Multivariable Fractional Polynomials with Extensions

Edwin Kipruto Michael Kammer Patrick Royston Willi Sauerbrei

June 24, 2023

Contents

1 Introduction to Multivariable Fractional Polynomial(MFP)

1.1 Overview of MFP

Multivariable regression models are widely used across various fields of science where empirical data is analyzed. In model building, many researchers often assume a linear function for continuous variables, sometimes after applying "standard" transformations such as logarithmic, or divide the variable into several categories. However, assuming linearity without considering non-linear relationships may hinder the detection of stronger effects or even cause the effects to be mis-modeled. Categorizing continuous variables, which results in modeling implausible step functions, is a common practice but widely criticized (Sauerbrei et al. 2023; Royston et al. 2006).

When building a descriptive model with the aim of identifying predictors of an outcome and understanding the the relationship between the predictors and the outcome, two components are often considered: variable selection, which involves identifying the subset of "important" predictors, and identification of possible non-linearity in continuous predictors.

The MFP approach has been proposed as a pragmatic method for dealing with non-linearity in multivariable model-building. This approach retains continuous predictors as continuous, identifies non-linear functions if sufficiently supported by the data, and eliminates weakly influential predictors using backward elimination (BE). Despite its simplicity and ease of understanding for researchers familiar with regression models, the selected models often capture the essential information from the data. The MFP models are relatively straightforward to interpret and report, which is essential for their transportability and practical applicability. In summary, the MFP procedure combines:

- variable selection through backward elimination (BE) with
- Selection of fractional polynomial (FP) functions for continuous variables

The analyst must decide on a nominal significance level (α) for both components. The choice of these two significance levels has a strong influence on the complexity of the final model. While it is possible to use the same α level for both components, they can also differ. The decision regarding these significance levels heavily depends on the specific aim of the analysis.

Section 1.2 provides an overview of fractional polynomial functions for a single continuous variable in the model, including the function selection procedure (FSP). In Section 1.3, the MFP approach is described, focusing on models that involve two or more variables. Section 2 is an introduction to our package which covers the installation process and provides instructions for utilizing it in various linear regression models. Section 3 presents a comparison between our package and other R packages that implement MFP. Section 4 introduces an extension of MFP using the approximate cumulative distribution (ACD) transformation of a continuous covariate. This extension allows for modeling a sigmoid relationship between covariates and an outcome variable. Subsection 4.1 describes the function selection procedure with ACD transformation (FSPA), while section 4.2 offers a guide on how to implement MFPA using our package. Lastly, Section 5 describes an additional extension of MFP that is not currently implemented in our package but is available in the STATA software.

For more comprehensive information about MFP and its extensions, please visit our website at https://mfp.imbi.uni-freiburg.de/.

1.2 Fractional polynomial models for a continuous variable

Suppose that we have an outcome variable, a single continuous covariate x, and a regression model relating them. A starting point is the straight-line model, $\beta_1 x$ (for simplicity, we suppress the constant term, β_0). Often, a straight line is an adequate description of the relationship, but other models should be investigated

for possible improvements in fit. A simple extension of the straight line is a power transformation model, $\beta_1 x^p$. The latter model has often been used by practitioners in an ad hoc way, utilizing different choices of p. Royston and Altman (1994) formalized the model by calling it a first-degree fractional polynomial or FP1 function. The power p is chosen from a pragmatically restricted set of eight elements: $S = \{-2, 1, 0.5, 0, 0.5, 1, 2, 3\}$, where x^0 denotes natural logarithm of x, $log(x)$.

As with polynomial regression, extension from one-term FP1 functions to more complex and flexible two-term FP2 functions is straightforward. The quadratic function $\beta_1 x^1 + \beta_2 x^2$ is written as $\beta_1 x^{p1} + \beta_2 x^{p2}$ in FP terminology. The powers $p1 = 1$ and $p2 = 2$ are members of S. Royston and Altman extended the class of FP2 functions with different powers to cases with equal powers ($p1 = p2 = p$) by defining them as $\beta_1 x^p + \beta_2 x^p log(x)$. These are known as repeated-powers functions. Detailed definition of FP functions or models is given in Section 4.3.1 of Royston and Sauerbrei (2008). For formal definitions, we use notation from Royston and Sauerbrei (2008). Throughout the rest of this article, we use abbreviations (R&S, year) and (S&R, year) for papers published by these two author.

FP1 functions are always monotonic and those with power $p < 0$ have an asymptote as $x \to \infty$. FP2 functions may be monotonic or unimodal (i.e., have one maximum or one minimum for some positive values of x), and they have an asymptote as $x \to \infty$ when both $p1$ and $p2$ are negative. For more details, see R&S (2008), Section 4.4. Figure 1 shows FP1 and some FP2 curves. The subset of FP2 powers is chosen to illustrate the flexibility available with a few pairs of powers $(p1, p2)$.

In total, there are 44 models available within the set of FP powers (S), consisting of 8 FP1 models and 36 FP2 models. Although the allowed class of FP functions may seem limited, it encompasses a wide range of diverse shapes. This is illustrated in Figure 1, with the left panel displaying eight FP1 powers and the middle panel depicting a subset of FP2 powers. The right panel demonstrates the variations in shape for a fixed FP2 power (-2, 2) but different regression coefficients.

Based on extensive experience with real data and several simulation studies, FP1 and FP2 are generally considered adequate in the context of multivariable model building, particularly when variable selection and functional forms is required. The content of this article has been previously published in two encyclopedia articles (Sauerbrei and Royston, 2011; Sauerbrei and Royston, 2016).

