

Package ‘gam’

October 21, 2009

Title Generalized Additive Models

Date 2008-05-28

Version 1.01

Author Trevor Hastie

Description Functions for fitting and working with generalized additive models, as described in chapter 7 of “Statistical Models in S” (Chambers and Hastie (eds), 1991), and “Generalized Additive Models” (Hastie and Tibshirani, 1990).

Maintainer Trevor Hastie <hastie@stanford.edu>

Depends R (>= 2.0), stats, splines, akima

License GPL-2

Repository CRAN

Date/Publication 2009-10-21 06:42:51

R topics documented:

anova.gam	2
gam	3
gam.control	7
gam.data	8
gam.exact	8
kyphosis	10
lo	10
na.gam.replace	12
plot.gam	13
predict.gam	15
s	17
step.gam	18

Index	21
--------------	-----------

Description

Produces an ANODEV table for a set of GAM models, or else a summary for a single GAM model

Usage

```
## S3 method for class 'gam':
anova(object, ..., test)
## S3 method for class 'gam':
summary(object, dispersion=NULL, ...)
```

Arguments

object	a fitted gam
...	other fitted gams for anova
test	a character string specifying the test statistic to be used. Can be one of "F", "Chisq" or "Cp", with partial matching allowed, or 'NULL' for no test.
dispersion	a dispersion parameter to be used in computing standard errors

Details

These are methods for the functions `anova` or `summary` for objects inheriting from class 'gam'. See [anova](#) for the general behavior of this function and for the interpretation of 'test'.

When called with a single 'gam' object, a special anova table for 'gam' models is returned. This gives a breakdown of the degrees of freedom for all the terms in the model, separating the projection part and nonparametric part of each. For example, a term specified by 's()' is broken down into a single degree of freedom for its linear component, and the remainder for the nonparametric component. In addition, a type of score test is performed for each of the nonparametric terms. The nonparametric component is set to zero, and the linear part is updated, holding the other nonparametric terms fixed. This is done efficiently and simultaneously for all terms.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992).

References

- Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
- Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
- Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. New York: Springer.

Examples

```
data(gam.data)
gam.object <- gam(y~s(x,6)+z,data=gam.data)
anova(gam.object)
gam.object2 <- update(gam.object, ~.-z)
anova(gam.object, gam.object2, test="Chisq")
```

gam

*Fitting Generalized Additive Models***Description**

`gam` is used to fit generalized additive models, specified by giving a symbolic description of the additive predictor and a description of the error distribution. `gam` uses the *backfitting algorithm* to combine different smoothing or fitting methods. The methods currently supported are local regression and smoothing splines.

Usage

```
gam(formula, family = gaussian, data, weights, subset, na.action,
     start, etastart, mustart, control = gam.control(...),
     model=FALSE, method, x=FALSE, y=TRUE, ...)

gam.fit(x, y, smooth.frame, weights = rep(1,nobs), start = NULL,
        etastart = NULL, mustart = NULL, offset = rep(0, nobs), family = gaussian(),
        control = gam.control())
```

Arguments

<code>formula</code>	a formula expression as for other regression models, of the form <code>response ~ predictors</code> . See the documentation of <code>lm</code> and <code>formula</code> for details. Built-in nonparametric smoothing terms are indicated by <code>s</code> for smoothing splines or <code>lo</code> for loess smooth terms. See the documentation for <code>s</code> and <code>lo</code> for their arguments. Additional smoothers can be added by creating the appropriate interface functions. Interactions with nonparametric smooth terms are not fully supported, but will not produce errors; they will simply produce the usual parametric interaction.
<code>family</code>	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See family for details of family functions.)
<code>data</code>	an optional data frame containing the variables in the model. If not found in <code>data</code> , the variables are taken from <code>environment(formula)</code> , typically the environment from which <code>gam</code> is called.
<code>weights</code>	an optional vector of weights to be used in the fitting process.

