How to Easily Select the Best Regression Model in R with regsubsets()

Authored by

stats writer

November 19, 2025

Regression analysis is a fundamental tool in statistical modeling, yet determining the optimal set of predictor variables--the inputs that best explain the response variable--remains a critical challenge. Including too many variables can lead to overfitting, where the model performs excellently on the training data but poorly on unseen data, while excluding key predictors results in biased estimates and poor predictive power. The process of finding the ideal balance between complexity and explanatory power is known as model selection.

Traditionally, researchers might rely on methods like forward, backward, or stepwise regression. However, these methods, while computationally efficient, are heuristic and do not guarantee finding the globally best model subset. For situations demanding rigor and an exhaustive exploration of all possibilities, specialized tools are required to evaluate every feasible combination of predictors against stringent statistical criteria such as the AIC (Akaike Information Criterion) or the BIC (Bayesian Information Criterion).

This need for comprehensive model search is precisely where the **regsubsets()** function, housed within the **leaps library** in R, proves invaluable. It provides statisticians and data scientists with a powerful, deterministic approach to systematically identifying the best subset of variables for a given regression problem. By leveraging this function, users can move beyond simple stepwise procedures and ensure that their final model is built upon the strongest statistical foundation possible.

The Role of `regsubsets()` in Model Selection

The core purpose of the **regsubsets()** function is to perform all-subsets regression, meaning it evaluates every single possible combination of predictor variables up to a specified maximum size. Unlike automated procedures that sequentially add or remove variables (like stepwise regression), this exhaustive approach guarantees that the globally optimal model for a given subset size is identified based on chosen metrics. This capability is essential when seeking the most parsimonious yet powerful explanatory model.

This functionality is provided by the **leaps** package in R, which is specifically designed for efficient subset selection in linear regression. When calling **regsubsets()**, the user defines the response variable and the pool of potential predictor variables. The function then processes these inputs and returns a structured object containing information on the best model identified for each possible number of predictors, ranging from one up to the total number of predictors specified.

Furthermore, the function offers flexibility in terms of the statistical criteria used for selection. While the summary output primarily displays model characteristics like R-squared and Adjusted R-squared, the function internally uses these measures, alongside metrics such as Mallow's Cp and BIC, to designate the "best" model for each subset size. Understanding these underlying metrics is key to correctly interpreting the final results and making an informed decision about the

dimensionality of the final model.

Setting Up the Regression Example with `mtcars`

To demonstrate the practical application of **regsubsets()**, we will utilize the well-known **mtcars dataset**, which is built into the R environment. This dataset provides comprehensive measurements on 11 different attributes for 32 automobiles (1973-74 models). It is frequently used in statistical teaching and practice due to its clean structure and relevant variables relating to automobile performance.

Before performing the subset selection, it is crucial to inspect the structure of the data. We use the standard **head()** function to view the initial rows of the dataset, confirming the variable names and data types. This initial exploration confirms the availability of relevant performance metrics such as horsepower (hp), miles per gallon (mpg), weight (wt), and other design characteristics.

For this specific example, our objective is to model the relationship where **hp** (horsepower) serves as the response variable. We will investigate a subset of four other variables as potential predictors. The goal is to determine the strongest combination of these predictors that collectively best explains the variation observed in horsepower among the different vehicle models.

#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

Our selected pool of four candidate predictor variables, which we hypothesize influence vehicle horsepower, includes the following attributes:

mpg (Miles per gallon)
wt (Weight)
drat (Rear axle ratio)
qsec (1/4 mile time)

Executing the Exhaustive Search Regression

The power of **regsubsets()** lies in its ability to systematically evaluate all 2k possible models derived from the k potential predictors. In our scenario, with four candidate predictors, the function will examine 24 = 16 different models (including the null model and the full model), identifying the best performing model for each size (1, 2, 3, and 4 predictors). This comprehensive evaluation contrasts sharply with stepwise regression, which follows a local optimization path that might miss the true global optimum.

To initiate this process, we first ensure the **leaps** package is loaded into the R session using the **library()** command. We then define our regression formula, specifying **hp** as the dependent variable predicted by the four independent variables: **mpg**, **wt**, **drat**, and **qsec**. The results of this extensive search are stored in the **bestSubsets** object.

Following the execution of the **regsubsets()** function, we use the **summary()** function on the resulting object to display the condensed statistical output. This summary provides a clear matrix view of the selected models, indicating which variables were chosen as optimal for models of increasing complexity. This initial output is essential for quick visual identification of the strongest predictor combinations.

library(leaps)

#find best regression model using exhaustive search
bestSubsets <- regsubsets(hp ~ mpg + wt + drat + qsec, data=mtcars)

#view basic results
summary(bestSubsets)

Subset selection object
Call: regsubsets.formula(hp ~ mpg + wt + drat + qsec, data = mtcars)
4 Variables (and intercept)
Forced in Forced out
mpg FALSE FALSE
wt FALSE FALSE
drat FALSE FALSE
qsec FALSE FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
mpg wt drat qsec
1 (1) "*" " " " " " "
2 (1) " " "*" " " "*"

3 (1) "*" "*" " " "*"
4 (1) "*" "*" "*" "*"

Interpreting the Subset Selection Output Matrix

The output generated by the **summary()** function, particularly the matrix at the bottom, is the primary source of information regarding the optimal subsets. This matrix lists model sizes (from 1 to 4 predictors) in the rows, and the potential predictor variables (mpg, wt, drat, qsec) in the columns. The presence of an asterisk (*****) in a column signifies that the corresponding predictor variable was included in the best model of that specific size, as determined by the evaluation criteria.

