

LAB 7 – Backward Stepwise Selection

In this lab, we will perform backward stepwise selection for linear regression. We will use a modified version of our usual dataset, please download it on Blackboard.

As usual, our features consist of the columns “Age”, “Height”, “Mental” and “Skill”, while the target variable is “Salary”.

The entire algorithm is as follows:

Algorithm 6.3 *Backward stepwise selection*

1. Let \mathcal{M}_p denote the *full* model, which contains all p predictors.
 2. For $k = p, p - 1, \dots, 1$:
 - (a) Consider all k models that contain all but one of the predictors in \mathcal{M}_k , for a total of $k - 1$ predictors.
 - (b) Choose the *best* among these k models, and call it \mathcal{M}_{k-1} . Here *best* is defined as having smallest RSS or highest R^2 .
 3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .
-

Implement this algorithm to the best of your ability.

Start by computing the cross-validation MSE for M_4 . You can then enter a loop for the rest.

We will use linear regression for every model M_k . For step (2a), use R^2 instead of RSS . For step (3), use cross-validation (10-fold) and store every cross-validation MSE for later use. You can use the cross-validation function from LAB 4.

Finally, print the following:

1. For each model M_0, M_1, M_2, M_3 and M_4 , the cross-validation MSE values,

2. Which model among M_0, M_1, M_2, M_3 and M_4 is the most optimal,
3. Features that are included in the most optimal model,
4. The order of the features in terms of importance, from high to low.

Important note: Same rules from LABS 1-5 apply.