1.2.1 Function selection procedure (FSP)

Choosing the best FP1 or FP2 function by grid search, minimizing the deviance (minus twice the maximized log-likelihood), is straightforward. However, having a suitable default function is important for increasing the parsimony, stability, and general usefulness of selected functions. In most of the algorithms implementing fractional polynomial (FP) modeling, the default function is linear-arguably, a natural choice. Therefore, unless the data support a more complex FP function, a straight line model is chosen.

There are occasional exceptions; for example, in modeling time-varying regression coefficients in the Cox model, Sauerbrei et al. (2007) chose a default time (t) transformation of $(log(t))$ rather than (t). It can be assumed that deviance difference between an FPm and an FP \(m-1\) model is distributed approximately as central χ^2 on 2 degrees of freedom (d.f.) (R&S 2008, Chapter 4.9; Ambler and Royston (2001). To select a specific function, a closed test procedure (other procedures had been proposed before) was proposed (R&S 2008, Section 4.10). The complexity of the finally chosen function is predicated on preliminary decisions as to the nominal significance level (α) and the degree (m) of the most complex FP model allowed. Typical choices are $\alpha = 0.05$ and FP2 (m = 2). We illustrate the strategy for m = 2, which runs as follows:

1. Test the best FP2 model for x at the (α) significance level against the null model using 4 d.f. If the test is not significant, stop and conclude that the effect of x is "not significant" at the α level. Otherwise continue.
2. Test the best FP2 for x against a straight line at the α level using 3 d.f. If the test is not significant, stop, the final model being a straight line. Otherwise continue.
3. Test the best FP2 for x against the best FP1 at the α level using 2 d.f. If the test is not significant, the final model is FP1, otherwise, the final model is FP2. This marks the end of procedure.

The test at step 1 is of overall association of the outcome with x. The test at step 2 examines the evidence for nonlinearity. The test at step 3 chooses between a simpler or more complex nonlinear model.

1.3 Multivariable fractional polynomial (MFP) procedure

When developing a multivariable model with a relatively large number of candidate covariates (say 20, we are not envisaging the case of high-dimensional data), an important distinction is between descriptive, predictive and explanatory modelling (Shmueli, 2010). MFP was mainly developed for descriptive modelling, aiming to capture the data structure parsimoniously. Nevertheless, a suitable descriptive model often has a fit similar to a model whose aim is good prediction. In some fields, the term explanatory modelling is used exclusively for testing causal theory. Unlike developing a predictive model based on acceptable statistical criteria, developing a model suitable for description is much more challenging (Sauerbrei et al., 2015).

In many areas of science, the main interest often lies in the identification of influential variables and determination of appropriate functional forms for continuous variables. Often, linearity is presumed without checking this important assumption, and much better-fitting nonlinear functions may not be considered. The MFP procedure was proposed as a pragmatic strategy to investigate whether nonlinear functions can improve the model fit (R&S, 2008; S&R, 1999). MFP combines backward elimination (BE) for the selection of variables with a systematic search for possible nonlinearity by the function selection procedure (FSP). The extension is feasible with any type of regression model to which BE is applicable. When developing models for description, it is important to consider factors such as model stability, generalizability, and practical usefulness. The philosophy behind MFP modeling is to create interpretable and relatively simple models (Sauerbrei et al 2007). Consequently, an analyst using MFP should be less concerned about failing to include variables with a weak effect or failing to identify minor curvature in a functional form of a continuous covariate. Modifications that may improve MFP models are combination with post-estimation shrinkage (Dunkler et al., 2016, R-package shrink) and a more systematic check for overlooked local features (Binder and Sauerbrei, 2010, currently not implemented in our package).Successful use of MFP requires only general knowledge about building regression models.

Two nominal significance level values are the main tuning parameters: α_1 for selecting variables with BE (in the first step of the FSP) and α_2 for comparing the fit of functions within the FSP. Often, $\alpha_1 = \alpha_2$ is a good choice. If available, subject-matter knowledge should replace or at least guide data-dependent model choice. Only minor modifications are required to incorporate various types of subject-matter knowledge into MFP modeling. For a detailed example, see S&R (1999). Recommendations for practitioners of MFP modeling are given in Sauerbrei et al. (2007b) and in R&S (2008, Section 12.2).

1.3.1 MFP – Key Issues and Approaches to Handling Them

Mainly focusing on the FP component, we briefly mention key issues of MFP modeling and refer to the literature for further reading. Regarding variable selection, we have summarized relevant issues and provided arguments for backward elimination as our preferred strategy (R&S 2008, Chapter 2). Even when a search for model improvement using a nonlinear function is not considered, that is, all functions are assumed linear, it is infeasible to derive a suitable and stable model for description in small datasets. Below we provide some information about sample size needed, but implicitly we assume that the sample size is "sufficient".

1.3.1.1 The variable has to be positive The class of FP1 and FP2 functions includes a log and other transformations which require that the continuous variable must be positive. A preliminary origin-shift transformation can be applied (R&S 2008, Chapters 4.7 and 11). For variables with a "spike" of probability mass at zero, a binary indicator variable may be added to the model and the FSP may be modified accordingly (Royston et al., 2010; Becher et al., 2012; Lorenz et al., 2018).

