

www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

TextbookMultivariate Adaptive Regression Splines

What can we help you find?

Search

Looking for info about statistics?

We wrote the book on it.
And you can read it for free!

- Elementary Concepts
- Statistics Glossary
- Basic Statistics
- ANOVA / MANOVA
- Association Rules
- Boosting Trees
- Canonical Analysis
- CHAID Analysis
- C & R Trees
- Classification Trees
- Cluster Analysis
- Correspondence Analysis
- Data Mining Techniques
- Discriminant Analysis
- Distribution Fitting
- Experimental Design
- Factor Analysis
- General Discrim. Analysis
- General Linear Models
- Generalized Additive Mod.
- Generalized Linear Mod.
- General Regression Mod.
- Graphical Techniques
- Ind.Components Analysis
- Linear Regression
- Log-Linear Analysis
- MARSplines
- Machine Learning
- Multidimensional Scaling
- Neural Networks
- Nonlinear Estimation
- Nonparametric Statistics
- Partial Least Squares
- Power Analysis
- Process Analysis
- Quality Control Charts
- Reliability / Item Analysis
- SEPATH (Structural eq.)
- Survival Analysis
- Text Mining
- Time Series / Forecasting

Multivariate Adaptive Regression Splines (MARSplines)









- [Introductory Overview](#)
 - [Regression Problems](#)
 - [Multivariate Adaptive Regression Splines](#)
 - [Model Selection and Pruning](#)
 - [Applications](#)
- [Technical Notes: The MARSplines Algorithm](#)
- [Technical Notes: The MARSplines Model](#)

Introductory Overview

Multivariate Adaptive Regression Splines (MARSplines) is an implementation of techniques popularized by Friedman (1991) for solving regression-type problems (see also, [Multiple Regression](#)), with the main purpose to predict the values of a continuous dependent or outcome variable from a set of independent or predictor variables. There are a large number of methods available for fitting models to continuous variables, such as a linear regression [e.g., [Multiple Regression](#), [General Linear Model \(GLM\)](#)], nonlinear regression ([Generalized Linear/Nonlinear Models](#)), regression trees (see [Classification and Regression Trees](#)), [CHAID](#), [Neural Networks](#), etc. (see also Hastie, Tibshirani, and Friedman, 2001, for an overview).

MARSplines is a nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, MARSplines constructs this relation from a set of coefficients and basis functions that are entirely "driven" from the regression data. In a sense, the method is based on the "divide and conquer" strategy, which partitions the input space into regions, each with its own regression equation. This makes MARSplines particularly suitable for problems with higher input dimensions (i.e., with more than 2 variables), where the [curse of dimensionality](#) would likely create problems for other techniques.

The MARSplines technique has become particularly popular in the area of [data mining](#) because it does not assume or impose any particular type or class of relationship (e.g., linear, logistic, etc.) between the predictor variables and the dependent (outcome) variable of interest. Instead,

-  [Variance Components](#)
-  [Statistical Advisor](#)
-  [Distribution Tables](#)
-  [References Cited](#)
-  [Send Comments](#)
-  [Business Solutions](#)
-  [Free Resources](#)
-  [About Textbook](#)

useful models (i.e., models that yield accurate predictions) can be derived even in situations where the relationship between the predictors and the dependent variables is non-monotone and difficult to approximate with parametric models. For more information about this technique and how it compares to other methods for nonlinear regression (or regression trees), see Hastie, Tibshirani, and Friedman (2001).

REGRESSION PROBLEMS

Regression problems are used to determine the relationship between a set of dependent variables (also called output, outcome, or response variables) and one or more independent variables (also known as input or predictor variables). The dependent variable is the one whose values you want to predict, based on the values of the independent (predictor) variables. For instance, one might be interested in the number of car accidents on the roads, which can be caused by 1) bad weather and 2) drunk driving. In this case one might write, for example,

$$\text{Number_of_Accidents} = \text{Some Constant} + 0.5 * \text{Bad_Weather} + 2.0 * \text{Drunk_Driving}$$

The variable *Number of Accidents* is the dependent variable that is thought to be caused by (among other variables) *Bad Weather* and *Drunk Driving* (hence the name dependent variable). Note that the independent variables are multiplied by factors, i.e., *0.5* and *2.0*. These are known as regression coefficients. The larger these coefficients, the stronger the influence of the independent variables on the dependent variable. If the two predictors in this simple (fictitious) example were measured on the same scale (e.g., if the variables were standardized to a mean of *0.0* and standard deviation *1.0*), then *Drunk Driving* could be inferred to contribute 4 times more to car accidents than *Bad Weather*. (If the variables are not measured on the same scale, then direct comparisons between these coefficients are not meaningful, and, usually, some other standardized measure of predictor "importance" is included in the results.)

