wordsummary.txt`. (3p)

4 Interfacing and parallel computations (4p)

a). Write code that reads matrix x (of dimension n x p) and performs p one-sided t-tests (i.e. tests if the column means are different from 0) in parallel. The code should be so flexible that it detects the number of cores automatically, and the output should be a vector with p-values. (2p)

b). Compare the computation time of your code in a) on a matrix of size n=1000 and p=1000 with the timing from the ordinary apply function (performing t-tests). The timing should be over the complete parallel code. Provide code and timing output. (2p)

5 Data mining (3 p)

a). Your task is to use the `glmnet` package to perform lasso analysis based on the data in `spambase.csv` data file. The variable spam in the last column should be used as response variable. All other variables should be used as predictors, you need to convert these to a matrix. Perform lasso analysis (using the binomial family) together with 10-fold cross-validation to determine which variables that will be selected. Provide the plots from the CV-analyses that shows how MSE is related to the regularization parameter (log(λ)), and a list of the selected regression coefficients. Explain your code carefully. (3p)