

How to Use stepAIC in R for Feature Selection?

Authored by
stats writer

November 19, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Use stepAIC in R for Feature Selection?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=97089>

The selection of optimal predictor variables is a critical stage in developing robust statistical models. In the world of data science and statistical computing, particularly within the R statistical environment, the **stepAIC** function serves as a powerful utility for automated feature selection. This function, residing in the **MASS** package, employs a stepwise algorithm to efficiently navigate the vast space of possible subsets of predictor variables, ultimately identifying the combination that yields the most parsimonious yet effective regression model.

The mechanism behind stepAIC involves iterative testing. It systematically adds or removes potential predictors one by one, evaluating the resulting model's quality using the Akaike information criterion (AIC) score. The process initiates typically from a full model (containing all potential predictors) or an empty model (containing none, except the intercept). It continues until the AIC score, which acts as the guiding metric, ceases to improve, thereby pinpointing the optimal subset of features. This methodology is versatile and applicable to various modeling frameworks, including both **linear regression** and **logistic regression** models.

Understanding Stepwise Feature Selection

Stepwise feature selection is a widely utilized methodology for tackling the challenge of **model overfitting**, particularly when dealing with datasets containing a high number of potentially correlated predictors. The fundamental objective is to balance predictive power with model simplicity. By reducing the number of input variables, we achieve models that are easier to interpret, computationally less demanding, and potentially more generalizable to new, unseen data.

There are three primary modes of stepwise selection, each dictating the strategy used for variable inclusion or exclusion. **Forward selection** starts with no predictors and sequentially adds the variable that provides the greatest statistical improvement. Conversely, **Backward elimination** begins with a full model (containing all predictors) and systematically removes the least significant variable at each step. The most common and often preferred approach implemented by the stepAIC function is **mixed selection** (or "both"), which allows for both adding and subtracting variables during the process, providing a more robust search mechanism across the possible model space.

While stepwise procedures offer an efficient, automated method for feature reduction, it is crucial to recognize that they are heuristic in nature. They do not guarantee finding the absolute best global model, as they only explore a path of locally optimal changes. However, when guided by a strong statistical metric like the Akaike information criterion, stepwise methods provide an excellent practical compromise between exhaustive model searching and computational feasibility, especially in preliminary data analysis phases.

The Role of the Akaike Information Criterion (AIC)

The true power of the `stepAIC` function lies in its reliance on the **Akaike information criterion (AIC)**. The **AIC** is an estimator of the relative quality of statistical models for a given set of data. It provides a means for model selection by estimating the out-of-sample prediction error, thereby quantifying how well a model fits the dataset while simultaneously penalizing model complexity.

In essence, the **AIC** operates on the principle of parsimony. It seeks to find the model that maximizes the amount of variation explained in the data (goodness of fit) while imposing a penalty for the inclusion of an excessive number of parameters (complexity). A model with too many parameters, even if it fits the training data perfectly, is highly prone to **overfitting**. Therefore, when comparing multiple models, the model that reports the **lowest AIC score** is generally considered the preferred choice, representing the optimal balance between fit and complexity.

The adoption of AIC contrasts sharply with methods that rely solely on measures like R-squared, which always increases or stays the same when new variables are added, regardless of their predictive value. Because the **AIC** includes a penalty term directly proportional to the number of parameters, it naturally guides the selection process towards simpler, more interpretable models that still retain strong explanatory power.

Mathematical Foundation of AIC Calculation

To fully appreciate how `stepwise regression` achieves its goal, it is beneficial to examine the mathematical formula defining the AIC. The calculation is designed to incorporate both the model's fit to the data and its structural complexity. This mathematical definition is standardized across most statistical software packages, ensuring consistent results during model comparison.

The standard formula for calculating the AIC is defined as follows:

$$AIC = 2K - 2\ln(L)$$

Let's precisely define the two critical components of this formula:

K: This represents the number of estimated **model parameters**. This count includes all predictor variables used in the `regression model`, plus the intercept term, and sometimes the error variance, depending on the specific model type. The penalty term, $2K$, therefore increases linearly with model complexity.

$\ln(L)$: This term is the **maximum value of the log-likelihood function** for the estimated model. The `log-likelihood` measures the probability of observing the data given the model and its estimated parameters. A higher `log-likelihood` indicates a better fit of the model to the training data. Most statistical software can automatically calculate this value.

The interplay between these two terms is what drives the optimization process. As the model complexity (K) increases, the penalty grows larger (positive $2K$). While adding more predictors generally increases the log-likelihood, this improvement must be substantial enough to overcome the complexity penalty. If a new predictor adds negligible fitting power, the resulting AIC score will increase, signaling that the variable should not be included.

Introducing the stepAIC() Function in R

The `stepAIC()` function is integral to advanced statistical modeling in R, though it requires loading the **MASS** package (Modern Applied Statistics with S) before execution. This function is specifically designed to accept a fitted model object, such as those derived from the `lm()` (linear model) or `glm()` (generalized linear model) functions, providing an automated wrapper around the iterative model selection search.

The basic operational syntax of the function is critical to understand, as it defines the scope and nature of the model search:

stepAIC(object, direction, ...)

The arguments define the critical inputs for the stepwise process:

object: This is the mandatory input--the name of a previously fitted full regression model. This initial model should contain all candidate predictor variables under consideration.

direction: This argument dictates the type of stepwise search employed. The possible values are **"backward"** (where the search starts full and removes variables), **"forward"** (where the search starts empty and adds variables), or **"both"** (mixed selection). Using "both" is generally the most comprehensive approach as it allows the algorithm to correct for suboptimal decisions made in earlier steps.