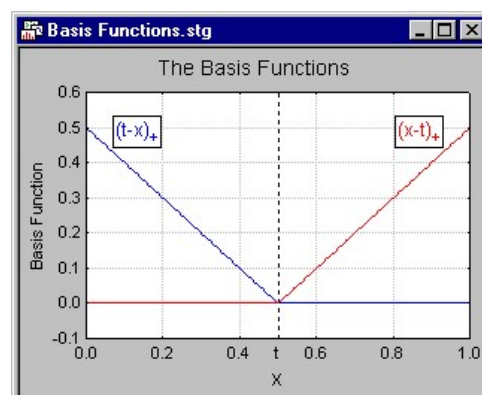
For additional details regarding these types of statistical models, refer to [Multiple Regression](#) or [General Linear Models \(GLM\)](#), as well as [General Regression Models \(GRM\)](#). *In general, the social and natural sciences regression procedures are widely used in research. Regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ...". For example, educational researchers might want to learn what the best predictors of success in high-school are. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether a new immigrant group will adapt and be absorbed into society.*

MULTIVARIATE ADAPTIVE REGRESSION SPLINES

The car accident example we considered previously is a typical application for linear regression, where the response variable is hypothesized to depend linearly on the predictor variables. Linear regression also falls into the category of so-called parametric regression, which assumes that the nature of the relationships (but not the specific parameters) between the dependent and independent variables is known *a priori* (e.g., is linear). By contrast, nonparametric regression (see [Nonparametrics](#)) does not make any such assumption as to how the dependent variables are related to the predictors. Instead it allows the regression function to be "driven" directly from data.

Multivariate Adaptive Regression Splines is a nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. Instead, MARSplines constructs this relation from a set of coefficients and so-called basis functions that are entirely determined from the regression data. You can think of the general "mechanism" by which the MARSplines algorithm operates as multiple piecewise linear regression (see [Nonlinear Estimation](#)), where each breakpoint (estimated from the data) defines the "region of application" for a particular (very simple) linear regression equation.

Basis functions. Specifically, MARSplines uses two-sided truncated functions of the form (as shown below) as basis functions for linear or nonlinear expansion, which approximates the relationships between the response and predictor variables.



Shown above is a simple example of two basis functions $(t-x)_+$ and $(x-t)_+$ (adapted from Hastie, et al., 2001, Figure 9.9). Parameter t is the knot of the basis functions (defining the "pieces" of the piecewise linear regression); these knots (parameters) are also determined from the data. The "+" signs next to the terms $(t-x)$ and $(x-t)$ simply denote that only positive results of the respective equations are considered; otherwise the respective functions evaluate to zero. This can also be seen in the illustration.

The MARSplines model. The basis functions together with the model parameters (estimated via [least squares estimation](#)) are combined to produce the predictions given the inputs. The general MARSplines model equation (see Hastie et al., 2001, equation 9.19) is given as:

$$y = f(X) = \beta_0 + \sum_{m=1}^M \beta_m h_m(X)$$

where the summation is over the M nonconstant terms in the model (further details regarding the model are also provided in [Technical Notes](#)). To summarize, y is predicted as a function of the predictor variables X (and their interactions); this function consists of an intercept parameter (β_0) and the weighted (by β_m) sum of one or more basis functions $h_m(X)$, of the kind illustrated earlier. You can also think of this model as "selecting" a weighted sum of basis functions from the set of (a large number of) basis functions that span all values of each predictor (i.e., that set would consist of one basis function, and parameter t , for each distinct value for each predictor variable). The MARSplines algorithm then searches over the space of all inputs and predictor values (knot locations t) as well as interactions between variables. During this search, an increasingly larger number of basis functions are added to the model (selected from the set of possible basis functions), to maximize an overall least squares goodness-of-fit criterion. As a result of these operations, MARSplines automatically determines the most important independent variables as well as the most significant interactions among them. The details of this algorithm are further described in [Technical Notes](#), as well as in Hastie et al., 2001).

Categorical predictors. In practice, both continuous and categorical predictors could be used, and will often yield useful results. However, the basic MARSplines algorithm assumes that the predictor variables are continuous in nature, and, for example, the computed knots program will usually not coincide with actual class codes found in the categorical predictors. For a detailed discussion of categorical predictor variables in MARSplines, see Friedman (1993).

Multiple dependent (outcome) variables. The MARSplines algorithm can be applied to multiple dependent (outcome) variables. In this case, the algorithm will determine a common set of basis functions in the predictors, but estimate different coefficients for each dependent variable. This method of treating multiple outcome variables is not unlike some [neural networks](#) architectures, where multiple outcome variables can be predicted from common neurons and hidden layers; in

the case of MARSplines, multiple outcome variables are predicted from common basis functions, with different coefficients.

MARSplines and classification problems. Because MARSplines can handle multiple dependent variables, it is easy to apply the algorithm to classification problems as well. First, code the classes in the categorical response variable into multiple indicator variables (e.g., 1 = observation belongs to class k, 0 = observation does not belong to class k); then apply the MARSplines algorithm to fit a model, and compute predicted (continuous) values or scores; finally, for prediction, assign each case to the class for which the highest score is predicted (see also Hastie, Tibshirani, and Freedman, 2001, for a description of this procedure). Note that this type of application will yield heuristic classifications that may work very well in practice, but is not based on a statistical model for deriving classification probabilities.