1.3.1.2 Sample size, influential observations and replicability of MFP models

All statistical models are potentially adversely affected by influential observations or "outliers". However, compared with models that comprise only linear functions, the situation may be more critical for FP functions because logarithmic or negative power transformations may produce extreme functional estimates at small values of x. Conversely, the same may happen with large positive powers at large values of x. Such transformations may create influential observations that may affect parts of the FSP. To mitigate the impact of influential observations, it is important to assess the robustness of FP functions. Some suggestions for investigating influential points (IPs) and handling such issues in MFP modeling can be found in R&S (2008, Chapters 5 and 10) and their paper on improving the robustness of FP models (R&S, 2007). Using synthetic data, a more detailed investigation of IPs is given in Sauerbrei et al. (2023). The authors conclude that for smaller sample sizes, IPs and low power are important reasons that the MFP approach may not be able to identify underlying functional relationships for continuous variables and selected models might differ substantially from the true model. However, for larger sample sizes (about 50 or more observations per variable) a carefully conducted MFP analysis is often a suitable way to select a multivariable regression model which includes continuous variables. Figure 2 illustrates the effects of influential observation 151 on the art dataset. The variable of interest is x5 with extremely large values, and the outcome is a continuous variable. The inclusion of observation 151 in the data results in the selection of the FP2 function, while its elimination leads to the selection of the FP1 function. Therefore, it is sufficient to describe variable x5 using a simpler FP1 function rather than a complex FP2 function. For more details, refer to Sauerbrei et al. (2023).

1.3.2 Fractional polynomials are unsuitable for modeling some types of functions

1.3.2.1 MFPA to model sigmoid functions

Sigmoid (doubly asymptotic) functions are not represented in the class of standard FP functions. Royston (2014) introduced the approximate cumulative distribution (ACD) transformation of a continuous covariate x as a route toward modeling a sigmoid relationship between x and an outcome variable. In addition, R&S (2016) proposed the MFPA procedure, which extends standard MFP by permitting selection of sigmoid functions derived from the ACD transformation when supported by sufficiently strongly evidence in the data.

2 Introduction to mfp2 package

`mfp2` is a package that selects the MFP model. In addition, it has the ability to model a sigmoid relationship between `x` and an outcome variable `y` using the ACD transformation proposed by Royston (2016). The package offers three options for variable and function selection: p-value, Akaike information criterion (AIC), and Bayesian information criterion (BIC). Furthermore, it provides functions for prediction and plotting. Currently, the package implements linear, logistic, Poisson, and Cox regression models. However, the package is designed in such a way that it can easily incorporate other generalized linear models or parametric survival models.

The main function, `mfp2()`, implements both MFP and MFP with ACD transformation. It offers two interfaces for input data. The first interface allows direct input of the predictor matrix `x` and the outcome vector `y`. The second interface uses a formula object in conjunction with a data.frame, similar to the `glm()` function with slight modifications. Both interfaces are equivalent in terms of functionality.

The authors of `mfp2` are Edwin Kipruto, Michael Kammer, Patrick Royston, and Willi Sauerbrei, with contribution from Gregory Steiner and Georg Heinze. The R package is maintained by Edwin Kipruto, while the STATA version of `mfp` is maintained by Patrick Royston.

This vignette describes basic usage of `mfp2` in R. There are additional vignettes available that will further enhance your understanding on the mfp2 package.

- "Introduction to Multivariable Fractional Polynomial" provides an overview of multivariable fractional polynomials.

2.1 Estimation algorithm

The estimation algorithm employed in `mfp2` sequentially processes the predictors using a back-fitting approach. It calculates the p-values of each predictor using the likelihood ratio test, assuming linearity. Subsequently, the predictors are arranged based on these p-values. By default, the predictors are arranged in order of decreasing statistical significance. This ordering aims to prioritize modeling relatively important variables before less important ones. This approach may help mitigate potential challenges in model fitting arising from collinearity or more generally, the presence of "concurvity" among the predictors (stata??). Although alternative options for predictor ordering are available, we prefer the default option.

If a predictor contains nonpositive values, the program by default shifts the location of the predictor x to ensure positivity. In addition, it scales the shifted predictor before the first cycle of the algorithm. For more information on shifting and scaling, please refer to section xx.

At the initial cycle, the best-fitting FP function for the first variable (after ordering) is determined, with all the other variables assumed to be linear. The functional form (but not the estimated regression coefficients) is kept, and the process is repeated for the other variables. 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 except the one currently being processed

A variable whose functional form is prespecified to be linear is tested for exclusion within the above procedure when its nominal p-value is less than 1 or argument `keep = FALSE`; otherwise, it is included. 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 1–5 cycles.

2.2 Installation

To install the `mfp2` package, enter the following command in the R console:

```
install.packages("mfp2")
```

2.3 Quick Start

The purpose of this section is to provide users with a comprehensive understanding of the `mfp2` package. We will provide a concise overview of its key functions and resulting outputs. We will delve into each function in detail, highlighting their specific use. This will provide users with a deeper understanding of the package's functionality. The package includes a built-in dataset that is specifically designed for analysis within the `mfp2` framework.

To begin, let's load the `mfp2` package:

```
library(mfp2)
```

2.4 Linear Regression

The default `family` in `mfp2` package is `Gaussian`, which fits a Gaussian linear model. In this section, we will demonstrate how to fit this model. We will use the prostate cancer data (Stamey et al., 1989) included in our package. The dataset contains seven predictors (six continuous variables and one binary variable) and a continuous outcome variable (log prostate-specific antigen (lpsa)) of 97 patients with prostate cancer. Our aim is to determine whether non-linear functional relationships exist between the predictors and the outcome variable.

