1 Introduction

The \texttt{mfp} package is a collection of R \cite{7} functions targeted at the use of fractional polynomials (FP) for modelling the influence of continuous covariates on the outcome in regression models, as introduced by Royston \& Altman (1994) \cite{8} and modified by Sauerbrei \& Royston (1999) \cite{9}. The model may include binary, categorical or further continuous covariates which are included in the variable selection process but without need of FP transformation. It combines backward elimination with a systematic search for a 'suitable' transformation to represent the influence of each continuous covariate on the outcome. An application of multivariable fractional polynomials (MFP) in modelling prognostic and diagnostic factors in breast cancer is given by \cite{9}. The stability of the models selected is investigated in \cite{10}. Briefly, fractional polynomials models are useful when one wishes to preserve the continuous nature of the covariates in a regression model, but suspects that some or all of the relationships may be non-linear. At each step of a 'backfitting' algorithm MFP constructs a fractional polynomial transformation for each continuous covariate while fixing the current functional forms of the other covariates. The algorithm terminates when no more covariate is excluded and the functional forms of the continuous covariates do not change anymore.

2 Inventory of functions

\texttt{mfp.object} is an object representing a fitted \texttt{mfp} model. Class \texttt{mfp} inherits from either \texttt{glm} or \texttt{coxph} depending on the type of model fitted. In addition to the standard \texttt{glm/coxph} components the following components are included in an \texttt{mfp} object

\texttt{x} the final FP transformations that are contained in the design matrix \texttt{x}. The covariate "z" with 4 df (second-degree FP) has corresponding columns "z.1" and "z.2" in \texttt{x}. A first-degree FP covariate "z" would have one column "z.1".
powers a matrix containing the best FP powers for each covariate. If a covariate has less than two
powers NAs will fill the appropriate cell of the matrix.

pvalues a matrix containing the P-values from the "closed test procedure" together with the best
powers chosen. Briefly p.null is the P-value for the test of inclusion (see mfp), p.lin corresponds
to the test of nonlinearity and p.FP the test of simplification by comparing first degree (FP1)
and second degree (FP2) transformations. The best first degree FP power (power2) and best
second degree FP powers (power4.1 and power4.2) are also given. The numbers 2 and 4 describe
the corresponding degrees of freedom.

scale all covariates are shifted and rescaled before being power transformed if nonpositive values are
encountered or the range of values of the covariates 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 covariate corresponding to the
degree of FP (m=4 for second degree FP, m=2 for first degree FP).

df.final a vector containing the degrees of freedom of each covariate at convergence of the backfitting
algorithm (m=4 for second degree FP, m=2 for first degree FP, m=1 for untransformed variable,
m=0 if covariate was excluded).

dev the deviance of the final model.

dev.lin the deviance of the model that uses the linear predictor of untransformed covariates.

dev.null the deviance of the null model.

fptable the table of the final fp transformations.

fit a call of the corresponding glm or cox model using the selected and (possibly) FP transformed
variables of the final model.

3 Usage in R

Start with

>library(mfp)

An mfp.object will be created by application of function mfp.
A typical call of an mfp model has the form response ~ terms where response is the (numeric)
response vector and terms is a series of terms, separated by + operators, which specifies a linear
predictor for response provided by the formula argument of the function call.

>str(mfp)

function (formula = formula(data), data = parent.frame(), family = gaussian,
method = c("efron", "breslow"), subset = NULL, na.action = na.omit,
init = NULL, alpha = 0.05, select = 1, maxits = 20, keep = NULL, rescale = TRUE,
verbose = FALSE, x = TRUE, y = TRUE)
Fractional polynomial terms are indicated by \texttt{fp}.

For \texttt{binomial} models the response can also be specified as a \texttt{factor}. If a Cox proportional hazards model is required then the outcome need to be specified using the \texttt{Surv()} notation.

The argument \texttt{family} describes 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.

Argument \texttt{alpha} sets the global FP selection level for all covariates. Different selection levels for individual covariates can be chosen by using the \texttt{fp} function. The variable selection level for all covariates is set by \texttt{select}. Values for individual fractional polynomials may be set using the \texttt{fp} function.

The function \texttt{fp} defines a fractional polynomial object for a single input variable.

\begin{verbatim}
>str(fp)
function (x, df = 4, select = NA, alpha = NA, scale = TRUE)
\end{verbatim}

In addition to \texttt{alpha} and \texttt{select} the \texttt{scale} argument of the \texttt{fp} function denotes the use of pre-transformation scaling to avoid possible numerical problems or for variables with non-positive values.

### 3.1 Model selection

The original Stata implementation of \texttt{mfp} uses two different selection procedures for a single continuous covariate \(x\), a sequential selection procedure and a closed testing selection procedure (\texttt{ra2}, [1]). In the R implementation only the \texttt{ra2} algorithm is used which is also the default in the Stata and SAS implementations of \texttt{mfp}.