<code>subset</code>	an optional vector specifying a subset of observations to be used in the fitting process.
<code>na.action</code>	a function which indicates what should happen when the data contain NAs. The default is set by the <code>na.action</code> setting of <code>options</code> , and is <code>na.fail</code> if that is unset. The “factory-fresh” default is <code>na.omit</code> . A special method <code>na.gam.replace</code> allows for mean-imputation of missing values (assumes missing at random), and works gracefully with <code>gam</code>
<code>start</code>	starting values for the parameters in the additive predictor.
<code>etastart</code>	starting values for the additive predictor.
<code>mustart</code>	starting values for the vector of means.
<code>offset</code>	this can be used to specify an <i>a priori</i> known component to be included in the additive predictor during fitting.
<code>control</code>	a list of parameters for controlling the fitting process. See the documentation for <code>gam.control</code> for details. These can also be set as arguments to <code>gam()</code> itself.
<code>model</code>	a logical value indicating whether <i>model frame</i> should be included as a component of the returned value.
<code>method</code>	the method to be used in fitting the parametric part of the model. The default method <code>"glm.fit"</code> uses iteratively reweighted least squares (IWLS). The only current alternative is <code>"model.frame"</code> which returns the model frame and does no fitting.
<code>x, y</code>	For <code>gam</code> : logical values indicating whether the response vector and model matrix used in the fitting process should be returned as components of the returned value. For <code>gam.fit</code> : <code>x</code> is a model matrix of dimension $n * p$, and <code>y</code> is a vector of observations of length n .
<code>smooth.frame</code>	for <code>gam.fit</code> only. This is essentially a subset of the model frame corresponding to the smooth terms, and has the ingredients needed for smoothing each variable in the backfitting algorithm. The elements of this frame are produced by the formula functions <code>lo</code> and <code>s</code> .
<code>...</code>	further arguments passed to or from other methods.

Details

The `gam` model is fit using the local scoring algorithm, which iteratively fits weighted additive models by backfitting. The backfitting algorithm is a Gauss-Seidel method for fitting additive models, by iteratively smoothing partial residuals. The algorithm separates the parametric from the non-parametric part of the fit, and fits the parametric part using weighted linear least squares within the backfitting algorithm. This version of `gam` remains faithful to the philosophy of GAM models as outlined in the references below.

An object `gam.slist` (currently set to `c("lo", "s", "random")`) lists the smoothers supported by `gam`. Corresponding to each of these is a smoothing function `gam.lo`, `gam.s` etc that take particular arguments and produce particular output, custom built to serve as building blocks in the backfitting algorithm. This allows users to add their own smoothing methods. See the documentation for these methods for further information. In addition, the object `gam.wlist` (currently

set to `c("s", "lo")`) lists the smoothers for which efficient backfitters are provided. These are invoked if all the smoothing methods are of one kind (either all "lo" or all "s").

Value

`gam` returns an object of class `gam`, which inherits from both `glm` and `lm`.

Gam objects can be examined by `print`, `summary`, `plot`, and `anova`. Components can be extracted using extractor functions `predict`, `fitted`, `residuals`, `deviance`, `formula`, and `family`. Can be modified using `update`. It has all the components of a `glm` object, with a few more. This also means it can be queried, summarized etc by methods for `glm` and `lm` objects. Other generic functions that have methods for `gam` objects are `step` and `preplot`.

The following components must be included in a legitimate ‘gam’ object. The residuals, fitted values, coefficients and effects should be extracted by the generic functions of the same name, rather than by the "\$" operator. The `family` function returns the entire family object used in the fitting, and `deviance` can be used to extract the deviance of the fit.

`coefficients` the coefficients of the parametric part of the additive predictors, which multiply the columns of the model matrix. The names of the coefficients are the names of the single-degree-of-freedom effects (the columns of the model matrix). If the model is overdetermined there will be missing values in the coefficients corresponding to inestimable coefficients.

`additive.predictors` the additive fit, given by the product of the model matrix and the coefficients, plus the columns of the `$smooth` component.

`fitted.values` the fitted mean values, obtained by transforming the component `additive.predictors` using the inverse link function.

`smooth`, `nl.df`, `nl.chisq`, `var` these four characterize the nonparametric aspect of the fit. `smooth` is a matrix of smooth terms, with a column corresponding to each smooth term in the model; if no smooth terms are in the `gam` model, all these components will be missing. Each column corresponds to the strictly nonparametric part of the term, while the parametric part is obtained from the model matrix. `nl.df` is a vector giving the approximate degrees of freedom for each column of `smooth`. For smoothing splines specified by $s(x)$, the approximate `df` will be the trace of the implicit smoother matrix minus 2. `nl.chisq` is a vector containing a type of score test for the removal of each of the columns of `smooth`. `var` is a matrix like `smooth`, containing the approximate pointwise variances for the columns of `smooth`.

