Computational Statistics II
Assignment 1: Bootstrap and cross-validation for linear regression

Alessia Pini

The R dataset `BostonHousing` (accessible through the library `mlbench`) contains housing data for 506 census tracts of Boston from the 1970 census. You can access the data as follows:

```
library(mlbench)
data(BostonHousing)
head(BostonHousing)
help(BostonHousing)
```

We are interested in exploring the relationship between a set of variables and the house median value (medv) as outcome, using the Bootstrap for inference on the regression model.

1. Give a theoretical description of the Bootstrap for regression, specifying the model and assumptions. Describe the two methods of bootstapping residuals and pairs, underlying their differences (max. 1 page).

2. Fit a linear model on the entire dataset, using the variable `medv` as outcome. Give an estimate of the standard error or each estimated coefficient using Bootstrap. Use both the bootstrapping of pairs and of residuals.

3. Compute the classical t-distribution confidence intervals, and the Bootstrap-t intervals for all model coefficients. Comment on the differences between the two bootstrapping methods (residuals and pairs), if any. Choose one of the two bootsrtapping method and reduce the model (if possible) keeping significant covariates only.

4. Perform a 10-fold cross-validation to estimate the test set MSE of the full model (all covariates) and reduced model (obtained at point 3). Compare the results.

5. **Bonus.** Perform a test on the effect of the variable `rm` based on Bootstrapping the residuals of the null model. Compute the p-value of the test, and compare it with the classical t-test p-value. Comment on the result.