What is Backward Selection and can you provide a definition and example?

Authored by

stats writer

June 28, 2024

Backward selection is a statistical method used for selecting relevant variables in a regression model by eliminating non-significant variables one by one. It is also known as stepwise regression. This method starts with a full model including all possible variables and then removes the least significant variable until the best fitting model is obtained. This process helps to simplify the model and improve its predictive power. For example, in a study examining factors affecting student GPA, backward selection may be used to determine which variables, such as study habits, extracurricular activities, or family income, have the most significant impact on GPA.

What is Backward Selection? (Definition & Example)

In statistics, stepwise selection is a procedure we can use to build a regression model from a set of predictor variables by entering and removing predictors in a stepwise manner into the model until there is no statistically valid reason to enter or remove any more.

The goal of stepwise selection is to build a regression model that includes all of the predictor variables that are statistically significantly related to the response variable.

One of the most commonly used stepwise selection methods is known as backward selection, which works as follows:

Step 1: Fit a regression model using all *p* predictor variables. Calculate the AIC* value for the model.

Step 2: Remove the predictor variable that leads to the largest reduction in AIC and also leads to a statistically significant reduction in AIC compared to the model with all *p* predictor variables.

Step 3: Remove the predictor variable that leads to the largest reduction in AIC and also leads to a statistically significant reduction in AIC compared to the model with *p-1* predictor variables.

Repeat the process until removing any predictor variable no longer longer leads to a statistically significant reduction in AIC.

***There are several metrics you could use to calculate the quality of fit of a regression model including cross-validation prediction error, Cp, BIC, AIC, or adjusted R2. In the example below we choose to use AIC.**

The following example shows how to perform backward selection in R.

Example: Backward Selection in R

For this example we'll use the built-in in R:

#view first six rows of *mtcars*
head(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb
Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4
Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4
Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1
Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1
Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2
Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

We will fit a multiple linear regression model using *mpg* (miles per gallon) as our response variable and all of the other 10 variables in the dataset as potential predictors variables.

The following code shows how to perform backward stepwise selection:

#define intercept-only model
intercept_only <- lm(mpg ~ 1, data=mtcars)

#define model with all predictors
all <- lm(mpg ~ ., data=mtcars)

```
#perform backward stepwise regression
backward <- step(all, direction='backward',
scope=formula(all), trace=0)

#view results of backward stepwise regression
backward$anova

Step Df Deviance Resid. Df Resid. Dev AIC
1 NA NA 21 147.4944 70.89774
2 - cyl 1 0.07987121 22 147.5743 68.91507
3 - vs 1 0.26852280 23 147.8428 66.97324
4 - carb 1 0.68546077 24 148.5283 65.12126
5 - gear 1 1.56497053 25 150.0933 63.45667
6 - drat 1 3.34455117 26 153.4378 62.16190
7 - disp 1 6.62865369 27 160.0665 61.51530
8 - hp 1 9.21946935 28 169.2859 61.30730

#view final model
backward$coefficients

(Intercept) wt qsec am
9.617781 -3.916504 1.225886 2.935837
```

Here is how to interpret the results:

First, we fit a model using all 10 predictor variables and calculate the AIC of the model.

Next, we removed the variable (vs) that lead to the greatest reduction in AIC and also had a statistically significant reduction in AIC compared to the 9-predictor variable model.

Next, we removed the variable (carb) that lead to the greatest reduction in AIC and also had a statistically significant reduction in AIC compared to the 8-predictor variable model.

We repeated this process until removing any variable no longer led to a statistically significant reduction in AIC.

The final model turns out to be:

mpg = 9.62 - 3.92*wt + 1.23*qsec + 2.94*am

A Note on Using AIC

In the previous example, we chose to use AIC as the metric for evaluating the fit of various regression models.

AIC stands for Akaike information criterion and is calculated as:

AIC = 2K - 2*ln*(L)

where:

K: The number of model parameters.*ln*(L): The log-likelihood of the model. This tells us how likely the model is, given the data.

However, there are other metrics you might choose to use to evaluate the fit of regression models including cross-validation prediction error, Cp, BIC, AIC, or adjusted R2.

Fortunately, most statistical software allows you to specify which metric you would like to use when performing backward selection.

Additional Resources

The following tutorials provide additional information about regression models:

Introduction to Forward Selection