Once executed, the function iteratively refines the model, providing detailed output at each step, showcasing which variables are dropped or added, and the resulting change in the AIC value. The ultimate result of the function is the statistically preferred model object, ready for deeper analysis and interpretation.

Practical Example: Implementing stepAIC with the mtcars Dataset

To demonstrate the practical application of stepAIC, we will use the built-in **mtcars** dataset in R. This dataset contains comprehensive measurements on 11 different attributes for 32 different cars, serving as an excellent test case for feature selection in a linear modeling context.

We aim to find the best predictors for a car's gross horsepower (**hp**). We start by considering four potential predictor variables: miles per gallon (**mpg**), weight (**wt**), rear axle ratio (**drat**), and quarter

mile time (**qsec**). First, we view the data to familiarize ourselves with its structure:

#view first six rows of mtcars dataset

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 proceed by fitting the initial model that includes all four candidate predictors. We then execute the ``stepAIC()`` function, specifying the search direction as **"both"** to allow for maximum path flexibility in finding the model with the minimum AIC value:

library(MASS)

#fit initial multiple linear regression model

```
model <- lm(hp ~ mpg + wt + drat + qsec, data=mtcars)
```

#use both forward and backward selection to find model with lowest AIC

```
stepAIC(model, direction="both")
```

Start: AIC=226.88

```
hp ~ mpg + wt + drat + qsec
```

Df Sum of Sq RSS AIC

```
- drat 1 94.9 28183 224.98
```

```
- mpg 1 1519.4 29608 226.56
```

```
none 28088 226.88
```

```
- wt 1 3861.9 31950 229.00
```

```
- qsec 1 28102.2 56190 247.06
```

Step: AIC=224.98

```
hp ~ mpg + wt + qsec
```

Df Sum of Sq RSS AIC

```
- mpg 1 1424.5 29608 224.56
```

```
none 28183 224.98
```

```
+ drat 1 94.9 28088 226.88
```

```
- wt 1 3797.9 31981 227.03
- qsec 1 29625.1 57808 245.97
```

Step: AIC=224.56

```
hp ~ wt + qsec
```

Df Sum of Sq RSS AIC

```
none 29608 224.56
+ mpg 1 1425 28183 224.98
+ drat 1 0 29608 226.56
- wt 1 43026 72633 251.28
- qsec 1 52881 82489 255.35
```

Call:

```
lm(formula = hp ~ wt + qsec, data = mtcars)
```

Coefficients:

```
(Intercept) wt qsec
441.26 38.67 -23.47
```

Analyzing Stepwise Output and Model Convergence

The resulting output clearly maps the iterative decisions made by the stepwise algorithm. We observe the process of starting with a complex model and sequentially moving toward a simpler, parsimonious solution guided by the objective of minimizing the AIC score.

The interpretation of the iterative steps is as follows:

Starting Point: The initial model, including all four predictors ($hp \sim mpg + wt + drat + qsec$), yields an AIC value of **226.88**. The algorithm then evaluates the effect of removing each variable one by one. Removing **drat** results in the lowest subsequent AIC (224.98).

First Refinement: The model is reduced to $hp \sim mpg + wt + qsec$. Starting from this new model (AIC 224.98), the function again checks potential moves (removing **mpg**, removing **wt**, etc., or adding **drat** back). It determines that removing **mpg** leads to the next greatest reduction, achieving an AIC of **224.56**.

Final Convergence: The model is now $hp \sim wt + qsec$. In the final step, the algorithm checks if adding back the removed variables (**mpg** or **drat**) or removing the remaining ones would further reduce the AIC. Since the "none" option (retaining the current model structure) results in the lowest AIC (224.56), the search terminates, indicating that **wt** and **qsec** are the optimal feature subset.

The final regression model, having converged at the minimum local AIC, is defined by only two

highly significant predictors: vehicle weight (**wt**) and quarter mile time (**qsec**). This results in the final, simplified predictive equation:

$$hp = 441.26 + 38.67(wt) - 23.47(qsec)$$

This model is the most efficient choice among the subsets tested, achieving an AIC of **224.56**.

Advantages and Caveats of Using stepAIC

The `stepAIC()` function offers substantial advantages, primarily its ability to automate and accelerate the model selection process, which is invaluable when dealing with high-dimensional datasets. Its use of the AIC as a criterion provides a sound statistical basis for balancing the model's goodness-of-fit with its inherent complexity, leading to models that typically exhibit better generalization capabilities than those selected purely by maximizing R-squared values.

However, analysts must be aware of the inherent limitations of automated stepwise regression. Since the algorithm performs a sequential, local search, the final model is highly dependent on the initial set of variables and the specific path chosen during the iterations. It is mathematically possible that a configuration of variables skipped by the sequential path could result in a lower global AIC, meaning that the model found by `stepAIC()` is only locally optimal.

Furthermore, stepwise procedures can introduce a form of variable selection bias, leading to inflated measures of fit and reduced statistical power for hypothesis testing regarding the remaining variables. Therefore, while stepAIC is an invaluable tool for preliminary exploratory modeling, the resulting feature set should always be validated through robust techniques like cross-validation, and confirmed against domain knowledge to ensure the final, simple model is both statistically sound and theoretically justifiable.