# Statistical Learning, MT7049

# Project 1

**Instructions:** This project consists of two tasks that should be solved individually. Any code inspired by fellow students or other sources must be clearly acknowledged.

The solution should be submitted at the course webpage as a pdf file (where the obtained results are presented and described) together with a source file.

## Data

For this project, real financial data from the Swedish capital market including stock prices and capital market index will be used. A link is provided below to download data for the Swedish capital market index OMX and **25 stocks** of your interest. The data series should contain at least 5-years of daily values (for instance, for 5-year data, the period of study is 2011-2015). This link at OMX30 data from WSJ will guide you to all the stocks registered at Swedish stock exchange. Choose **25 stocks** of your interest. Once you click any particular stock from `Movers OMXS30`, a new window will open and the `Historical Prices` option under `RESEARCH & RATINGS` option on the bottom menu will guide you to download historical prices. Choose a date period of your interest and click on `Download` button. There is an option to download the data as a spreadsheet at the bottom of this page, from here you can save data into your directory. In these data, you will find many different type of `prices`. It is suggested to work with the data in column `Close`, since it gives us the closing value of that stock prices at a particular day. The data for the index of the Swedish capital market `OMX Stockholm 30 Index` can be downloaded from the link at OMXS30. Here again, you will find many different types of `prices`, but as above, choose to work with the data in column `Close`. Before proceeding any further, make sure, the frequency and length of each stock data and the market stock is same. Finally, for each stock as well as for the index of the Swedish capital market compute the log-returns defined by

$$r_t = \ln(P_t) - \ln(P_{t-1}),$$

where $P_t$ denote the price/index value on day $t$ and $r_t$ is the log-return on day $t$.

An example of getting data and computation of the log-returns in the case of the OMX index, ABB and Nordea can be found in the provided R-code on the course web-page.

## Task 1: Linear regression model

The return of the market index is determined as a linear combination of the returns on stocks included in the index. One of economical hypotheses states that there exist only a few stocks on a capital market such that a portfolio constructed by using these stocks will duplicate the behaviour of the capital market index. One of the aims of this task is to determine the portfolio based on the stocks included in the OMX index that can mimic its behavior. For that reason, several linear regressions studied in the course will be fitted and the corresponding stocks will be specified. In order to achieve the goal, perform the following steps using the return data obtained as described in section "Data" :

(a) Fit a linear regression model by using the log-returns of the capital market index as a response variable and the log-returns of 25 stocks as predictors. Determine which stocks have an influence on the return of the capital market. Explain your answer.

(b) Can the results of part (a) be used to answer the question which of the stocks have to be included in the portfolio to mimic the behavior of the Swedish capital market index? Provide a statistical explanation of your answer.

(c) Use the forward-stepwise selection to the regression of part (a). Which stocks should be included in the portfolio that mimics the behavior of the Swedish capital market index?

(d) Repeat part (c) with the backward-stepwise selection. Do you receive the same set of stocks as in part (c)?

(e) Fit the ridge regression to the data. Determine the regularization parameter $\lambda$ by using leave-one-out cross-validation. Also, include the plots with the estimated coefficients as a function of $\lambda$. Is it possible to determine the stocks and the structure of the portfolio that mimics the behavior of the Swedish capital market index by using the ridge regression? Explain your answer.

(f) Repeat the previous step by using the Lasso regression instead of the ridge regression. Is it possible to determine the stocks and the structure of the portfolio that mimics the behavior of the Swedish capital market index by using the Lasso regression? Explain your answer.

(g) Summarize the obtained results by providing the stocks to be included in the portfolio that mimics the behavior of the Swedish capital market index together with the corresponding regression coefficients.

## Task 2: Linear classification

One of the most important tasks of traders on capital markets is to determine proper times to open the position (to buy assets) and to close the position (to sell assets) for each stock. It is important to open the position when the price of a stock starts to go up and to close the position when the price of a stock drops down. In this task using the logistic regression model, we aim to check whether it is advantageous to use a part of stocks determined by the Lasso regression in comparison to all 25 stocks in order to describe the movement of the Swedish capital market index. For that reason, we use the first (around) 80% to fit two logistic regression models, while the other 20% of data should be used for the prediction purposes, where dependent variable in both logistic regressions is defined by

$$G_t = \begin{cases} 1, & \text{market return} \geq 0 \text{ on day } t, \text{ i.e., the market index goes up on day } t, \\ 0, & \text{market return} < 0 \text{ on day } t, \text{ i.e., the market index goes down on day } t. \end{cases}$$

(a) Fit two logistic regression models to the first (around) 80% of returns data created in section "Data", where the first one uses all 25 stock returns as predictors while the second one uses only those stocks which are determined by the Lasso regression in Task 1. Explain the output of the fitted models.

(b) Using the remaining 20% of return data, determine the error rate by using the two logistic regressions from part (a), i.e., compute the proportion of cases with incorrect classifications to the total number of considered cases for each of the model.

(c) Which conclusions can be drawn by using your results from part (b)? Explain your answer.