`smooth.frame` This is essentially a subset of the model frame corresponding to the smooth terms, and has the ingredients needed for making predictions from a `gam` object

`residuals` the residuals from the final weighted additive fit; also known as residuals, these are typically not interpretable without rescaling by the weights.

`deviance` up to a constant, minus twice the maximized log-likelihood. Similar to the residual sum of squares. Where sensible, the constant is chosen so that a saturated model has deviance zero.

<code>null.deviance</code>	The deviance for the null model, comparable with <code>deviance</code> . The null model will include the offset, and an intercept if there is one in the model
<code>iter</code>	the number of local scoring iterations used to compute the estimates.
<code>family</code>	a three-element character vector giving the name of the family, the link, and the variance function; mainly for printing purposes.
<code>weights</code>	the <i>working</i> weights, that is the weights in the final iteration of the local scoring fit.
<code>prior.weights</code>	the case weights initially supplied.
<code>df.residual</code>	the residual degrees of freedom.
<code>df.null</code>	the residual degrees of freedom for the null model.

The object will also have the components of a `lm` object: `coefficients`, `residuals`, `fitted.values`, `call`, `terms`, and some others involving the numerical fit. See `lm.object`.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992), and the philosophy in Hastie and Tibshirani (1991). This version of `gam` is adapted from the S version to match the `glm` and `lm` functions in R.

Note that this version of `gam` is different from the function with the same name in the R library `mgcv`, which uses only smoothing splines with a focus on automatic smoothing parameter selection via GCV.

References

Hastie, T. J. (1991) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. New York: Springer.

See Also

[glm](#), [family](#), [lm](#).

Examples

```
data(kyphosis)
gam(Kyphosis ~ s(Age,4) + Number, family = binomial, data=kyphosis,
trace=TRUE)
data(airquality)
gam(Ozone^(1/3) ~ lo(Solar.R) + lo(Wind, Temp), data=airquality, na=na.gam.replace)
gam(Kyphosis ~ poly(Age,2) + s(Start), data=kyphosis, family=binomial, subset=Number>2)
data(gam.data)
gam.object <- gam(y ~ s(x,6) + z, data=gam.data)
summary(gam.object)
plot(gam.object, se=TRUE)
```

```
data(gam.newdata)
predict(gam.object, type="terms", newdata=gam.newdata)
```

gam.control	<i>Auxilliary for controlling GAM fitting</i>
-------------	---

Description

Auxiliary function as user interface for 'gam' fitting. Typically only used when calling 'gam' or 'gam.fit'.

Usage

```
gam.control(epsilon=1e-07, bf.epsilon = 1e-07, maxit=30, bf.maxit = 30, trace=FALSE)
```

Arguments

epsilon	convergence threshold for local scoring iterations
bf.epsilon	convergence threshold for backfitting iterations
maxit	maximum number of local scoring iterations
bf.maxit	maximum number of backfitting iterations
trace	should iteration details be printed while gam is fitting the model.
...	

Value

a list is returned, consisting of the five parameters, conveniently packaged up to supply the control argument to gam. The values for gam.control can be supplied directly in a call to gam; these are then filtered through gam.control inside gam.

References

Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Examples

```
## Not run: gam(formula, family, control = gam.control(bf.maxit=15))
## Not run: gam(formula, family, bf.maxit = 15) # these are equivalent
```

`gam.data`*Simulated dataset for gam*

Description

A simple simulated dataset, used to test out the gam functions

Usage

```
data(gam.data)
data(gam.newdata)
```

Format

A data frame with 100 observations on the following 6 variables.

x a numeric vector - predictor
y a numeric vector - the response
z a numeric vector - noise predictor
f a numeric vector - true function
probf a numeric vector - probability function
ybin a numeric vector - binary response

Details

This dataset is artificial, and is used to test out some of the features of gam.

Examples

```
data(gam.data)
gam(y ~ s(x) + z, data=gam.data)
```

`gam.exact`*A method for gam producing asymptotically exact standard errors for linear estimates*

Description

This function is a "wrapper" for a gam object, and produces exact standard errors for each linear term in the gam call (except for the intercept).

Usage

```
gam.exact(gam.obj)
```

Arguments

gam.obj a gam object

Details

Only standard errors for the linear terms are produced. There is a print method for the gamex class.

Value

A list (of class gamex) containing a table of coefficients and a variance covariance matrix for the linear terms in the formula of the gam call.