MODEL SELECTION AND PRUNING

In general, nonparametric models are adaptive and can exhibit a high degree of flexibility that may ultimately result in [overfitting](#) if no measures are taken to counteract it. Although such models can achieve zero error on training data, they have the tendency to perform poorly when presented with new observations or instances (i.e., they do not generalize well to the prediction of "new" cases). MARSplines, like most methods of this kind, tend to overfit the data as well. To combat this problem, MARSplines uses a pruning technique (similar to [pruning](#) in classification trees) to limit the complexity of the model by reducing the number of its basis functions.

MARSplines as a predictor (feature) selection method. This feature - the selection of and pruning of basis functions - makes this method a very powerful tool for predictor selection. The MARSplines algorithm will pick up only those basis functions (and those predictor variables) that make a "sizeable" contribution to the prediction (refer to [Technical Notes](#) for details).

APPLICATIONS

Multivariate Adaptive Regression Splines have become very popular recently for finding predictive models for "difficult" [data mining](#) problems, i.e., when the predictor variables do not exhibit simple and/or monotone relationships to the dependent variable of interest. Alternative models or approaches that you can consider for such cases are [CHAID](#), [Classification and Regression Trees](#), or any of the many [Neural Networks](#) architectures available. Because of the specific manner in which MARSplines selects predictors ([basis functions](#)) for the model, it does generally "well" in situations where regression-tree models are also appropriate, i.e., where hierarchically organized successive splits on the predictor variables yield good (accurate) predictions. In fact, instead of considering this technique as a generalization of multiple regression (as it was presented in this introduction), you may consider MARSplines as a generalization of regression trees, where the "hard" binary splits are replaced by "smooth" basis functions. Refer to Hastie, Tibshirani, and Friedman (2001) for additional details.

Technical Notes: The MARSplines Algorithm

[To index](#)

Implementing MARSplines involves a two step procedure that is applied successively until a desired model is found. In the first step, we build the model, i.e. increase its complexity by adding basis functions until a preset (user-defined) maximum level of complexity has been reached. Then we begin a backward procedure to remove the least significant basis functions from the model, i.e. those whose removal will lead to the least reduction in the (least-squares) goodness of fit. This algorithm is implemented as follows:

1. Start with the simplest model involving only the constant basis function.
2. Search the space of basis functions, for each variable and for all possible knots, and add those which maximize a certain measure of goodness of fit (minimize prediction error).

3. Step 2 is recursively applied until a model of pre-determined maximum complexity is derived.
4. Finally, in the last stage, a pruning procedure is applied where those basis functions are removed that contribute least to the overall (least squares) goodness of fit.

Technical Notes: The Multivariate Adaptive Regression Splines (MARSplines) Model [To index](#)

The MARSplines algorithm builds models from two sided truncated functions of the predictors (x) of the form:

$$(x-t)_+ = \begin{cases} x-t & x > t \\ 0 & \text{otherwise} \end{cases}$$

These serve as **basis functions** for linear or nonlinear expansion that approximates some true underlying function $f(x)$.

The MARSplines model for a dependent (outcome) variable y , and M terms, can be summarized in the following equation:

$$y = f(x) = \beta_0 + \sum_{m=1}^M \beta_m H_{km}(x_{v(k,m)})$$

where the summation is over the M terms in the model, and b_0 and b_m are parameters of the model (along with the knots t for each **basis function**, which are also estimated from the data). Function H is defined as:

$$H_{km}(x_{v(k,m)}) = \prod_{k=1}^K h_{km}$$

where $xv(k,m)$ is the predictor in the k 'th of the m 'th product. For order of interactions $K=1$, the model is additive and for $K=2$ the model pairwise interactive.

During forward stepwise, a number of basis functions are added to the model according to a pre-determined maximum which should be considerably larger (twice as much at least) than the optimal (best least-squares fit).

After implementing the forward stepwise selection of basis functions, a backward procedure is applied in which the model is pruned by removing those basis functions that are associated with the smallest increase in the (least squares) goodness-of-fit. A least squares error function (inverse of goodness-of-fit) is computed. The so-called Generalized Cross Validation error is a measure of the goodness of fit that takes into account not only the residual error but also the model complexity as well. It is given by

$$GCV = \frac{\sum_{i=1}^N (y_i - f(x_i))^2}{\left(1 - \frac{C}{N}\right)^2}$$

with

$$C = 1 + cd$$

where N is the number of cases in the data set, d is the effective degrees of freedom, which is equal to the number of independent basis functions. The quantity c is the penalty for adding a basis function. Experiments have shown that the best value for C can be found somewhere in the range $2 < d < 3$ (see Hastie et al., 2001).

[To index](#)