The \texttt{ra2} algorithm is described in [1] and [7]. It uses a closed test procedure [2] which maintains approximately the correct Type I error rate for each component test. The procedure allows the complexity of candidate models to increase progressively from a prespecified minimum (a null model) to a prespecified maximum (an FP) according to an ordered sequence of test results.

The algorithm works as follows:

1. Perform a 4 df test at the \(\alpha\) level of the best-fitting second-degree FP against the null model. If the test is not significant, drop \(x\) and stop, otherwise continue.

2. Perform a 3 df test at the \(\alpha\) level of the best-fitting second-degree FP against a straight line. If the test is not significant, stop (the final model is a straight line), otherwise continue.

3. Perform a 2 df test at the \(\alpha\) level of the best-fitting second-degree FP against the best-fitting first-degree FP. If the test is significant, the final model is the FP with \(m = 2\), otherwise the FP with \(m = 1\).

The tests in step 1, 2 and 3 are of overall association, non-linearity and between a simpler or more complex FP model, respectively.

### 4 Example

#### 4.1 Cox proportional hazards model

We use the dataset \texttt{GBSG} which contains data from a study of the German Breast Cancer Study Group for patients with node-positive breast cancer.

\begin{verbatim}
>data(GBSG)
>str(GBSG)
\end{verbatim}
The response variable is recurrence free survival time ($\text{Surv(rfst, cens)}$). Complete data on 7 prognostic factors is available for 686 patients. The median follow-up was about 5 years, 299 events were observed for recurrence free survival time. We use a Cox proportional hazards regression to model the hazard of recurrence by the 7 prognostic factors of which 5 are continuous, age of the patients in years ($\text{age}$), tumor size in mm ($\text{tumsize}$), number of positive lymphnodes ($\text{posnodal}$), progesterone receptor in fmol ($\text{prm}$), estrogen receptor in fmol ($\text{esm}$); one is binary, menopausal status ($\text{menostat}$); and one is ordered categorical with three levels, tumor grade ($\text{tumgrad}$). The additional variable $\text{htreat}$ describes if a hormonal therapy was applied and is used as stratification variable.

We use $\text{mfp}$ to build a model from the initial set of 7 covariates using the backfitting model selection algorithm. We set the global variable selection level to 0.05 and use the default FP selection level. By using $\text{fp()}$ in the model formula a fractional polynomial transformation with possibly pre-transformation scaling is estimated. This is done here for $\text{tumsize}$, $\text{posnodal}$, $\text{prm}$, and $\text{esm}$. Otherwise a linear form of the unscaled variable is used, as for $\text{age}$. Categorical factors are included without transformation.

Hormonal therapy ($\text{htreat}$) was used as stratification variable.

By $\text{verbose=TRUE}$ the process of FP and variable selection is printed.

```
> f <- mfp(Surv(rfst, cens) ~ strata(htreat)+age+fp(tumsize)+fp(posnodal)+fp(prm)+fp(esm) + +menostat+tumgrad, family = cox, data = GBSG, select=0.05, verbose=TRUE)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Deviance</th>
<th>Power(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>posnodal</td>
<td>3135.218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3103.245</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3081.123</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3074.213</td>
<td>0.5 3</td>
</tr>
<tr>
<td>prm</td>
<td>3095.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3074.213</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3067.746</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>3066.502</td>
<td>-2 0.5</td>
</tr>
<tr>
<td>tumgrad2</td>
<td>3081.253</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3074.213</td>
<td>1</td>
</tr>
</tbody>
</table>

4
<table>
<thead>
<tr>
<th>Variable</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>tumgrad3</td>
<td>3082.613</td>
<td>3074.213</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tumsize</td>
<td>3075.813</td>
<td>3074.213</td>
<td>1</td>
<td>-1</td>
<td>-1 3</td>
</tr>
<tr>
<td>menostat2</td>
<td>3076.922</td>
<td>3075.813</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>3076.922</td>
<td>3076.922</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>esm</td>
<td>3077.795</td>
<td>3076.922</td>
<td>1</td>
<td>3</td>
<td>-0.5 3</td>
</tr>
<tr>
<td>Cycle 2</td>
<td>posnodal</td>
<td>3152.737</td>
<td>3108.965</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3085.051</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3077.795</td>
<td>0.5 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prm</td>
<td>3099.562</td>
<td>3077.795</td>
<td>1</td>
<td>0.5</td>
<td>0 0.5</td>
</tr>
<tr>
<td>tumgrad2</td>
<td>3085.024</td>
<td>3077.795</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>df.initial</td>
<td>select</td>
<td>alpha</td>
<td>df.final</td>
<td>power1</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>--------</td>
<td>-------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>tumgrad3</td>
<td>3086.686</td>
<td>3077.795</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tumsize</td>
<td>3077.795</td>
<td>3064.711</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>menostat2</td>
<td>3077.795</td>
<td>3066.973</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>3077.795</td>
<td>3077.737</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Transformation

<table>
<thead>
<tr>
<th>Variable</th>
<th>shift</th>
<th>scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>posnodal</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>prm</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>tumgrad2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tumgrad3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tumsize</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>menostat2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>age</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>esm</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

### Fractional Polynomials

<table>
<thead>
<tr>
<th>Variable</th>
<th>df.initial</th>
<th>select</th>
<th>alpha</th>
<th>df.final</th>
<th>power1</th>
<th>power2</th>
</tr>
</thead>
<tbody>
<tr>
<td>posnodal</td>
<td>4</td>
<td>0.05</td>
<td>0.05</td>
<td>4</td>
<td>0.5</td>
<td>3</td>
</tr>
<tr>
<td>prm</td>
<td>4</td>
<td>0.05</td>
<td>0.05</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>tumgrad2</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>tumgrad3</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>tumsize</td>
<td>4</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>menostat2</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>esm</td>
<td>4</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Transformations of covariates:

```
formula
age <NA>
tumsize <NA>
posnodal I((posnodal/10) - 0.5) + I((posnodal/10)^3)
prm I(((prm + 1)/100)^1)
esm <NA>
menostat <NA>
tumgrad tumgrad
```

Deviance table:

<table>
<thead>
<tr>
<th>Model</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model</td>
<td>3198.026</td>
</tr>
<tr>
<td>Linear model</td>
<td>3103.245</td>
</tr>
<tr>
<td>Final model</td>
<td>3077.795</td>
</tr>
</tbody>
</table>

After two cycles the final model is selected. Of the possible input variables tumor size (tumsize), menopausal status (menostat), age and estrogen receptor (esm) were excluded from the model. Only for variable posnodal a nonlinear transformation was chosen. Prescaling was used for esm, prm and tumsize.

Details of the model fit are given by `summary`.

```R
> summary(f)
```

Call:
coxph(formula = Surv(rfst, cens) ~ strata(htreat) + I((posnodal/10)^0.5) + I((posnodal/10)^3) + I(((prm + 1)/100)^1) + tumgrad, data = GBSG)

n= 686, number of events= 299

```
coef exp(coef) se(coef)      z  Pr(>|z|)
I((posnodal/10)^0.5) 1.79086   5.99461  0.21339  8.392 < 2e-16 ***
I((posnodal/10)^3)  -0.03251   0.96801  0.01334 -2.438  0.01478 *
I(((prm + 1)/100)^1) -0.21321   0.80799  0.05381 -3.963 7.41e-05 ***
tumgrad2 0.61624   1.85195  0.24877  2.477  0.01324 *
tumgrad3 0.74897   2.11482  0.26798  2.795  0.00519 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef) exp(-coef) lower .95 upper .95
I((posnodal/10)^0.5)  5.995   0.1668   3.9457   9.1075
I((posnodal/10)^3) 0.968    1.0330   0.9430   0.9936
I(((prm + 1)/100)^1) 0.808    1.2376   0.7271   0.8978
tumgrad2  1.852    0.5400   1.1373   3.0157
tumgrad3  2.115    0.4729   1.2507   3.5758
```

Concordance= 0.679  (se = 0.024 )
$R^2 = 0.161$ (max possible = 0.991)

Likelihood ratio test = 120.2 on 5 df, $p=0$

Wald test = 116.4 on 5 df, $p=0$

Score (logrank) test = 122.9 on 5 df, $p=0$

Details of the FP transformations are given in the \texttt{fptable} value of the resulting \texttt{mfp.object}.

\begin{verbatim}
> f$fptable
                  df.initial select alpha df.final power1 power2
    posnodal       4 0.05 0.05       4   0.5   3
     prm           4 0.05 0.05       1   1    .
   tumgrad2       1 0.05 0.05       1   1    .
   tumgrad3       1 0.05 0.05       1   1    .
     tumsize      4 0.05 0.05       0    .    .
menostat2       1 0.05 0.05       0    .    .
       age        1 0.05 0.05       0    .    .
       esm        4 0.05 0.05       0    .    .
\end{verbatim}

The final model uses a second degree fractional polynomial for the number of positive lymphnodes with powers 0.5 and 3.

The value \texttt{fit} of the resulting \texttt{mfp} object can be used for survival curve estimation of the final model \texttt{fit} (1).

**References**


Figure 1: Predicted survival curves of the final mfp model for the two strata defined by hormonal treatment (red line = no hormonal treatment, green line = hormonal treatment).