Package 'mfp'

October 13, 2022

Version 1.5.2.2

Date 2022-01-19

Title Multivariable Fractional Polynomials

Depends survival

Description Fractional polynomials are used to represent curvature in regression models. A key reference is Royston and Altman, 1994.

License GPL (>= 2)

BugReports https://github.com/georgheinze/mfp/issues/

NeedsCompilation no

Author Georg Heinze [cre],
Gareth Ambler [aut],
Axel Benner [aut]

Maintainer Georg Heinze <georg.heinze@meduniwien.ac.at>

Repository CRAN

Date/Publication 2022-01-20 19:12:42 UTC

R topics documented:

| bodyfat | *percentage of body fat determined by underwater weighing* |

Description

A data frame containing the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men.

Source: Roger W. Johnson (1996), "Fitting Percentage of Body Fat to Simple Body Measurements", Journal of Statistics Education. Original data are from K. Penrose, A. Nelson, and A. Fisher (1985), "Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques" (abstract), Medicine and Science in Sports and Exercise, 17(2), 189.

Data were supplied by Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, who gave permission to freely distribute the data and use them for non-commercial purposes. Note, however, that there seem to be a few errors. For instance, body densities for cases 48, 76, and 96 seem to have one digit in error in the two body fat percentage values. Also note a man (case 42) over 200 pounds in weight who is less than 3 feet tall (the height should presumably be 69.5 inches, not 29.5 inches). Percent body fat estimates are truncated to zero when negative (case 182).

Usage

```
data(bodyfat)
```

Format

This data frame contains the observations of 252 men:

case Case number.

brozek Percent body fat using Brozek's equation: 457/Density - 414.2

siri Percent body fat using Siri's equation: 495/Density - 450

density Density determined from underwater weighing (gm/cm**3).

age Age (years).

weight Weight (lbs).

height Height (inches).

neck Neck circumference (cm).

chest Chest circumference (cm).

abdomen Abdomen circumference (cm) "at the umbilicus and level with the iliac crest".

hip Hip circumference (cm).

thigh Thigh circumference (cm).

knee Knee circumference (cm).

ankle Ankle circumference (cm).

biceps Biceps (extended) circumference (cm).

forearm Forearm circumference (cm).

wrist Wrist circumference (cm) "distal to the styloid processes".

Source

e.g. <http://lib.stat.cmu.edu/datasets/bodyfat>

References

R.W. Johnson (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education [Online]*, **4(1)**.

K.W. Penrose, A.G. Nelson, A.G. Fisher (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine and Science in Sports and Exercise*, **17**, 189.

P. Royston, W. Sauerbrei (2004). Improving the robustness of fractional polynomial models by preliminary covariate transformation. Submitted.

Examples

```
data(bodyfat)
bodyfat$height[bodyfat$case==42] <- 69.5   # apparent error
bodyfat <- bodyfat[-which(bodyfat$case==39),]  # cp. Royston $\amp$ Sauerbrei, 2004
mfp(siri ~ fp(age, df = 4, select = 0.1) + fp(weight, df = 4, select = 0.1)
              + fp(height, df = 4, select = 0.1), family = gaussian, data = bodyfat)
```

---

cox   *Family Objects for Cox Proportional Regression Models*

Description

Family objects provide a convenient way to specify the details of the models used by functions such as 'glm'. See the documentation for 'glm' for details.

Usage

```
cox()
```

---

fp   *Fractional Polynomial Transformation*

Description

This function defines a fractional polynomial object for a quantitative input variable x.

Usage

```
fp(x, df = 4, select = NA, alpha = NA, scale=TRUE)
```

Arguments

x	quantitative input variable.
df	degrees of freedom of the FP model. df=4: FP model with maximum permitted degree m=2 (default), df=2: FP model with maximum permitted degree m=1, df=1: Linear FP model.
select	sets the variable selection level for the input variable.
alpha	sets the FP selection level for the input variable.
scale	use pre-transformation scaling to avoid numerical problems (default=TRUE).

Examples

```
## Not run:
fp(x, df = 4, select = 0.05, scale = FALSE)

## End(Not run)
```

GBSG   *German Breast Cancer Study Group*

Description

A data frame containing the observations from the GBSG study.