Author(s)

Aidan McDermott, Department of Biostatistics, Johns Hopkins University. See <http://ihapss.biostat.jhsph.edu/software/gam.exact/gam.exact.htm> Modified by Trevor Hastie for R

References

[1] Issues in Semiparametric Regression: A Case Study of Time Series Models in Air Pollution and Mortality, Dominici F, McDermott A., Hastie T.J., *JASA*, December 2004, 99(468), 938-948. See <http://www-stat.stanford.edu/~hastie/Papers/dominiciR2.pdf>

Examples

```
set.seed(31)
n <- 200
x <- rnorm(n)
y <- rnorm(n)
a <- rep(1:10, length=n)
b <- rnorm(n)
z <- 1.4 + 2.1*a + 1.2*b + 0.2*sin(x/(3*max(x))) + 0.3*cos(y/(5*max(y))) + 0.5 * rnorm(n)
dat <- data.frame(x,y,a,b,z,testit=b*2)
### Model 1: Basic
gam.o <- gam(z ~ a + b + s(x,3) + s(y,5), data=dat)
coefficients(summary.glm(gam.o))
gam.exact(gam.o)
### Model 2: Poisson
gam.o <- gam(round(abs(z)) ~ a + b + s(x,3) + s(y,5), data=dat, family=poisson)
coefficients(summary.glm(gam.o))
gam.exact(gam.o)
```

`kyphosis`*A classic example dataset for GAMs*

Description

Data on the results of a spinal operation "laminectomy" on children, to correct for a condition called "kyphosis"; see Hastie and Tibshirani (1990) for details

Usage

```
data(kyphosis)
```

Format

A data frame with 81 observations on the following 4 variables.

Kyphosis a response factor with levels `absent` `present`.

Age of child in months, a numeric vector

Number of vertebra involved in the operation, a numeric vector

Start level of the operation, a numeric vector

Source

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

`lo`*Specify a loess fit in a GAM formula*

Description

A symbolic wrapper to indicate a smooth term in a formula argument to `gam`

Usage

```
lo(..., span=0.5, degree=1)
gam.lo(x, y, w, span, degree, ncols, xeval)
```

Arguments

...	the unspecified ... can be a comma-separated list of numeric vectors, numeric matrix, or expressions that evaluate to either of these. If it is a list of vectors, they must all have the same length.
span	the number of observations in a neighborhood. This is the smoothing parameter for a <code>loess</code> fit.
degree	the degree of local polynomial to be fit; currently restricted to be 1 or 2.
x	for <code>gam.lo</code> , the appropriate basis of polynomials generated from the arguments to <code>lo</code> . These are also the variables that receive linear coefficients in the GAM fit.
y	a response variable passed to <code>gam.lo</code> during backfitting
w	weights
ncols	for <code>gam.lo</code> the number of columns in <code>x</code> used as the smoothing inputs to local regression. For example, if <code>degree=2</code> , then <code>x</code> has two columns defining a degree-2 polynomial basis. Both are needed for the parametric part of the fit, but <code>ncol=1</code> telling the local regression routine that the first column is the actually smoothing variable.
xeval	If this argument is present, then <code>gam.lo</code> produces a prediction at <code>xeval</code> .

Details

A smoother in `gam` separates out the parametric part of the fit from the non-parametric part. For local regression, the parametric part of the fit is specified by the particular polynomial being fit locally. The workhorse function `gam.lo` fits the local polynomial, then strips off this parametric part. All the parametric pieces from all the terms in the additive model are fit simultaneously in one operation for each loop of the backfitting algorithm.

Value

`lo` returns a numeric matrix. The simplest case is when there is a single argument to `lo` and `degree=1`; a one-column matrix is returned, consisting of a normalized version of the vector. If `degree=2` in this case, a two-column matrix is returned, consisting of a degree-2 polynomial basis. Similarly, if there are two arguments, or the single argument is a two-column matrix, either a two-column matrix is returned if `degree=1`, or a five-column matrix consisting of powers and products up to degree 2. Any dimensional argument is allowed, but typically one or two vectors are used in practice.

The matrix is endowed with a number of attributes; the matrix itself is used in the construction of the model matrix, while the attributes are needed for the backfitting algorithms `all.wam` (weighted additive model) or `lo.wam` (currently not implemented). Local-linear curve or surface fits reproduce linear responses, while local-quadratic fits reproduce quadratic curves or surfaces. These parts of the `loess` fit are computed exactly together with the other parametric linear parts

Note that `lo` itself does no smoothing; it simply sets things up for `gam`; `gam.lo` does the actual smoothing. of the model.