Load the `prostate` dataset from the `mfp2` package and display the first few rows of the dataset

```
# Load prostate data
data("prostate")
head(prostate)
#> # A tibble: 6 x 9
#>   obsno   age   svi pgg45 cavol weight   bph    cp   lpsa
#>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>
#> 1     1    50     0     0 0.560   16.0  0.25  0.25 -0.431
#> 2     2    58     0     0 0.370   27.7  0.25  0.25 -0.163
#> 3     3    74     0    20 0.600   14.7  0.25  0.25 -0.163
#> 4     4    58     0     0 0.300   26.6  0.25  0.25 -0.163
#> 5     5    62     0     0 2.12    31.0  0.25  0.25  0.372
#> 6     6    50     0     0 0.350   25.2  0.25  0.25  0.765

# create predictor matrix x and numeric vector y
x <- as.matrix(prostate[,2:8])
y <- as.numeric(prostate$lpsa)
```

The command loads a dataframe from the **R** data archive since the `mfp2` package is already loaded. We create a matrix `x` and a numeric vector `y` from the dataframe.

2.4.1 Fitting MFP Models Using Default and Formula Interface

The default interface of `mfp2()` requires a matrix of predictors `x` and a numeric vector of response `y` for continuous outcomes. If you have one predictor, make sure you convert it into a matrix with a single column.

We demonstrate how to fit the Gaussian linear model using the default interface of `mfp2()` with default parameters.

```
fit <- mfp2(x, y)
```

The fit is an object of class `mfp2` that inherits from `glm` and `lm`. This means that `mfp2` inherits properties and methods from `glm` and `lm`, which allows the `mfp2` to utilize methods and functions specific to `glm` and `lm`. Various methods are defined for extracting components from the `mfp2` object, such as coef, print, summary, fracplot, and predict. These methods allow users to access different components of the mfp2 object, including coefficient estimates, plotting functionalities, and predictions.

To fit the same model using the formula interface, we use the `fp()` function.

```
fit <- mfp2(lpsa ~ fp(age) + svi + fp(pgg45) + fp(cavol) + fp(weight) + fp(bph) +
    fp(cp), data = prostate)
```

The main distinction between the `mfp2()` and `glm()` functions in **R** is the inclusion of the `fp()` function within the formula. The presence of the `fp()` function in the formula indicates that the variables included within it should undergo fractional polynomial (FP) transformation, provided that the degree of freedom (df) is not equal to 1. A df of 1 indicates a linear relationship, which does not require transformation. Note that df is an argument in the `fp()` function. For more details on the fp() function, please refer to section xx

The variable `svi` is a binary variable and is therefore not passed to the `fp()` function. This is because binary or factor variables do not undergo FP transformation. If a binary variable is passed to the `fp()` function, the program will automatically set the df to 1, treating the variable as linear. However, passing a factor variable to the `fp()` function will result in an error. For more details on how `mfp2` handles factor variables, refer to section XX.

2.4.2 Shifting and Scaling of Predictors

Fractional polynomials are defined only for positive variables due to the use of logarithms and other powers such as square root. Thus, `mfp2()`function estimates shifting factors for each variables to ensure positivity. The function `find_shift_factor()`, used internally by `mfp2`, automatically estimates shifting factors for each continuous variables. The formula used to estimate the shifting factor for a variable, say `x1`, is given by:

$$shift_{x1} = \gamma - min(x1)$$

where $min(x1)$ is the smallest observed value of `x1`, while γ is the minimum increment between successive ordered sample values of `x1`, excluding 0 (Royston and Sauerbrei,2008). The new variable $x1' = x1 + shift_{x1}$ will then be used by `mfp2()` in estimating the FP powers.

For example, to estimate shifting factors for predictor matrix `x` from prostate data in R, you can run the following code:

```
# minimum values for each predictor
apply(x, 2, min)
#>    age    svi  pgg45  cavol weight    bph     cp
#>  41.00   0.00   0.00   0.26  10.75   0.25   0.25

# shifting values for each predictor
apply(x, 2, find_shift_factor)
#>    age    svi  pgg45  cavol weight    bph     cp
#>      0      0      1      0      0      0      0
```

We see that among the continuous variables, only the variable `pgg45` is shifted by a factor of 1, which is attributed to its minimum value being 0. Even though the variable `svi` also has a minimum value of 0, it is not shifted because its a binary variable. The user can manually set the shifting factors for each variable in `mfp2()` function.

If the values of the variables are too large or too small, it is important to scale the variables to reduce the chances of numerical underflow or overflow which can lead to inaccuracies and difficulties in estimating the model. Scaling can be done automatically or by directly specifying the scaling values for each variables so that the magnitude of the `x` values are not too extreme. By default scaling factors are estimated by the program as follows.

