# Computer lab ML 1-2: Programming in R and regression.

## Learning objectives

The main objective of this computer lab is to make the student familiar with basic R handling as well as performing regression analysis.

After completing the lab the student shall be able to:

1. Start and run a basic R session
2. Have an understanding of important programing structures, including loops and application of functions.
3. Perform regression and prediction of data split into training and test partitions.

## Recommended reading

Chapter 1 - 8.5 in Matloff (2011). The art of R programming.

Chapter 1 - 4 in James et al. (2013). An Introduction to Statistical Learning.

## Assignment 1: Introductory R session

The book “The art of R programming” by Norman Matloff (<http://diytranscriptomics.com/Reading/files/The%20Art%20of%20R%20Programming.pdf>) provides text for a first R session that starts at page 4. In this assignment you will perform an interactive R session based on the commands given in the text (between lines at >) for section 1.2 to 1.4 and 1.7.1 to 1.7.4, and save this session. You shall also provide comments (using the hash-mark #) describing what each command does. For example: a <- c(3,4) # Combines values 3 and 4 into a vector named a. Don’t forget to set your working directory after start of R.

***Assignment 2: Lists and mode transformations***

Lists are collections of vectors that can be of different mode. Your task in this assignment is to first create a list (named mylist) that consists of one character vector (“J”,“A”,“G”,”B”,”H”,”E”,”C”,”F”,”D”,”I”), one numeric vector (10 random values from rnorm), and a matrix of size 10 x 10 (containing integer 1 to 100). After that you will provide a way of sorting the rows of all components (character, numeric and matrix) of mylist based on the order of the sorted character list. Finally, do a matrix times vector multiplication of the sorted second component and the third (integer) component (you will need to extract and convert these components to suitable modes).

***Assignment 3: Regression***

The mushroom data set <http://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.data> includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. Hence, the classes of the binary variable is poisonous (p) and edible (e), and located in the first column. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom. Your task is to find out if this statement is true.

Read the data directly into R from the webpage. Perform logistic regression of all variables against the edibility response using the GLM function. Take a look at the result and delete the unimportant variables in a new data set. Create training and test data sets be randomly splitting the new data set into 70% training and 30% test. Hint: use the sample.int() function with replace=F. Perform a new logistic regression on the training data and predict edibility for the test data. Calculate the Prediction Error Rate between the test predictions and original test observations. What’s your conclusion? Present your code with clear explanations (in the form of # comments) as well as a discussion of the result and your main conclusion.

## To hand in

A written report (preferably a Word or .pdf document) where you summarize your main findings in the assignments. Submit your report to Patrik.Waldmann@slu.se.