Homework 1 (part 1)

This is the first half. The second half will appear at the end of the next laboratory (Regression 03 - Regularization). The aim here is to apply linear regression on the Home prices dataset from Kaggle House Prices: Advanced Regression Techniques. As explained in the documentation there, this dataset is a test bed for many prediction methods, thus it is not to be expected that linear regression will be particularly successful. Later on in the course we will try other methods.
As a warm-up, take the similar but smaller Boston home prices dataset. This is a very classic example and it is easy to find its study with linear regression in many online resources. Googling "Boston home prices" will yield many implementations, both in R and in python.
You can learn from and even reproduce parts of these sources. This is acceptable provided that you:

1. Give credit where it is due, in particular, including full reference (URL) of any cited work

2. Do not copy/paste "*in extenso*" large chunks of code.

3. Understand everything you write, explaining with sufficient details the steps you take and the results you obtain.

The Boston home prices data are intended for prediction of home values in suburbs of the namesake city. `medv` is the response variable and the other variables are predictors. See their description in the dataset help. Some of these predictors are intrinsic habitational characteristics, such as number of bedrooms; other predictors have a socioeconomic nature and, finally, other predictors are geographical or environmental.
In R you find this dataset in the MASS package, from the book by W. N. Venables and B. D. Ripley (2002), *Modern Applied Statistics with S:*

```
data(Boston)
str(Boston)
```

Guidelines

Perform a statistical description of the data. In the first place individually, summarizing each variable, both graphically, e.g. with boxplot and histogram, and numerically. Do these variables have a normal appearance? Or, rather, do these variables show an asymmetric shape? Check correlations between pairs of

predictors and between individual predictors and response. It will be useful to truncate to 2 or 1 decimal places, to avoid clutter:
In this way we can see at a glance which correlations are large or small. Is there a danger of multicollinearity?
Fit a linear regression model of `medv` on the remaining variables. From the model summary, can we state the model fits well? which variables appear to be better or worse predictors?
Prepare an optimal model with the better predictors. The `ResSS` of this model is much larger than the one from the full model? Note that this model still has a non significant predictor. Discard it.
Fit another linear regression model with response `log(medv)` on the remaining predictors or with their logarithms. Which one is better?

Homework 1 (part 2)

Adjust the regression y~x1+x2+x3+x4+x5+x6 with the Fearn dataset using:

1. Ordinary Least Squares (OLS), selecting the best predictors subset
2. Ridge regression. Compare prediction errors. Which one is better?
3. The lasso.
4. PCR
5. PLS

NOTE: the data frames Fearn.1 and Fearn.2 were used as train and test subsets in the original paper. You may choose to follow this selection or merge both subsets and partition the joint dataset in some other way.