One important attribute is named `call`. For example, `lo(x)` has a `call` component `gam.lo(data[["lo(x)"]], z, w, span = 0.5, degree = 1, ncols = 1)`. This is an expression that gets evaluated repeatedly in `all.wam` (the backfitting algorithm).

gam.lo returns an object with components

residuals	The residuals from the smooth fit. Note that the smoother removes the parametric part of the fit (using a linear fit with the columns in x), so these residual represent the nonlinear part of the fit.
nl.df	the nonlinear degrees of freedom
var	the pointwise variance for the nonlinear fit

When gam.lo is evaluated with an xeval argument, it returns a matrix of predictions.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992).

References

Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

See Also

[s](#), [bs](#), [ns](#), [poly](#), [loess](#)

Examples

```
y ~ Age + lo(Start, span=.5)
# fit Start using a loess smooth with a span of 0.5.
y ~ lo(Age) + lo(Start, Number)
y ~ lo(Age, 0.5) # the argument name for span is not needed.
```

na.gam.replace *Missing Data Filter for GAMs*

Description

A method for dealing with missing values, friendly to GAM models.

Usage

```
na.gam.replace(frame)
```

Arguments

frame	a model or data frame
-------	-----------------------

Value

a model or data frame is returned, with the missing observations (NAs) replaced. The following rules are used. A factor with missing data is replaced by a new factor with one more level, labelled "NA", which records the missing data. Ordered factors are treated similarly, except the result is an unordered factor. A missing numeric vector has its missing entries replaced by the mean of the non-missing entries. Similarly, a matrix with missing entries has each missing entry replaced by the mean of its column. If `frame` is a model frame, the response variable can be identified, as can the weights (if present). Any rows for which the response or weight is missing are removed entirely from the model frame.

The word "gam" in the name is relevant, because `gam()` makes special use of this filter. All columns of a model frame that were created by a call to `lo()` or `s()` have an attribute names "NAs" if NAs are present in their columns. Despite the replacement by means, these attributes remain on the object, and `gam()` takes appropriate action when smoothing against these columns. See section 7.3.2 in Hastie (1992) for more details.

Author(s)

Trevor Hastie

References

Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

See Also

[na.fail](#), [na.omit](#), [gam](#)

Examples

```
data(airquality)
gam(Ozone^(1/3) ~ lo(Solar.R) + lo(Wind, Temp), data=airquality, na=na.gam.replace)
```

plot.gam

Plot Components of a GAM Object

Description

A plot method for GAM objects, which can be used on GLM and LM objects as well. It focuses on terms (main-effects), and produces a suitable plot for terms of different types

Usage

```
## S3 method for class 'gam':
plot(x, residuals, rugplot, se, scale, ask = FALSE, terms, ...)
## S3 method for class 'gam':
preplot(object, newdata, terms, ...)
```

Arguments

<code>x</code>	a gam object, or a <code>preplot.gam</code> object. The first thing <code>plot.gam()</code> does is check if <code>x</code> has a component called <code>preplot</code> ; if not, it computes one using <code>preplot.gam()</code> . Either way, it is this <code>preplot.gam</code> object that is required for plotting a gam object.
<code>object</code>	same as <code>x</code>
<code>residuals</code>	if TRUE, partial deviance residuals are plotted along with the fitted terms—default is FALSE. If <code>residuals</code> is a vector with the same length as each fitted term in <code>x</code> , then these are taken to be the overall residuals to be used for constructing the partial residuals.
<code>rugplot</code>	if TRUE (the default), a univariate histogram or rugplot is displayed along the base of each plot, showing the occurrence of each 'x'; ties are broken by jittering.
<code>se</code>	if TRUE, upper and lower pointwise twice-standard-error curves are included for each plot. The default is FALSE.
<code>scale</code>	a lower limit for the number of units covered by the limits on the 'y' for each plot. The default is <code>scale=0</code> , in which case each plot uses the range of the functions being plotted to create their <code>ylim</code> . By setting <code>scale</code> to be the maximum value of <code>diff(ylim)</code> for all the plots, then all subsequent plots will be produced in the same vertical units. This is essential for comparing the importance of fitted terms in additive models.
<code>ask</code>	if TRUE, <code>plot.gam()</code> operates in interactive mode.
<code>newdata</code>	if supplied to <code>preplot.gam</code> , the <code>preplot</code> object is based on them rather than the original.
<code>terms</code>	subsets of the terms can be selected
<code>...</code>	

Value

a plot is produced for each of the terms in the object `x`. The function currently knows how to plot all main-effect functions of one or two predictors. So in particular, interactions are not plotted. An appropriate 'x-y' is produced to display each of the terms, adorned with residuals, standard-error curves, and a rugplot, depending on the choice of options. The form of the plot is different, depending on whether the 'x'-value for each plot is numeric, a factor, or a matrix.