Usage

```
data(GBSG)
```

Format

This data frame contains the observations of 686 women:

id patient id 1...686.

htreat hormonal therapy, a factor at two levels 0 (no) and 1 (yes).

age of the patients in years.

menostat menopausal status, a factor at two levels 1 (premenopausal) and 2 (postmenopausal).

tumsize tumor size (in mm).

tumgrad tumor grade, a ordered factor at levels 1 < 2 < 3.

posnodal number of positive nodes.

prm progesterone receptor (in fmol).

esm estrogen receptor (in fmol).

rfst recurrence free survival time (in days).

cens censoring indicator (0 censored, 1 event).

References

M. Schumacher, G. Basert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L.A. Neumann and H.F. Rauschecker for the German Breast Cancer Study Group (1994). Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, **12**, 2086–2093.

W. Sauerbrei and P. Royston (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistics Society Series A*, Volume **162**(1), 71–94.

Examples

```
data(GBSG)
mfp(Surv(rfst, cens) ~ fp(age, df = 2, select = 0.05)
            + fp(prm, df = 4, select = 0.05), family = cox, data = GBSG)
```

---

mfp   *Fit a Multiple Fractional Polynomial Model*

Description

Selects the multiple fractional polynomial (MFP) model which best predicts the outcome. The model may be a generalized linear model or a proportional hazards (Cox) model.

Usage

```
mfp(formula, data, family = gaussian, method = c("efron", "breslow"),
        subset, na.action, init, alpha=0.05, select = 1, maxits = 20,
        keep = NULL, rescale = TRUE, verbose = FALSE, x = TRUE, y = TRUE)
```

Arguments

formula	a formula object, with the response of the left of a ~ operator, and the terms, separated by + operators, on the right. Fractional polynomial terms are indicated by fp. If a Cox PH model is required then the outcome should be specified using the Surv() notation used by coxph.
data	a data frame containing the variables occurring in the formula. If this is missing, the variables should be on the search list.
family	a family object - a list of functions and expressions for defining the link and variance functions, initialization and iterative weights. Families supported are gaussian, binomial, poisson, Gamma, inverse.gaussian and quasi. Additionally Cox models are specified using "cox".
method	a character string specifying the method for tie handling. This argument is used for Cox models only and has no effect for other model families. See 'coxph' for details.

subset	expression saying which subset of the rows of the data should be used in the fit. All observations are included by default.
na.action	function to filter missing data. This is applied to the model.frame after any subset argument has been used. The default (with na.fail) is to create an error if any missing values are found.
init	vector of initial values of the iteration (in Cox models only).
alpha	sets the FP selection level for all predictors. Values for individual predictors may be changed via the fp function in the formula.
select	sets the variable selection level for all predictors. Values for individual predictors may be changed via the fp function in the formula.
maxits	maximum number of iterations for the backfitting stage.
keep	keep one or more variables in the model. The selection level for these variables will be set to 1.
rescale	logical; uses re-scaling to show the parameters for covariates on their original scale (default TRUE).
verbose	logical; run in verbose mode (default FALSE).
x	logical; return the design matrix in the model object?
y	logical; return the response in the model object?

Details

The estimation algorithm processes the predictors in turn. Initially, mfp silently arranges the predictors in order of increasing P-value (i.e. of decreasing statistical significance) for omitting each predictor from the model comprising all the predictors with each term linear. The aim is to model relatively important variables before unimportant ones.

At the initial cycle, the best-fitting FP function for the first predictor is determined, with all the other variables assumed linear. The FP selection procedure is described below. The functional form (but NOT the estimated regression coefficients) for this predictor is kept, and the process is repeated for the other predictors in turn. The first iteration concludes when all the variables have been processed in this way. The next cycle is similar, except that the functional forms from the initial cycle are retained for all variables excepting the one currently being processed.

A variable whose functional form is prespecified to be linear (i.e. to have 1 df) is tested only for exclusion within the above procedure when its nominal P-value (selection level) according to select() is less than 1.

Updating of FP functions and candidate variables continues until the functions and variables included in the overall model do not change (convergence). Convergence is usually achieved within 1-4 cycles.