The interpretation of this matrix is straightforward and allows us to quickly identify the statistically superior variable combinations for models of varying complexity.

Model Size 1: For a model restricted to using only one predictor variable, the exhaustive search determined that **mpg** (Miles per Gallon) provides the highest explanatory power.

Model Size 2: When selecting the best model using exactly two predictors, the optimal combination consists of **wt** (Weight) and **qsec** (1/4 mile time). This two-variable model is statistically superior to any other possible pairing among the four candidates.

Model Size 3: The best model incorporating three predictor variables includes **mpg**, **wt**, and **qsec**. The addition of **drat** did not improve the model sufficiently to be selected over the other three variables at this size.

Model Size 4: The model with four predictors is the full model, incorporating **mpg**, **wt**, **drat**, and **qsec**.

Evaluating Model Quality Using Statistical Metrics

While the star matrix identifies the variables included in the best models of each size, it does not explicitly provide the numerical metrics used for comparison. To truly select the overall best model, we must delve into the various statistical measures calculated by **regsubsets()**. These metrics quantify the goodness-of-fit while often penalizing complexity, helping us determine the ideal trade-off. Key metrics available for extraction include the coefficient of determination (R-squared), the Residual Sum of Squares (RSS), the Adjusted R-squared, Mallow's Cp, and the BIC.

The R-squared value measures the proportion of variance in the response variable (hp) that is predictable from the predictor variables. However, R-squared always increases as more predictors are added, regardless of whether those predictors are meaningful. This inherent bias makes it a

poor metric for comparing models with different numbers of predictors, as it favors overly complex models.

To address the shortcomings of standard R-squared, the **Adjusted R-squared** metric introduces a penalty for the inclusion of unnecessary predictor variables. It is generally considered a superior measure for comparing models of varying sizes, as it increases only if the newly added variable significantly improves the model beyond what would be expected by chance. Similarly, Mallow's Cp and the BIC are powerful metrics used for penalized model selection, preferring models that strike a balance between high fit and low complexity.

By extracting these summary metrics, we can quantify the performance of each model subset identified. The selection process typically involves choosing the model that maximizes the Adjusted R-squared or minimizes criteria like RSS, Cp, or BIC. This quantitative evaluation step is crucial for transitioning from identifying potential subsets to selecting the final, definitive model for prediction or inference.

rsq: The standard R-squared for each model.
RSS: The Residual Sum of Squares for each model, measuring unexplained variance.
adjr2: The **Adjusted R-squared** for each model, penalized for complexity.
cp: Mallow's **Cp** statistic for each model, where values close to the number of predictors plus one are desirable.
bic: The **BIC** statistic for each model, where the lowest value indicates the preferred model.

Analyzing Adjusted R-squared for Optimal Selection

In many practical applications, the Adjusted R-squared (adjr2) is the preferred metric for initial model comparison because it is easy to interpret--it still represents the variance explained, but normalized for the degrees of freedom lost due to increased complexity. To select the best overall model, we must extract these values from the **bestSubsets** object and identify which subset size yields the maximum Adjusted R-squared value.

We access the Adjusted R-squared values using the syntax `summary(bestSubsets)$adjr2`. This returns a vector where each element corresponds to the maximum Adjusted R-squared achieved by the best model of that specific size (1 predictor, 2 predictors, 3 predictors, and so on).

#view adjusted R-squared value of each model
summary(bestSubsets)$adjr2

0.5891853 0.7828169 0.7858829 0.7787005

These numerical results allow us to perform a detailed comparison of the predictive power across

the identified subsets. We can now precisely quantify the performance associated with each optimal model size:

The one-predictor model (using **mpg**) achieved an Adjusted R-squared of approximately **0.589**.

The two-predictor model (using **wt** and **qsec**) achieved an Adjusted R-squared of approximately **0.783**.

The three-predictor model (using **mpg**, **wt**, and **qsec**) achieved the highest Adjusted R-squared value, approximately **0.786**.

The four-predictor model (the full model, including **drat**) saw a slight decrease in performance, yielding an Adjusted R-squared of approximately **0.779**.

Drawing Conclusions on the Optimal Model

The analysis of the Adjusted R-squared values provides clear evidence regarding the optimal model complexity for predicting horsepower (hp) based on the four candidate variables. The model with three predictors (**mpg**, **wt**, and **qsec**) yields the highest Adjusted R-squared (0.786). This implies that this specific combination of variables explains the largest proportion of variance in horsepower, after accounting for the degrees of freedom used.

It is particularly insightful to note that while the two-predictor model (0.783) performed extremely well, the addition of **mpg** in the three-variable model resulted in a marginal, yet measurable, improvement in explanatory power. Crucially, the inclusion of the fourth variable, **drat**, in the full model (Size 4) did not justify its presence, as evidenced by the subsequent drop in the Adjusted R-squared value from 0.786 to 0.779. This demonstrates that **drat** is likely redundant or contributes noise rather than signal when combined with the other three variables.

Ultimately, the selected model should be the one that maximizes the chosen metric, thereby offering the best balance of fit and parsimony. Based on the Adjusted R-squared criterion, the three-variable model containing **mpg**, **wt**, and **qsec** is the most statistically robust choice among all potential subsets. These values are essential inputs for practitioners, offering a quantitative basis for model selection when building reliable predictive statistical models.

By utilizing **regsubsets()** and carefully interpreting its output alongside appropriate statistical metrics, we can confidently identify a powerful and concise set of predictor variables, moving beyond reliance on guesswork or suboptimal stepwise procedures.