After adjusting the location of `x` (if necessary) so that its minimum value is positive, creating `x'` automatic scaling will divide each value of `x'` by 10^p where the exponent `p` is given by

$$p = sign(k) \times floor(|k|) \quad \text{where} \quad k = log_{10}(max(x') - min(x'))$$

The `mfp2()` function uses this formula to scale `x` matrix, and the scaling process is implemented through the `find_scale_factor()` function. The following R code demonstrates the estimation of scaling factors for `x`. From the output below, we see that the variables `age`, `cavol`, and `cp` have scaling factors of 10 each, while the variables `pgg45` and `weight` have scaling factors of 100 each. Each variable will be divided by its corresponding scaling factor. A scaling factor of 1 implies no scaling.

```
# shift x if nonpositive values exist
shift <- apply(x, 2, find_shift_factor)
xnew <- sweep(x, 2, shift, "+")

# scaling factors
apply(xnew, 2, find_scale_factor)
#>    age    svi  pgg45  cavol weight    bph     cp
#>     10      1    100     10    100      1     10
```

To manually enter shifting and scaling factors, the `mfp2()` function provides the `shift` and `scale` arguments. In the default usage of `mfp2()`, a vector of shifting or scaling factors, with a length equal to the number of predictors, can be provided. In the formula interface, shifting or scaling factors can be directly specified within the`fp()` function. Below is an example to illustrate this:

```
# Default interface
mfp2(x, y, shift = c(0, 0, 1, 0, 0, 0, 0), scale = c(10, 1, 100, 10, 100, 1,
    10))

# Formula interface
mfp2(lpsa ~ fp(age, shift = 0, scale = 10) + svi + fp(pgg45, shift = 1, scale = 100) +
    fp(cavol, shift = 0, scale = 10) + fp(weight, shift = 0, scale = 100) +
    fp(bph, shift = 0, scale = 1) + fp(cp, shift = 0, scale = 10), data = prostate)
```

In the default interface, each variable in the `x` matrix is assigned a shifting and scaling factor based on their respective positions. For instance, the first variable in the column of `x`, which is `age`, is assigned a shifting factor of 0 and a scaling factor of 10. The second variable, `svi`, is assigned a shifting factor of 0 and a scaling factor of 1, and so on.

2.4.3 Setting degrees of freedom for each variable

The degrees of freedom (df) for each predictor (excluding the intercept) are twice the degrees of freedom of the FP. For instance, if the maximum allowed complexity of variable x_1 is a second-degree FP (FP2), then the degrees of freedom assigned to this variable should be 4. The default `df` is 4 for each predictor.

After assigning the default degrees of freedom to each variable, the program proceeds and overrides the default value based on the number of unique values for a given variable. The rules for overriding the default degrees of freedom are as follows:

- If a variable has 2-3 distinct values, it is assigned `df` = 1 (linear).
- If a variable has 4-5 distinct values, it is assigned `df` = min(2, default).
- If a variable has 6 or more distinct values, it is assigned `df` = default.

These rules ensure that the appropriate `df` are assigned to variables. For instance, it is not sensible to fit an FP2 function to a variable with only 3 distinct values.

The following code illustrates how to set different `df` for each variable. In the default interface, the `df` of the binary variable `svi` is explicitly set to 1, while In the formula interface, there is no need to specify the `df`, as the program automatically assigns `df = 1` for binary variables.

If the user attempts to enter `df = 4` for `svi` in the default interface, the program will reset the `df` to 1 and issue a warning.

```
# Default Interface
mfp2(x, y, df = c(4, 1, 4, 4, 4, 4, 4))

# Formula Interface
mfp2(lpsa ~ fp(age, df = 4) + svi + fp(pgg45, df = 4) + fp(cavol, df = 4) +
    fp(weight, df = 4) + fp(bph, df = 4) + fp(cp, df = 4), data = prostate)
```

If the user does not explicitly assign `df` to variables, the program will automatically assign a `df = 4` to each variable. It is important to note that even if a continuous variable is not passed to the `fp()function in the`

`formula interface, the`df‘ of that variable will still be set to the default value and will later be adjusted based on the unique values. The following three examples are equivalent:

```
# Default Interface
mfp2(x, y)

# Formula Interface
mfp2(lpsa ~ fp(age) + svi + fp(pgg45) + fp(cavol) + fp(weight) + fp(bph) + fp(cp),
    data = prostate)

# Formula Interface but `cp` not passed to the fp() function
mfp2(lpsa ~ fp(age) + svi + fp(pgg45) + fp(cavol) + fp(weight) + fp(bph) + cp,
    data = prostate)
```

2.4.4 Tuning parameters for MFP

The two key components of MFP are variable selection with backward elimination (BE) and function selection for continuous variables through the function selection procedure (FSP). These components require two nominal significance levels: α_1 for variable selection with BE and α_2 for comparing the fit of the functions within the FSP. The choice of these significance levels strongly influences the complexity and stability of the final model. While it is possible to use the same nominal significance level for both components, they can also differ based on the aims of the analysis.

The `mfp2()` function has an argument called `criterion`, which allows the user to specify the criteria for variable and function selection. The default criterion is `pvalue`, which enables the user to set the two nominal significance levels. Please refer to section section 1.4.4.1 for instructions on setting the nominal significance levels.

Information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), have been proposed for selecting models fitted on the same data. The `mfp2` package offers an alternative approach to variable and function selection by utilizing information criteria. In the MFP framework, each predictor is evaluated univariately while accounting for the other predictors within an overarching backfitting algorithm. This algorithm iteratively assesses each predictor and selects the model (null, linear, FP1, FP2, etc.) with the minimum AIC or BIC for each variable.

The “null” model refers to a model without the predictor of interest. A “linear” model assumes a linear relationship with the outcome variable, while an FP1 model assumes a non-linear relationship using the FP1 function. The `criterion` argument allows users to specify whether they want to use AIC or BIC criteria for both variable and function selection. If AIC or BIC is selected as the criterion, the nominal significance levels set through the `select` and `alpha` arguments will be ignored. For details on using AIC and BIC in variable and function selection, please refer to section 1.4.4.2.

