# Computer lab ML 3-4: Regularization and non-linear modelling.

## Learning objectives

The main objective of this computer lab is to make the student familiar with variable selection and regularization, as well as some important non-linear regression techniques.

After completing the lab the student shall be able to:

1. Perform shrinkage analyses with ridge regression and the lasso.
2. Apply different spline procedures for non-linear regression and evaluate their usefulness.

## Recommended reading

Chapter 5 - 7 in James et al. (2013). An Introduction to Statistical Learning.

## Assignment 1: Regularization with ridge regression and the lasso for genome-wide prediction

Genome-wide prediction is a technique in animal breeding where genetic SNP markers (predictors) are scored on a set of individuals in order to predict phenotypes (response). The number of SNPs are often much larger than the number of individuals, resulting in the *p* >> *n* problem. Regularization is a useful procedure in this situation.

The glmnet package in R fits a wide range of regularized regression models based on the coordinate descent methodology. glmnet computes the solutions for a decreasing sequence of values for the regularization parameter , starting at the smallest value for which the entire vector (i.e. where all regression coefficients are regularized so heavily that none are in the model). In glmnet, the lasso is obtained by setting the parameter α to 1. Ridge regression is obtained by setting α to 0.

Your task is to compare the prediction MSE between ridge regression and the lasso on a simulated data in the **QTLMAS2010SNP.Rdata** and **QTLMAS2010Phen.Rdata** file. Load the data files into R and create training data of observations 1:2326 and test data of observation 2327:3226. Perform ridge regression and lasso (use 10 –fold CV), and calculate predicted MSE based on the test data for both methods. Discuss your results and present a plot of the regression coefficients of the method with lowest MSE.

***Assignment 2: Analysis of mortality rates using splines***

Gompertz (1825) theory that mortality rates (probability of dying per unit time) of many organisms increase at an exponential rate was examined in an experiment involving fruit flies. A total of 1,203,646 fruit flies comprised the population for this experiment and the number of flies found dead each day was recorded. The data set **mortality\_rate.csv** contains the mortality rate (*Rate*) of the flies for each day (*Day*).

**2.a.** Compute the *Log-Mortality-Rate* (*LMR*) as the natural logarithm of *Rate*. Make a scatter plot of LMR versus Day. Given the log-transformation, does it seem as if the data would follow an exponential mortality rate?

**2.b.** Fit natural cubic splines using the ns() function in package splines() (LMR as response variable and Day as predictor variable). Fit four models with 1, 2, 15 and 50 knots. Produce four plots with data points, predicted lines and their knots. Use all the data points to calculate the mean squared error (MSE). Which number of knots seems to give the most reasonable fit based on MSE?

**2.c.** The next task is to divide the data into 70% training and 30% test partitions. Use the sample.int() function to partition the original data into 95 train and 41 test data samples. Fit the models to the train data and perform the predictions on the test data. Calculate MSE based on the difference between the real test data and the predicted test data. Repeat 10 times (i.e. create 10 different training and test data) and calculate the mean MSE. Compare the results with these obtained in **2.b**.

**2.d.** The final task of this assignment is to fit smoothing splines using the smooth.spline() function in R. Use generalized cross-validation to find the optimal degree of smoothing on each of the training data. Provide a plot of the data points and fitted spline curve for the first training data. Also, present the average of the estimated effective degrees of freedom and the smoothing parameter λ, as well as the number of proper knots. Compare the average predicted MSE with the best MSE from **2.c**.

## 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.