What is Stepwise Selection? (Explanation & Examples)

Authored by

stats writer

April 22, 2024

RECOMMENDED CITATION

stats writer (2024). *What is Stepwise Selection? (Explanation & Examples)*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=137870

Stepwise selection is a statistical method used in data analysis to select the most significant variables for a predictive model. It involves systematically adding or removing variables from a model based on their individual contribution to the model's overall performance. This process continues until the most optimal combination of variables is achieved. Stepwise selection is commonly used in regression analysis and can help improve the accuracy and simplicity of a model. For example, in a study on the factors affecting student grades, stepwise selection can be used to identify the most influential variables such as study hours, attendance, and extracurricular activities. This method allows researchers to efficiently identify the key factors impacting the outcome of interest.

What is Stepwise Selection? (Explanation & Examples)

In the field of machine learning, our goal is to build a model that can effectively use a set of predictor variables to predict the value of some response variable.

Given a set of *p* total predictor variables, there are many models that we could potentially build. One method that we can use to pick the best model is known as best subset selection, which attempts to choose the best model from *all* possible models that could be built with the set of predictors.

Unfortunately this method suffers from two drawbacks:

It can be computationally intense. For a set of *p* predictor variables, there are 2p possible models. For example, with 10 predictor variables there are 210 =

1,000 possible models to consider.Because it considers such a large number of models, it could potentially find a model that performs well on training data but not on future data. This could result in **overfitting**.

An alternative to best subset selection is known as stepwise selection, which compares a much more restricted set of models.

There are two types of stepwise selection methods: forward stepwise selection and backward stepwise selection.

Forward Stepwise Selection

Forward stepwise selection works as follows:

1. Let M0 denote the null model, which contains no predictor variables.

2. For k = 0, 2, ... p-1:

Fit all p-k models that augment the predictors in Mk with one additional predictor variable.Pick the best among these p-k models and call it Mk+1. Define "best" as the model with the highest R2 or equivalently the

lowest RSS.

3. Select a single best model from among M0...Mp using cross-validation prediction error, Cp, BIC, AIC, or adjusted R2.

Backward Stepwise Selection

Backward stepwise selection works as follows:

1. Let Mp denote the full model, which contains all *p* predictor variables.

2. For k = p, p-1, ... 1:

Fit all k models that contain all but one of the predictors in Mk, for a total of k-1 predictor variables.Pick the best among these k models and call it Mk-1. Define "best" as the model with the highest R2 or equivalently the lowest RSS.

Criteria for Choosing the "Best" Model

The last step of both forward and backward stepwise selection involves choosing the model with the lowest prediction error, lowest Cp, lowest BIC, lowest AIC, or highest adjusted R2.

Here are the formulas used to calculate each of these metrics:

Cp: (RSS+2dσ?) / n

AIC: (RSS+2dσ?2) / (nσ?2)

BIC: (RSS+log(n)dσ?2) / n

Adjusted R2: 1 - ((RSS/(n-d-1)) / (TSS / (n-1)))

where:

d: The number of predictorsn: Total observationsσ?: Estimate of the variance of the error associate with each response measurement in a regression modelRSS: Residual sum of squares of the regression modelTSS: Total sum of squares of the regression model

Pros & Cons of Stepwise Selection

Stepwise selection offers the following benefit:

It is more computationally efficient than best subset selection. Given *p* predictor variables, best subset selection must fit 2p models.

Conversely, stepwise selection only has to fit 1+p(p+1)/2 models. For p = 10 predictor variables, best subset selection must fit 1,000 models while stepwise selection only has to fit 56 models.

However, stepwise selection has the following potential drawback:

It is not guaranteed to find the best possible model out of all 2p potential models.

For example, suppose we have a dataset with p = 3 predictors. The best possible one-predictor model may contain x1 and the best possible two-predictor model may instead contain x1 and x2.

In this case, forward stepwise selection will fail to select the best possible two-predictor model because M1 will contain x1, so M2 must also contain x1 along with some other variable.