2.4.4.1 Nominal significance levels The `mfp2()` function has the arguments `select` and `alpha` for setting the nominal significance level for variable selection by BE and for testing between FP models of different degrees, respectively. It is important to note that when using these arguments, you should ensure that the `criterion` is set to `"pvalue"` to correctly use the specified significance levels.

For variable selection, a significance level can be set using the `select` argument. A value of 1 (`select = 1`) for all variables forces them all into the model. Setting the nominal significance level to be 1 for a given variable forces it into the model, leaving others to be selected or not. A variable is dropped if its removal leads to a nonsignificant increase in deviance

On the other hand, the `alpha` argument is used to determine the complexity of the selected FP function. A value of 1 (`alpha = 1`) will choose the most complex FP function permitted for a given variable. For

example, if FP2 is the most complex function allowed for variable x1, setting alpha = 1 will select the best FP2 function.

The rules for setting `select` and `alpha` are the same as those for setting `df` (see section 1.4.3). The following R codes shows how to set equal nominal significance levels for variable and function selection ($\alpha_1 = \alpha_2 = 0.05$) for each variable and produces identical results. Setting different nominal significance levels is straightforward. Simply replace the value 0.05 with the desired significance level of your choice.

```
# Default Interface
mfp2(x, y, select = rep(0.05, ncol(x)), alpha = rep(0.05, ncol(x)))

# Formula Interface
mfp2(lpsa ~ fp(age, select = 0.05) + svi + fp(pgg45, select = 0.05) + fp(cavol,
    select = 0.05) + fp(weight, select = 0.05) + fp(bph, select = 0.05) + fp(cp,
    select = 0.05), select = 0.05, alpha = 0.05, data = prostate)
```

In the formula interface, binary variables such as `svi` that are not passed in the `fp()` function utilize the global `select` argument. In our example, the global parameter is set to `select = 0.05`. If several binary variables exist in the model, the global parameters will be used for all of them. However, if specific parameters need to be set for individual binary variables, the user can use the `fp()`function. In summary, if a variable is not passed through the `fp` function, it will utilize the global parameters.

Suppose we want to force the variables "age" and "svi" in the model. To achieve this, we have two options:

- Set `select = 1` for `age` and `svi` in the `mfp2()` function.

- Alternatively, we can use the `keep` argument, which resets the nominal significance levels for `age` and `svi` to 1 if they are different from 1. This ensures that these variables are retained in the model.

```
#------Default Interface
# Set select to 1 for age and svi
mfp2(x, y, select = c(1, 1, 0.05, 0.05, 0.05, 0.05, 0.05), alpha = rep(0.05,
    ncol(x)))

# use keep argument
mfp2(x, y, select = c(0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05), alpha = rep(0.05,
    ncol(x)), keep = c("age", "svi"))

#-------Formula Interface
# use fp() function and set select to 1 for age and svi
mfp2(lpsa ~ fp(age, select = 1) + fp(svi, df = 1, select = 1) + fp(pgg45, select = 0.05) +
    fp(cavol, select = 0.05) + fp(weight, select = 0.05) + fp(bph, select = 0.05) +
    fp(cp, select = 0.05), select = 0.05, alpha = 0.05, data = prostate)

# use keep argument
mfp2(lpsa ~ fp(age, select = 0.05) + svi + fp(pgg45, select = 0.05) + fp(cavol,
    select = 0.05) + fp(weight, select = 0.05) + fp(bph, select = 0.05) + fp(cp,
    select = 0.05), select = 0.05, alpha = 0.05, keep = c("age", "svi"), data = prostate)
```

2.4.4.2 Information criterion Instead of using nominal significance levels (α_1, α_2) for variable and function selection, an alternative approach is to directly utilize the AIC or BIC, defined below.

$$AIC = -2\log(L) + 2k$$
$$BIC = -2\log(L) + \log(n) \times k$$

Where $\log(L)$ is the maximum log-likelihood of the fitted model, which measures how well the model fits the data. The parameter k corresponds to the number of estimated parameters in the model (regression estimates and FP powers) and n is the sample size or the number of observations in the dataset. AIC and BIC consider both model fit and complexity, with lower values indicating better-fitting models.

For instance, when selecting the best model for a variable of interest (z) with fixed adjustment variables x1 (with power p3) and x2 (linear), we can compare the AIC and BIC of different models, such as FP2, FP1, linear, and null models. The adjustment models all have the same number of parameters (4, including the intercept if it exists). However, the FP2, FP1, and linear models have additional 4, 2, and 1 parameters, respectively. The total number of parameters for each model type is used in calculating the AIC and BIC. The model with the smallest AIC or BIC is then selected.

In `mfp2` package, this can be achieved by setting the `criterion` argument to either "aic" or "bic" in the `mfp2()` function. Additionally, if there is a need to force certain variables, such as "age" and "svi", into the model, the `keep` argument can be used.

The following R code demonstrates how to implement this approach using both the default and formula interfaces, as well as how to force specific variables into the model:

```
# Default Interface
mfp2(x, y, criterion = "aic", keep = c("age", "svi"))

# Formula Interface
mfp2(lpsa ~ fp(age) + fp(svi, df = 1) + fp(pgg45) + fp(cavol) + fp(weight) + fp(bph) +
    fp(cp), criterion = "aic", keep = c("age", "svi"), data = prostate)
```

2.4.5 Model comparison tests

The FSP in `mfp2()` function compares various models for the variable of interest. For instance, if the most complex allowed FP function is FP2, the FSP will compare the best FP2 model with the null model, the best FP2 model with the Linear model, and the best FP2 model with the best FP1 model, when the `criterion = "pvalue"`. The deviance for each model (NULL, Linear, FP1, and FP2) and their corresponding differences and p-values will be calculated.