When `ask=TRUE`, rather than produce each plot sequentially, `plot.gam()` displays a menu listing all the terms that can be plotted, as well as switches for all the options.

A `preplot.gam` object is a list of precomputed terms. Each such term (also a `preplot.gam` object) is a list with components `x`, `y` and others—the basic ingredients needed for each term plot. These are in turn handed to the specialized plotting function `gplot()`, which has methods for different classes of the leading `x` argument. In particular, a different plot is produced if `x` is numeric, a category or factor, a matrix, or a list. Experienced users can extend this range by creating more `gplot()` methods for other classes. Graphical parameters (see [par](#)) may also be supplied as arguments to this function. This function is a method for the generic function `plot()` for class "gam".

It can be invoked by calling `plot(x)` for an object `x` of the appropriate class, or directly by calling `plot.gam(x)` regardless of the class of the object.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992).

References

Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

See Also

[preplot](#), [predict.gam](#)

Examples

```
data(gam.data)
gam.object <- gam(y ~ s(x, 6) + z, data=gam.data)
plot(gam.object, se=TRUE)
data(gam.newdata)
preplot(gam.object, newdata=gam.newdata)
```

predict.gam

Predict method for GAM fits

Description

Obtains predictions and optionally estimates standard errors of those predictions from a fitted generalized additive model object.

Usage

```
predict.gam(object, newdata, type, dispersion, se.fit = FALSE, na.action, terms,...)
```

Arguments

object	a fitted <code>gam</code> object, or one of its inheritants, such as a <code>glm</code> or <code>lm</code> object.
newdata	a data frame containing the values at which predictions are required. This argument can be missing, in which case predictions are made at the same values used to compute the object. Only those predictors, referred to in the right side of the formula in object need be present by name in <code>newdata</code> .
type	type of predictions, with choices "link" (the default), "response", or "terms". The default produces predictions on the scale of the additive predictors, and with <code>newdata</code> missing, <code>predict</code> is simply an extractor function for this component of a <code>gam</code> object. If "response" is selected, the predictions are on the scale of the response, and are monotone transformations of the additive predictors, using the inverse link function. If <code>type="terms"</code> is selected, a matrix of predictions is produced, one column for each term in the model.

<code>se.fit</code>	if TRUE, pointwise standard errors are computed along with the predictions.
<code>dispersion</code>	the dispersion of the GLM fit to be assumed in computing the standard errors. If omitted, that returned by 'summary' applied to the object is used
<code>terms</code>	if <code>type="terms"</code> , the <code>terms=</code> argument can be used to specify which terms should be included; the default is <code>labels(object)</code> .
<code>na.action</code>	function determining what should be done with missing values in 'newdata'. The default is to predict 'NA'.
...	

Value

a vector or matrix of predictions, or a list consisting of the predictions and their standard errors if `se.fit = TRUE`. If `type="terms"`, a matrix of fitted terms is produced, with one column for each term in the model (or subset of these if the `terms=` argument is used). There is no column for the intercept, if present in the model, and each of the terms is centered so that their average over the original data is zero. The matrix of fitted terms has a "constant" attribute which, when added to the sum of these centered terms, gives the additive predictor. See the documentation of `predict` for more details on the components returned.

When `newdata` are supplied, `predict.gam` simply invokes inheritance and gets `predict.glm` to produce the parametric part of the predictions. For each nonparametric term, `predict.gam` reconstructs the partial residuals and weights from the final iteration of the local scoring algorithm. The appropriate smoother is called for each term, with the appropriate `seval` argument (see [s](#) or [lo](#)), and the prediction for that term is produced.

The standard errors are based on an approximation given in Hastie (1992). Currently `predict.gam` does not produce standard errors for predictions at `newdata`.

Warning: naive use of the generic `predict` can produce incorrect predictions when the `newdata` argument is used, if the formula in `object` involves transformations such as `sqrt(Age - min(Age))`.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992). This version of `predict.gam` is adapted from the S version to match the corresponding `predict` methods for `glm` and `lm` objects in R. The `safe.predict.gam` function in S is no longer required, primarily because a safe prediction method is in place for functions like `ns`, `bs`, and `poly`.

