

# Package ‘VariableSelection’

February 17, 2026

**Title** Select Variables for Linear Models

**Version** 1.0.0

**Description** Provides variable selection for linear models and generalized linear models using Bayesian information criterion (BIC) and model posterior probability (MPP).

Given a set of candidate predictors, it evaluates candidate models and returns model-level summaries (BIC and MPP) and predictor-level posterior inclusion probabilities (PIP).

For more details see Xu, S., Ferreira, M. A., & Tegge, A. N. (2025) <[doi:10.48550/arXiv.2510.02628](https://doi.org/10.48550/arXiv.2510.02628)>.

**License** GPL-3

**Encoding** UTF-8

**RoxygenNote** 7.3.2

**Imports** stats (>= 4.2.2), GA (>= 3.2.3), memoise (>= 2.0.1)

**Suggests** knitr, rmarkdown, testthat (>= 3.0.0)

**VignetteBuilder** knitr

**LazyData** true

**Depends** R (>= 3.5.0)

**Config/testthat/edition** 3

**NeedsCompilation** no

**Author** Shuangshuang Xu [aut, cre]

**Maintainer** Shuangshuang Xu <xshuangshuang@vt.edu>

**Repository** CRAN

**Date/Publication** 2026-02-17 16:00:08 UTC

## Contents

dat	2
glm.best	2
glmdat	4
lm.best	4
modelselect.glm	5
modelselect.lm	6

---

dat	<i>A data frame contains dependent variable and continuous independent variables</i>
-----	--

---

**Description**

A data frame with seven columns. The independent variables are in the first six columns. The dependent variable is in the seventh column.

**Usage**

```
dat
```

**Format**

**dat:**  
A data frame.

**Description**

Description: `glm.best` is used to fit generalized linear model for the best model provided by `modelselect.glm`.

**Usage**

```
glm.best(  
  object,  
  family,  
  method = "models",  
  threshold = 0.95,  
  x = FALSE,  
  y = FALSE  
)
```

## Arguments

object	the model selection result from <code>modelselect.glm</code> .
family	a character string naming a family function describing the error distribution to be used in the model.
method	the criteria to do model select. <code>method = "models"</code> selects the best model by the highest posterior probabilities. <code>method = "variables"</code> selects the variables in the best model by the posterior inclusion probabilities which are larger than the threshold.
threshold	The threshold for variable selection. The variables with posterior inclusion probability larger than the threshold are selected in the best model. The default is 0.95.
x, y	logicals. If TRUE the corresponding components (the best model predictor matrix, the response) of the fit are returned.

## Value

An object of class "glm", which is a list containing the following components:

- `coefficients` a named vector of coefficients.
- `residuals` the working residuals, that is the residuals in the final iteration of the IWLS fit.
- `fitted.values` the fitted mean values, obtained by transforming the linear predictors by the inverse of the link function.
- `rank` the numeric rank of the fitted linear model.
- `family` the family object used.
- `linear.predictors` the linear fit on the link scale.
- `deviance` up to a constant, minus twice the maximized log-likelihood.
- `aic` A version of Akaike's An Information Criterion, minus twice the maximized log-likelihood plus twice the number of parameters, computed by the `aic` component of the family.
- `null.deviance` The deviance for the null model, comparable with `deviance`. The null model will include the offset, and an intercept if there is one in the model.
- `iter` the number of iterations of IWLS used.
- `weights` the working weights, that is the weights in the final iteration of the IWLS fit.
- `prior.weights` the weights initially supplied, a vector of 1s if none were.
- `df.residual` the residual degrees of freedom.
- `df.null` the residual degrees of freedom for the null model.
- `y` if requested, the response vector used.
- `converged` logical. Was the IWLS algorithm judged to have converged?
- `boundary` logical. Is the fitted value on the boundary of the allowable values?
- `model` if requested (the default), the model frame used.
- `call` the matched call.
- `formula` the formula supplied.
- `terms` the `terms.object` used.
- `data` the data argument.
- `threshold` the threshold used for `method = "variables"`.

---

**glmdat***A data frame contains dependent variable and binary independent variables*

---

**Description**

A data frame with seven columns. The independent variables are in the first six columns. The dependent variable is in the seventh column.

**Usage**

```
glmdat
```

**Format**

**glmdat:**

A data frame.

---

**lm.best***Title: Fitting linear models for the best model*

---

**Description**

Description: `lm.best` is used to fit linear model for the best model provided by `modelselect.lm`.

**Usage**

```
lm.best(object, method = "models", threshold = 0.95, x = FALSE, y = FALSE)
```

**Arguments**

**object** the model selection result from `modelselect.lm`.

**method** the criteria to do model select. `method = "models"` selects the best model by the highest posterior probabilities. `method = "variables"` selects the variables in the best model by the posterior inclusion probabilities which are larger than the threshold.

**threshold** The threshold for variable selection. The variables with posterior inclusion probability larger than the threshold are selected in the best model. The default is 0.95.

**x, y** logicals. If TRUE the corresponding components (the best model predictor matrix, the response) of the fit are returned.