Model Selection

mfp uses a form of backward elimination. It start from a most complex permitted FP model and attempt to simplify it by reducing the df. The selection algorithm is inspired by the so-called "closed test procedure", a sequence of tests in each of which the "familywise error rate" or P-value is maintained at a prespecified nominal value such as 0.05.

The "closed test" algorithm for choosing an FP model with maximum permitted degree m=2 (4 df) for a single continuous predictor, x, is as follows:

1. Inclusion: test the FP in x for possible omission of x (4 df test, significance level determined by select). If x is significant, continue, otherwise drop x from the model.

2. Non-linearity: test the FP in x against a straight line in x (3 df test, significance level determined by alpha). If significant, continue, otherwise the chosen model is a straight line.

3. Simplification: test the FP with m=2 (4 df) against the best FP with m=1 (2 df) (2 df test at alpha level). If significant, choose m=2, otherwise choose m=1.

All significance tests are carried out using an approximate P-value calculation based on a difference in deviances (-2 x log likelihood) having a chi-squared or F distribution, depending on the regression in use. Therefore, each of the tests in the procedure maintains a significance level only approximately equal to select. The algorithm is thus not truly a closed procedure. However, for a given significance level it does provide some protection against over-fitting, that is against choosing over-complex MFP models.

Value

an object of class `mfp` is returned which either inherits from both glm and lm or coxph.

Side Effects

details are produced on the screen regarding the progress of the backfitting routine. At completion of the algorithm a table is displayed showing the final powers selected for each variable along with other details.

Known Bugs

glm models should not be specified without an intercept term as the software does not yet allow for that possibility.

Author(s)

Gareth Ambler and Axel Benner

References

Ambler G, Royston P (2001) Fractional polynomial model selection procedures: investigation of Type I error rate. *Journal of Statistical Simulation and Computation* 69: 89–108.

Benner A (2005) mfp: Multivariable fractional polynomials. *R News* 5(2): 20–23.

Royston P, Altman D (1994) Regression using fractional polynomials of continuous covariates. *Appl Stat.* 3: 429–467.

Sauerbrei W, Royston P (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society (Series A)* 162: 71–94.

See Also

mfp.object, fp, glm

Examples

```
data(GBSG)
f <- mfp(Surv(rfst, cens) ~ fp(age, df = 4, select = 0.05)
                  + fp(prm, df = 4, select = 0.05), family = cox, data = GBSG)
print(f)
survfit(f$fit)  # use proposed coxph model fit for survival curve estimation
```

---

mfp.object	*Multiple Fractional Polynomial Model Object*

Description

Objects returned by fitting fractional polynomial model objects.

These are objects representing fitted `mfp` models. Class `mfp` inherits from either glm or coxph depending on the type of model fitted.

Value

In addition to the standard glm/coxph components the following components are included in a `mfp` object.

x	the final FP transformations that are contained in the design matrix x. The predictor "z" with 4 df would have corresponding columns "z.1" and "z.2" in x.
powers	a matrix containing the best FP powers for each predictor. If a predictor has less than two powers a NA will fill the appropriate cell of the matrix.
pvalues	a matrix containing the P-values from the closed tests. Briefly p.null is the P-value for the test of inclusion (see mfp), p.lin corresponds to the test of non-linearity and p.FP the test of simplification. The best m=1 power (power2) and best m=2 powers (power4.1 and power4.2) are also given.
scale	all predictors are shifted and rescaled before being power transformed if non-positive values are encountered or the range of the predictor is reasonably large. If x' would be used instead of x where x' = (x+a)/b the parameters a (shift) and b (scale) are contained in the matrix scale.
df.initial	a vector containing the degrees of freedom allocated to each predictor.
df.final	a vector containing the degrees of freedom of each predictor at convergence of the backfitting algorithm.
dev	the deviance of the final model.
dev.lin	the deviance of the model that has every predictor included with 1 df (i.e. linear).
dev.null	the deviance of the null model.
fptable	the table of the final fp transformations.
formula	the proposed formula for a call of glm/coxph.
fit	the fitted glm/coxph model using the proposed formula. This component can be used for prediction, etc.

See Also

mfp, glm.object

Index