A Comprehensive Guide to Model Selection in R Using the regsubsets() Function

Authored by

Mohammed looti

November 15, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *A Comprehensive Guide to Model Selection in R Using the regsubsets() Function*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=1707

Mastering Model Selection with R's `regsubsets()` Function

In the intricate world of regression analysis, success hinges on building a predictive model that is both highly accurate and suitably simple. This critical process, formally known as model selection, involves navigating a complex trade-off: maximizing the explanatory power derived from available predictor variables while rigorously avoiding common pitfalls such as overfitting. A model saturated with too many variables risks becoming overly specific to the training data, leading to poor generalization on new observations, whereas a model with too few might fail to capture essential, underlying data relationships.

Fortunately, modern statistical programming environments provide sophisticated tools to streamline this crucial decision. The statistical software R offers the powerful **`regsubsets()`** function, which is the cornerstone of the specialized **`leaps`** package. This function is expertly designed to execute best subsets regression, a systematic methodology that evaluates every potential combination of input variables. Its goal is to identify the statistically optimal linear regression model for any specified number of predictors.

This comprehensive guide will walk you through the practical application of **`regsubsets()`** within the R environment. We will demonstrate how to systematically identify the superior subset of predictor variables, accurately interpret the detailed results, and leverage quantitative statistical metrics--including Adjusted R-squared and Mallows' Cp--to inform the final model selection decision. By the end of this tutorial, you will possess the clarity and skills necessary to employ this function effectively, resulting in the creation of more robust, efficient, and highly interpretable regression models.

Preparation: Setting Up the R Environment and Data

To commence our practical exploration of best subsets regression, the first critical step is configuring the necessary R environment. The core functionality we intend to use, the **`regsubsets()`** function, is contained within the **`leaps`** package. If this package is not yet installed on your system, you must first execute the command `install.packages("leaps")`, and subsequently load it into your current R session using `library(leaps)`. This preliminary step ensures that the specialized algorithms required for efficient exploration of the model space are fully accessible.

For demonstration purposes, we will utilize the universally recognized, built-in **`mtcars`** dataset available in R. This dataset provides comprehensive specifications for 32 different automobiles, encompassing 11 key characteristics. These attributes include crucial variables such as miles per gallon (`mpg`), the number of cylinders (`cyl`), gross horsepower (`hp`), and vehicle weight (`wt`). Given its moderate size and diverse collection of quantitative metrics, the **`mtcars`** dataset serves as an

ideal platform for showcasing rigorous model selection techniques.

Prior to executing any modeling process, a fundamental step in the data analysis workflow is a thorough examination of the data structure. By viewing the initial rows of the **`mtcars`** dataset, we gain immediate insight into the format, scale, and specific data types of the variables we will manipulate. This crucial initial inspection confirms that the data is structured appropriately for linear regression modeling and the subsequent variable selection procedure.

 # Load the leaps package and view the first six rows of the 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

Defining the Model: Response and Candidate Predictors

Our core analytical objective in this demonstration is the development of a linear regression model specifically designed to predict vehicle horsepower. Consequently, the variable `hp` (horsepower) will be designated as our primary response variable. Our investigation seeks to quantify how various measurable vehicle characteristics influence this critical performance metric. To achieve a meaningful and robust prediction, we must judiciously select an initial pool of candidate predictor variables from the available attributes in the **`mtcars`** dataset.

The initial selection of potential variables is typically guided by established domain knowledge, empirical evidence from previous studies, or specific hypotheses concerning the relationships within the data. For the purposes of this walkthrough, we have chosen four variables that are logically and empirically connected to automotive performance and physical specifications. Our subsequent application of the **`regsubsets()`** function will serve to rigorously test which combination of these candidates ultimately yields the most powerful and efficient regression model for predicting horsepower.

The specific set of four potential predictor variables selected for exhaustive evaluation are:

`mpg`: Miles per gallon (a key inverse indicator of power consumption).
`wt`: Weight (measured in 1000 lbs, representing the vehicle's mass).
`drat`: Rear axle ratio (a mechanical specification influencing effective torque).

qsec: 1/4 mile time (a direct, quantitative measure of acceleration capability).