References

- Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
- Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.
- Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. New York: Springer.

See Also

[predict.glm](#), [fitted](#), [expand.grid](#)

Examples

```
data(gam.data)
gam.object <- gam(y ~ s(x,6) + z, data=gam.data)
predict(gam.object) # extract the additive predictors
data(gam.newdata)
predict(gam.object, gam.newdata, type="terms")
```

s

Specify a Smoothing Spline Fit in a GAM Formula

Description

A symbolic wrapper to indicate a smooth term in a formula argument to `gam`

Usage

```
s(x, df=4, spar=1)
gam.s(x, y, w, df, spar, xeval)
```

Arguments

<code>x</code>	the univariate predictor, or expression, that evaluates to a numeric vector.
<code>df</code>	the target equivalent degrees of freedom, used as a smoothing parameter. The real smoothing parameter (<code>spar</code> below) is found such that $df = \text{tr}(S) - 1$, where S is the implicit smoother matrix. Values for <code>df</code> should be greater than 1, with <code>df=1</code> implying a linear fit. If both <code>df</code> and <code>spar</code> are supplied, the former takes precedence.
<code>spar</code>	can be used as smoothing parameter, with values typically in $(0, 1]$. See smooth.spline for more details.
<code>y</code>	a response variable passed to <code>gam.s</code> during backfitting
<code>w</code>	weights
<code>xeval</code>	If this argument is present, then <code>gam.s</code> produces a prediction at <code>xeval</code> .

Value

`s` returns the vector `x`, endowed with a number of attributes. The vector itself is used in the construction of the model matrix, while the attributes are needed for the backfitting algorithms `all.wam` (weighted additive model) or `s.wam` (currently not implemented). Since smoothing splines reproduces linear fits, the linear part will be efficiently computed with the other parametric linear parts of the model.

Note that `s` itself does no smoothing; it simply sets things up for `gam`.

One important attribute is named `call`. For example, `s(x)` has a `call` component `gam.s(data[["s(x)"]], z, w, spar = 1, df = 4)`. This is an expression that gets evaluated repeatedly in `all.wam` (the backfitting algorithm).

`gam.s` returns an object with components

residuals	The residuals from the smooth fit. Note that the smoother removes the parametric part of the fit (using a linear fit in x), so these residual represent the nonlinear part of the fit.
nl.df	the nonlinear degrees of freedom
var	the pointwise variance for the nonlinear fit

When `gam.s` is evaluated with an `xeval` argument, it returns a vector of predictions.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992).

References

Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

See Also

[lo](#), [smooth.spline](#), [bs](#), [ns](#), [poly](#)

Examples

```
# fit Start using a smoothing spline with 4 df.
y ~ Age + s(Start, 4)
# fit log(Start) using a smoothing spline with 5 df.
y ~ Age + s(log(Start), df=5)
```

step.gam

Stepwise model builder for GAM

Description

Builds a GAM model in a step-wise fashion. For each "term" there is an ordered list of alternatives, and the function traverses these in a greedy fashion. Note: this is NOT a method for `step`, which used to be a generic, so must be invoked with the full name.

Usage

```
step.gam(object, scope, scale, direction, trace = TRUE, keep, steps, ...)
```

Arguments

<code>object</code>	An object of class <code>gam</code> or any of its inheritants.
<code>scope</code>	defines the range of models examined in the step-wise search. It is a list of formulas, with each formula corresponding to a term in the model. Each of these formulas specifies a "regimen" of candidate forms in which the particular term may enter the model. For example, a term formula might be <code>~1+ Income + log(Income) + s(Income)</code> This means that <code>Income</code> could either appear not at all, linearly, linearly in its logarithm, or as a smooth function estimated non-parametrically. A <code>1</code> in the formula allows the additional option of leaving the term out of the model entirely. Every term in the model is described by such a term formula, and the final model is built up by selecting a component from each formula. The supplied model <code>object</code> is used as the starting model, and hence there is the requirement that one term from each of the term formulas be present in <code>formula(object)</code> . This also implies that any terms in <code>formula(object)</code> not contained in any of the term formulas will be to be present in every model considered.
<code>scale</code>	an optional argument used in the definition of the AIC statistic used to evaluate models for selection. By default, the scaled Chi-squared statistic for the initial model is used, but if forward selection is to be performed, this is not necessarily a sound choice.
<code>direction</code>	The mode of step-wise search, can be one of "both", "backward", or "forward", with a default of "both". If <code>scope</code> is missing, the default for <code>direction</code> is "backward".
<code>trace</code>	If TRUE, information is printed during the running of <code>step.gam()</code> . This is an encouraging choice in general, since <code>step.gam()</code> can take some time to compute either for large models or when called with an an extensive <code>scope=</code> argument. A simple one line model summary is printed for each model visited in the search, and the selected model is noted at each step.
<code>keep</code>	A filter function whose input is a fitted <code>gam</code> object and the associated "AIC" statistic, and whose output is arbitrary. Typically <code>keep()</code> will select a subset of the components of the object and return them. The default is not to keep anything.
<code>steps</code>	The maximum number of steps to be considered. The default is 1000 (essentially as many as required). It is typically used to stop the process early.
<code>...</code>	Additional arguments to be passed on to <code>keep</code>