When comparing Gaussian models, the `mfp2()` function provides two options for calculating p-values: the F-test and the Chi-square test. The Chi-square test is the default option. For other model families like Cox or logistic regression models, the Chi-square test is used. For more detailed information, please refer to page 23 of the MFP Stata manual paper available at this link: https://www.stata.com/manuals/rfp.pdf

To use the F-test in **R**, we can set the `ftest` argument to TRUE (`ftest = TRUE`), as demonstrated in the example below. Conversely, to use the Chi-square test, set `ftest = FALSE`.

```
# Default Interface
mfp2(x, y, criterion = "pvalue", ftest = TRUE)

# Formula Interface
mfp2(lpsa ~ fp(age) + svi + fp(pgg45) + fp(cavol) + fp(weight) + fp(bph) + fp(cp),
    criterion = "pvalue", ftest = TRUE, data = prostate)
```

Please note that the p-values reported by the `mfp` program in `Stata` for Gaussian models are based on the F-test. If you intend to compare the results between the two software packages, it is crucial to ensure that `ftest = TRUE` is set in R.

However, it is important to be aware that the older version of the `mfp` package in R uses the Chi-square test for all model types. Therefore, when comparing the results between the two R packages, the user must set`ftest = FALSE`.

2.4.6 Fraction Polynomial Powers

Low order polynomials offer a limited family of shapes, and high order polynomials may fit poorly at the extreme values of the covariates. Due to these limitations, Royston an Altman (1994) proposed an extended family of curves known as fractional polynomials, whose power terms are restricted to a small predefined set, $S = \{2, 1, 0.5, 0, 0.5, 1, 2, 3\}$ of integer and non-integer values. The powers are selected so that conventional polynomials are a subset of the family. The power of 0 denotes the natural logarithm of x, log(x), while the power of 1 denotes no transformation. Furthermore, the set includes other powers such as the reciprocal (-1), square root (0.5), and square (2).

By default, the `mfp2()` function utilizes the predefined set S for each continuous covariate. However, there may be situations where users want to provide their own power terms based on their subject matter knowledge. In such cases, the `powers` argument which takes a list of distinct powers can be employed to specify the desired power terms. If the user provides identical powers for a variable, such as (0, 0, 2), the program will remove the duplicates and consider only the powers 0 and 2.

The following example illustrates how to assign different power terms to covariates. Two power terms (0 and 0.5) are evaluated for the "age" covariate. This set includes three linear models of degree 2 (with power combinations of (0, 0), (0, 0.5), and (0.5, 0.5)), and two models of degree 1 (with power of 0 and 0.5).

The "cavol" variable is assigned a single power term (0), indicating that FP1 is the most complex function for this variable. The remaining continuous variables, namely "pgg45," "weight," "bph," and "cp," are not assigned any power terms. Instead, the default powers defined in set S are utilized and a search through all possible fractional polynomials up to the degree set by `df` is performed.

It is important to note that when using the `fp()` function in the formula interface, the power terms provided must be a vector not a list since they are specific to a particular variable.

```
# create a list of power terms for covariates age and cavol
powx <- list(age = c(0, 0.5), cavol = 0)
# Default Interface
mfp2(x, y, criterion = "pvalue", powers = powx)

# Formula Interface
mfp2(lpsa ~ fp(age, powers = c(0, 0.5)) + svi + fp(pgg45) + fp(cavol, powers = 0) +
    fp(weight) + fp(bph) + fp(cp), data = prostate)
```

2.4.7 Explanation of output from model-selection algorithm

Using the prostate example, we briefly explain how the algorithm works. Similar to backward elimination, it starts with the full model (model with all variables) and investigates whether variables can be eliminated. However, for each of the continuous variables, the FSP is used to check whether a non-linear function fits the data significantly better than a linear function. After the first cycle, some variables may be eliminated from the model, and for some continuous variables, a more suitable non-linear function may be identified as a better fit for the data.

The algorithm then proceeds to the second cycle, but the new starting model now has fewer variables due to the elimination process in the previous cycle. Additionally, non-linear functions may have been identified for some of the continuous variables. In the second cycle, all variables will be reconsidered, even if they were not significant at the end of the first cycle, and FSP is used again to determine the 'best' fitting FP function (the functions may be different due to potential variations in the adjustment variables). The results obtained from the second cycle then serve as the starting model for the third cycle. In most cases, the variables and functions selected remain unchanged in cycles 3 or 4 and the algorithm stops with the final MFP model.

The order of 'searching' for model improvement by better fitting non-linear functions is important. Mismodelling the functional form of a variable with a strong effect is more critical than mismodelling the functional

form of a variable with a weak effect. Therefore, the order is determined by the p-values from the full model. Variables with small p-values are considered first.

