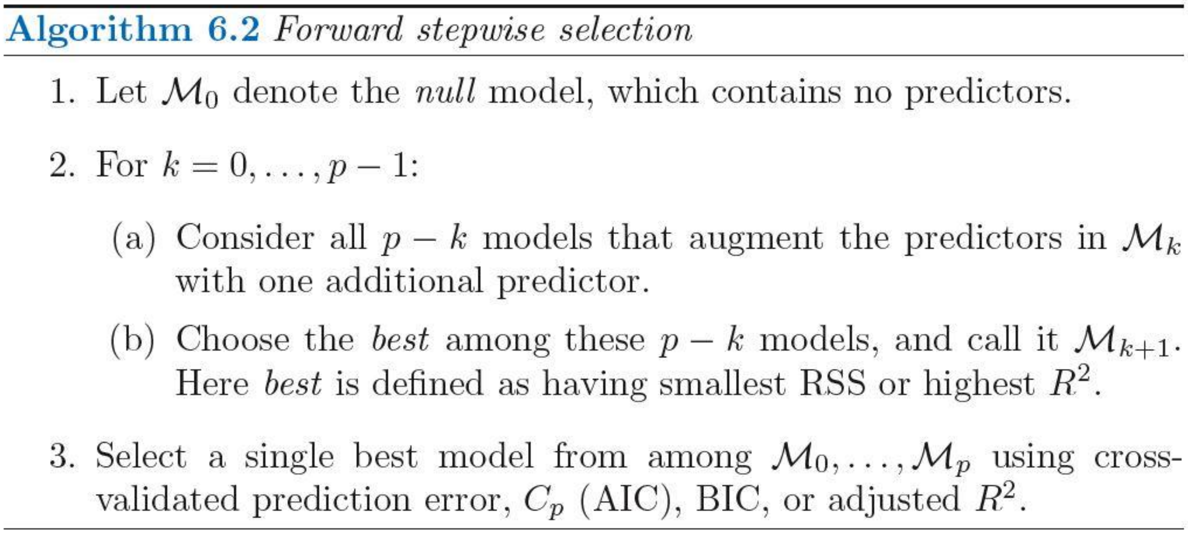
LAB 5 – Forward Stepwise Selection

In this lab, we will perform forward stepwise selection with multiple linear regression. To do this, we will need the .csv provided in the LAB 5 folder.

The entire algorithm is as follows:



Algorithm 6.2 *Forward stepwise selection*

```
1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
    (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$
        with one additional predictor.
    (b) Choose the *best* among these $p - k$ 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 $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

We are again interested in the effects of age, experience and power (x_1, x_2, x_3) to the player's salary (y). Since we have three variables plus the intercept, we should also calculate R^2 scores of according models M_0, M_1, M_2, M_3.

However, we are only going to calculate the first two of these, M_0 and M_1.

(20 pts) Define a function for calculating the adjusted R^2 scores This should take three parameters: the original output (y), predictions for this output ($\hat{y}$), and the number of variables in your model (d). In the body of the function, calculate and return the following:

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

n is equal to the number of elements in y (or $\hat{y}$, since they have the same size).

Also implement another function for calculating regular R^2 scores, if you haven't done so in the previous labs.

(30 pts) Calculate M_0:

Construct your input matrix as before. Then, for M_0, since your input should only be the intercept, take the ones column (the first column of your original input matrix) as input. Multiply it with the mean of your y (which is the list of salaries in this case). This acts as an array of predictions for M_0. Then calculate and display the adjusted R^2 score for M_0 using the function you implemented in the previous step. Notice that you're using no variables for M_0, which means that the parameter d of the function for adjusted R^2 score should be zero.

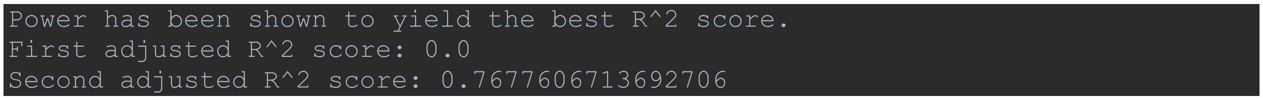
(50 pts) Calculate M_1:

For M_1, your input should be the intercept plus 1 variable. In order to choose the correct variable (you have three options in x_1, x_2, x_3), You need to calculate regular R^2 scores for each one. Implement a loop, where you:

- Choose one of x_1, x_2, x_3. Form a matrix with the chosen variable and x_0, which is the ones column.
- Perform multiple linear regression using this input as both train and test.
- Calculate the R^2 score using the predictions of this regression and store the result.

After the loop finishes, compare the R^2 scores. Display on the console which variable (age, experience or power) yielded the best R^2 score. Then, using the input which yielded the best R^2 score, perform another multiple linear regression using the input as both train and test. Using the predictions of this regression, calculate and display the adjusted R^2 score for M_1. Since you're using one variable besides the intercept, the parameter d of the function for adjusted R^2 score should be one.

Here is the desired (and correct) output:



```
Power has been shown to yield the best R^2 score.
First adjusted R^2 score: 0.0
Second adjusted R^2 score: 0.7677606713692706
```

Continuing with the insight section:

Model selection approaches help us select variables from a large set. Some of these variables can be of little help, some can even have adverse effects on our test error, therefore, these approaches come in very handy. They also help reduce the amount of data, which in turn reduces computational requirements.

We implemented a part of forward selection. Notice that for M_0, our adjusted R^2 score (or regular R^2 score) should always be zero. We are using the mean of the data as our predictions. Therefore, RSS should always be equal to TSS. And since our number of variables d is zero, the equation above for adjusted R^2 score equals to $1 - 1 = 0$.

The outline of the algorithm explains the rest. You have to extend the work we did to all variables. You can try expanding our lab work to the entire forward stepwise selection algorithm. If you do, you can contact me with any problems.

Keep in mind that in that case, you will need to keep track of which variable you have selected for M_i before moving on to M_{i+1}! You might have to keep multiple copies of the input matrix and perform different matrix operations on them.