This curated selection provides a diverse spectrum of mechanical and performance attributes. By systematically executing best subsets regression via **regsubsets()**, we can precisely evaluate both their individual predictive efficacy and their synergistic combined strength in modeling the response variable, hp.

Executing the Best Subsets Regression Search

We now proceed to the core operation: utilizing the **regsubsets()** function to conduct an exhaustive search across all possible subsets of our defined predictor variables. This method is crucial to best subsets regression because it guarantees the identification of the statistically optimal regression model for every possible size--that is, models containing exactly one, two, three, or four predictors. By ensuring that every combination is evaluated, we confirm that the strongest performing model for each level of complexity is found.

The implementation syntax for **regsubsets()** is intentionally designed to resemble other familiar R modeling functions. It requires the specification of the standard regression formula (`response ~ predictor1 + predictor2 + ...`) and the dataset to be analyzed. The function then automatically assesses models ranging from the most parsimonious single-predictor configuration up to the complete model incorporating all specified input variables.

library(leaps)

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

# Display the high-level summary of the selection process
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 ) "*" "*" "*" "*"
```

The resulting output generated by the `summary(bestSubsets)` command provides a concise summary of the systematic selection process. It details the precise formula utilized, the total number of variables assessed, and confirms the use of the rigorous exhaustive search algorithm. Most importantly, the matrix displayed at the bottom of the output clearly outlines the specific composition of the "best" regression model identified for each size, offering an immediate visual representation of variable inclusion.

Interpreting the Model Inclusion Matrix

The most instructive component of the **`regsubsets()`** output is the inclusion matrix, easily recognizable by the pattern of asterisks (**`*`**) and spaces. Each horizontal row in this matrix corresponds directly to the single best performing linear regression model found for a specific number of predictor variables. The presence of an asterisk in a column indicates that the corresponding predictor variable is an active component of the model for that size, while a space signifies its exclusion.

By carefully examining the structure of this matrix, we can draw precise conclusions about the optimal variable combinations at different complexity levels:

Model Size 1 (One Predictor): The analysis indicates that the strongest single-predictor model is constructed using only **`mpg`** (miles per gallon). This result demonstrates that `mpg`, when viewed in isolation, possesses the highest singular predictive power over the response variable, `hp`.
Model Size 2 (Two Predictors): The optimal two-variable combination includes **`wt`** (weight) and **`qsec`** (1/4 mile time). This pairing explains the observed variation in horsepower more effectively than any other possible pair of predictors.
Model Size 3 (Three Predictors): The top three-variable model incorporates **`mpg`**, **`wt`**, and **`qsec`**. Notably, `drat` is still excluded, suggesting that its marginal contribution to explaining horsepower is not sufficient to supersede the combined predictive strength offered by the other three variables.
Model Size 4 (Four Predictors): This final row represents the full model, which, by definition, includes all four potential predictor variables: **`mpg`**, **`wt`**, **`drat`**, and **`qsec`**. This configuration represents the maximum complexity considered in this specific exhaustive search.

While this matrix successfully identifies the best candidate models across various sizes, selecting the definitive "best" model requires moving beyond simple variable inclusion and evaluating these candidates using rigorous statistical criteria designed to balance fit and simplicity.

Evaluating Parsimony: The Role of Statistical Metrics

The **`regsubsets()`** function excels at identifying the best variable subsets for each model size, but it does not automatically designate which size is superior overall. To make a final, statistically sound decision, analysts must analyze the candidate models using established model selection metrics. These metrics provide quantitative measures of a model's goodness of fit, its parsimony (simplicity), and its reliability, allowing for a necessary comparison between models of differing complexity.

The `summary()` function, when applied to the **`regsubsets()`** output object, allows us to extract several crucial metrics. Key metrics used for evaluating the delicate balance between predictive strength and model parsimony include:

R-squared (`rsq`): Quantifies the proportion of variance in the response variable that is explained by the predictors. While higher values denote a better fit, R-squared will always increase with the addition of any predictor, regardless of its significance, making it unsuitable for comparing models of different sizes.
Residual Sum of Squares (`RSS`): Calculates the sum of squared errors between the observed data points and the model's predicted values. A lower RSS indicates a closer fit, but like R-squared, it inherently favors models with greater complexity.
Adjusted R-squared (`adjr2`): A modification of R-squared that introduces a penalty for the inclusion of extraneous variables. This crucial adjustment makes Adjusted R-squared the preferred metric for comparing models of varying sizes; the objective is always to **maximize** this value.
Mallows' Cp (`cp`): Provides an estimate of the total mean squared error of prediction. The ideal model is one where Mallows' Cp is approximately equal to *p* (where *p* is the number of predictors plus the intercept). We actively seek Cp values close to this benchmark.
Bayesian Information Criterion (`bic`): A powerful criterion derived from information theory that imposes a strong penalty on models containing more parameters. Generally, the model exhibiting the **lowest** BIC value is considered the most favorable, as it represents the optimal balance between fit and parsimony.

To make a data-driven choice, let's specifically analyze the Adjusted R-squared values for the top models identified by **`regsubsets()`**. Maximizing this metric is essential for selecting the model that provides the strongest predictive power while maintaining a desirable level of simplicity.

 # Extract and view Adjusted R-squared value of each optimal model
summary(bestSubsets)$adjr2

0.5891853 0.7828169 0.7858829 0.7787005

A detailed inspection of the calculated Adjusted R-squared values across the four candidate models reveals a clear optimal choice:

The one-predictor model (**mpg**) yields an Adjusted R-squared of approximately **0.589**.
The two-predictor model (**wt** and **qsec**) results in a substantial jump in explanatory power, reaching **0.783**.
The three-predictor model (**mpg**, **wt**, and **qsec**) achieves the highest Adjusted R-squared value at approximately **0.786**. This peak value confirms this model represents the best overall balance of fit and parsimony among all tested candidates.
The full four-predictor model (which includes **drat**) sees the Adjusted R-squared slightly decline to **0.779**. This quantitative decrease confirms that `drat` does not contribute enough unique explanatory power to justify the penalty incurred for increased complexity.

Based strictly on the criterion of maximizing the Adjusted R-squared metric, the three-predictor model, featuring the variables **mpg**, **wt**, and **qsec**, is identified as the most optimal and parsimonious choice for predicting horsepower in the **mtcars** dataset.

Conclusion: Summary and Considerations for Large Datasets

The **regsubsets()** function, housed within the **leaps** package, offers an indispensable, systematic methodology for rigorous model selection in regression analysis. By performing an exhaustive search, it empowers analysts to confidently identify the statistically strongest subset of predictor variables for any defined level of model complexity, effectively preventing the oversight of potentially superior combinations. This systematic capability is paramount for constructing statistical models that are both highly interpretable and possess robust out-of-sample predictive capabilities.

The proper interpretation of the **regsubsets()** inclusion matrix, coupled with analytical scrutiny of key statistical metrics such as Adjusted R-squared, Mallows' Cp, and the Bayesian Information Criterion (BIC), facilitates a comprehensive evaluation of all candidate models. Our demonstration using the **mtcars** dataset successfully located the optimal "sweet spot": a three-predictor model that maximized the explained variance of the response variable without introducing the penalties associated with unnecessary complexity, thereby maximizing its generalizability.

It is essential to recognize a key limitation: while **regsubsets()** excels at best subsets regression, the underlying exhaustive search method can become computationally prohibitive when dealing with datasets that contain a very large number of potential predictors (e.g., dozens or hundreds of variables). In such high-dimensional scenarios, analysts should pivot toward more computationally efficient feature selection alternatives. These methods commonly include forms of stepwise regression (forward, backward, or hybrid selection) or powerful regularization techniques such as Lasso or Ridge regression. Always align your chosen model selection strategy with the specific

computational constraints and analytical goals of your regression analysis project.

Additional Resources for Further Study

To further solidify your expertise in R programming and statistical modeling, we recommend exploring the official documentation and trusted informational resources listed below:

Official `regsubsets()` documentation
Wikipedia: Model Selection
Wikipedia: Linear Regression