**Value**

An object of class "lm", which is a list containing the following components:

**coefficients** A named vector of coefficients.  
**residuals** The residuals, that is the response minus the fitted values.  
**fitted.values** The fitted mean values.  
**rank** The numeric rank of the fitted linear model.  
**df.residual** The residual degrees of freedom.  
**call** The matched call.  
**terms** The **terms** object used.  
**model** (If requested) the model frame used.  
**qr** (If requested) the QR decomposition of the design matrix.  
**xlevels** (If the model formula includes factors) a record of the levels of the factors.  
**contrasts** (If the model formula includes factors) the contrasts used.  
**offset** The offset used.  
**threshold** the threshold used for method = "variables".

**Description**

Description: use BIC to do variable selection.

**Usage**

```
modelselect.glm(  
  formula,  
  data,  
  family,  
  GA_var = 16,  
  maxiterations = 2000,  
  runs_til_stop = 1000,  
  monitor = TRUE,  
  popSize = 100,  
  verbose = TRUE  
)
```

## Arguments

formula	an object of class "formula": a symbolic description of the model to be fitted. A typical model has the form <code>response ~ terms</code> where <code>response</code> is the (numeric) response vector and <code>terms</code> is a series of terms which specifies a linear predictor for <code>response</code> . A <code>terms</code> specification of the form <code>first + second</code> indicates all the terms in <code>first</code> together with all the terms in <code>second</code> with duplicates removed. A specification of the form <code>first:second</code> indicates the set of terms obtained by taking the interactions of all terms in <code>first</code> with all terms in <code>second</code> . The specification <code>first*second</code> indicates the cross of <code>first</code> and <code>second</code> . This is the same as <code>first + second + first:second</code> .
data	an data frame containing the variables in the model.
family	a character string naming a family function describing the error distribution to be used in the model.
GA_var	if the number of variables is smaller than <code>GA_var</code> , then do exhaustive model search, otherwise use genetic algorithm to do stochastic model search.
maxiterations	the maximum number of iterations to run before the GA search is halted.
runs_til_stop	the number of consecutive generations without any improvement in the best fitness value before the GA is stopped.
monitor	a logical defaulting to TRUE showing the evolution of the search. If <code>monitor</code> = FALSE, any output is suppressed.
popSize	the population size.
verbose	Logical; if TRUE, print a brief summary of results.

## Value

`modelselect.glm` returns a list containing the following components:

`models` A data frame of candidate models' BIC and posterior probabilities, sorted by decreasing posterior probability  
`variables` A data frame of candidate variables' posterior inclusion probabilities  
`data` The data with variables in the formula.

The function `glm.best` is used to obtain the linear fitting to the best model by posterior probability or by controlling variables' posterior inclusion probabilities.

## Description

Description: use BIC to do variable selection.

**Usage**

```
modelselect.lm(
  formula,
  data,
  GA_var = 16,
  maxiterations = 2000,
  runs_til_stop = 1000,
  monitor = TRUE,
  popSize = 100,
  verbose = TRUE
)
```

**Arguments**

formula	an object of class "formula": a symbolic description of the model to be fitted. A typical model has the form <code>response ~ terms</code> where <code>response</code> is the (numeric) response vector and <code>terms</code> is a series of terms which specifies a linear predictor for <code>response</code> . A <code>terms</code> specification of the form <code>first + second</code> indicates all the terms in <code>first</code> together with all the terms in <code>second</code> with duplicates removed. A specification of the form <code>first:second</code> indicates the set of terms obtained by taking the interactions of all terms in <code>first</code> with all terms in <code>second</code> . The specification <code>first*second</code> indicates the cross of <code>first</code> and <code>second</code> . This is the same as <code>first + second + first:second</code> .
data	an data frame containing the variables in the model.
GA_var	if the number of variables is smaller than <code>GA_var</code> , then do exhaustive model search, otherwise use genetic algorithm to do stochastic model search.
maxiterations	the maximum number of iterations to run before the GA search is halted.
runs_til_stop	the number of consecutive generations without any improvement in the best fitness value before the GA is stopped.
monitor	a logical defaulting to TRUE showing the evolution of the search. If <code>monitor = FALSE</code> , any output is suppressed.
popSize	the population size.
verbose	Logical; if TRUE, print a brief summary of results.

**Value**

`modelselect.lm` returns a list containing the following components:

- `models` A data frame of candidate models' BIC and posterior probabilities, sorted by decreasing posterior probability
- `variables` A data frame of candidate variables' posterior inclusion probabilities
- `data` The data with variables in the formula.

The function `lm.best` is used to obtain the linear fitting to the best model by posterior probability or by controlling variables' posterior inclusion probabilities.

# Index

\* **datasets**  
  dat, [2](#)  
  glmdat, [4](#)  
  
dat, [2](#)  
  
glm.best, [2](#)  
glmdat, [4](#)  
  
lm.best, [4](#)  
  
modelselect.glm, [5](#)  
modelselect.lm, [6](#)  
  
terms.object, [3](#)