Value

The step-wise-selected model is returned, with up to two additional components. There is an "anova" component corresponding to the steps taken in the search, as well as a "keep" component if the `keep=` argument was supplied in the call.

We describe the most general setup, when `direction = "both"`. At any stage there is a current model comprising a single term from each of the term formulas supplied in the `scope=` argument. A series of models is fitted, each corresponding to a formula obtained by moving each of the terms one step up or down in its regimen, relative to the formula of the current model. If the

current value for any term is at either of the extreme ends of its regimen, only one rather than two steps can be considered. So if there are p term formulas, at most $2 * p - 1$ models are considered. A record is kept of all the models ever visited (hence the -1 above), to avoid repetition. Once each of these models has been fit, the "best" model in terms of the AIC statistic is selected and defines the step. The entire process is repeated until either the maximum number of steps has been used, or until the AIC criterion can not be decreased by any of the eligible steps.

Author(s)

Written by Trevor Hastie, following closely the design in the "Generalized Additive Models" chapter (Hastie, 1992) in Chambers and Hastie (1992).

References

Hastie, T. J. (1992) *Generalized additive models*. Chapter 7 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

See Also

[step.glm](#), [gam](#), [dropl](#), [addl](#), [anova.gam](#)

Examples

```
data(gam.data)
gam.object <- gam(y~x+z, data=gam.data)
step.object <-step.gam(gam.object, scope=list("x"=~1+x+s(x,4)+s(x,6)+s(x,12), "z"=~1+z+s(z,4))
```

Index

*Topic **datasets**

gam.data, 8
kyphosis, 10

*Topic **models**

anova.gam, 1
gam, 3
gam.control, 7
gam.exact, 8
lo, 10
na.gam.replace, 12
plot.gam, 13
predict.gam, 15
s, 17
step.gam, 18

*Topic **nonparametric**

anova.gam, 1
gam, 3
gam.control, 7
gam.exact, 8
lo, 10
na.gam.replace, 12
plot.gam, 13
predict.gam, 15
s, 17
step.gam, 18

*Topic **regression**

anova.gam, 1
gam, 3
gam.control, 7
gam.exact, 8
lo, 10
na.gam.replace, 12
plot.gam, 13
predict.gam, 15
s, 17
step.gam, 18

*Topic **smooth**

anova.gam, 1
gam, 3

gam.control, 7
gam.exact, 8
lo, 10
na.gam.replace, 12
plot.gam, 13
predict.gam, 15
s, 17
step.gam, 18

add1, 20
anova, 2
anova.gam, 1, 20

bs, 12, 18

drop1, 20

expand.grid, 16

family, 3, 6
fitted, 16

gam, 3, 13, 20
gam.control, 4, 7
gam.data, 8
gam.exact, 8
gam.lo (lo), 10
gam.newdata (gam.data), 8
gam.s (s), 17
glm, 6, 20

kyphosis, 10

lm, 6
lo, 10, 16, 18
loess, 12

na.fail, 3, 13
na.gam.replace, 3, 12
na.omit, 3, 13
ns, 12, 18

options, 3

par, 14

plot.gam, 13

plot.preplot.gam(*plot.gam*), 13

poly, 12, 18

predict.gam, 15, 15

predict.glm, 16

preplot, 15

preplot.gam(*plot.gam*), 13

s, 12, 16, 17

smooth.spline, 17, 18

step, 20

step.gam, 18

summary.gam(*anova.gam*), 1