To explain the output of the MFP algorithm, we will use the default interface of the `mfp2()` function to build an MFP model with the default parameters. Specifically, all continuous variables are assigned a degree of freedom (df) of 4, implying that FP2 is the most complex permitted function. Moreover, we employ the 'pvalue' as the criterion for variable and function selection, and we set the nominal significance levels for both components to 0.05 for all variables. Finally, we use the F-test instead of Chi-square to calculate the p-values. The **R** output displays the df used for each variable, as denoted by the "initial degrees of freedom". The program correctly identifies "svi" as a binary variable and assigns it a df of 1. Additionally, the variables are ordered based on their p-values in descending order of significance, as indicated by the "visiting order". The variable "cavol" has the smallest p-value and will be evaluated first, while "age" has the largest p-value and will be evaluated last.

```
fit <- mfp2(x, y, criterion = "pvalue", select = 0.05, alpha = 0.05, ftest = TRUE)
#>
#> i Initial degrees of freedom:
#>    age svi pgg45 cavol weight bph cp
#> df   4   1     4     4      4   4  4
#>
#> i Visiting order: cavol, svi, pgg45, weight, bph, cp, age
#>
#> ----------------------
#> i Running MFP Cycle 1
#> ----------------------
#>
#> Variable: cavol (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -0.5, 1   12    196.3       .                .              .
#> null             NA        8     240.1       FP2              43.8           0.0000
#> linear           1         9     214.3       FP2              18.0           0.0011
#> FP1              0         10    199.7       FP2              3.4            0.2226
#> Selected: FP1
#>
#> Variable: svi (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> null             NA        8     208.6       .                .              .
#> linear           1         9     199.7       null             -9.0           0.0042
#> Selected: linear
#>
#> Variable: pgg45 (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -2, -2    12    196.7       .                .              .
#> null             NA        8     202.0       FP2              5.3            0.3091
#> Selected: null
#>
#> Variable: weight (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -2, -2    11    199.3       .                .              .
#> null             NA        7     209.7       FP2              10.4           0.0521
#> Selected: null
#>
#> Variable: bph (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
```

```
#> FP2              -1, 3     10    207.7       .                .                .
#> null             NA        6     217.7       FP2              10.0             0.0567
#> Selected: null
#>
#> Variable: cp (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              2, 3      9     213.2       .                .                .
#> null             NA        5     217.9       FP2              4.6              0.3655
#> Selected: null
#>
#> Variable: age (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -1, -1    8     216.6       .                .                .
#> null             NA        4     217.9       FP2              1.2              0.8839
#> Selected: null
#>
#> ---------------------
#> i Running MFP Cycle 2
#> ---------------------
#>
#> Variable: cavol (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -0.5, 1   7     215.0       .                .                .
#> null             NA        3     264.6       FP2              49.6             0.0000
#> linear           1         4     238.1       FP2              23.0             0.0001
#> FP1              0         5     217.9       FP2              2.8              0.2646
#> Selected: FP1
#>
#> Variable: svi (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> null             NA        3     226.9       .                .                .
#> linear           1         4     217.9       null             -9.0             0.0032
#> Selected: linear
#>
#> Variable: pgg45 (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              0.5, 3    8     214.3       .                .                .
#> null             NA        4     217.9       FP2              3.6              0.4973
#> Selected: null
#>
#> Variable: weight (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -2, -2    8     202.1       .                .                .
#> null             NA        4     217.9       FP2              15.7             0.0052
#> linear           1         5     205.0       FP2              2.9              0.4438
#> Selected: linear
#>
#> Variable: bph (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              0.5, 0.5  9     199.1       .                .                .
#> null             NA        5     205.0       FP2              5.9              0.2460
#> Selected: null
#>
```

```
#> Variable: cp (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              2, 3      9     200.3       .                .              .
#> null             NA        5     205.0       FP2              4.7            0.3617
#> Selected: null
#>
#> Variable: age (keep = FALSE)
#>                  Powers    DF    Deviance    Versus           Deviance diff. P-value
#> FP2              -0.5, 0   9     203.2       .                .              .
#> null             NA        5     205.0       FP2              1.8            0.7926
#> Selected: null
....
```

After the variables are ordered, the MFP algorithm starts by searching for a suitable function for "cavol". It compares the best-fitting FP2 (-0.5, 1) function for "cavol" against a null model that excludes "cavol." This comparison involves adjusting for all other six variables ("svi" to "age") using linear functions. The test is highly significant (p = 0.0000), indicating that the best FP2 function fits significantly better than a null model. Next, the model with the best FP2 function is compared to a model assuming linearity, and the test remains significant (p = 0.0011). This suggests that "cavol" can be better described by a non-linear function at this stage. Finally, the best FP2 function is compared to the best FP1 (0) function, and the test is not significant (p = 0.2226). Thus, at this stage in the model-selection procedure, the final function for "cavol" is FP1 with power 0, denoting a log function.

The next variable to be evaluated is "svi," which is a binary variable. In this case, only the test of null versus linear is appropriate. Both the null and linear models of "svi" include "log(cavol)" (the log function just selected) and the other five variables ("pgg45", "cp", "weight", "bph", and "age") that are still present in the model. The test of null versus linear is significant (p = 0.0042), indicating that "svi" remains in the model.

Next, we evaluate the continuous variable "pgg45" in the same model as for "svi". The p-value for the first test (FP2 versus null) is not significant (p = 0.3091), and the variable is eliminated from the model and will not be considered in the rest of the first cycle.

Following that, the algorithm evaluates the variable "weight" in a model adjusting for "log(cavol)", "svi", and the three variables ("cp", "bph", and "age") that have not yet been evaluated. The first test is non-significant (p = 0.0521), and the variable is eliminated for the rest of the first cycle. In the subsequent steps, "bph", "cp", and "age" are also eliminated.

At the end of the first cycle, only two variables were selected: "log(cavol)" and "svi".

The variables selected in the first cycle becomes the new starting model for the second cycle. In the second cycle, the effect of "cavol" is investigated again, but this time in a simpler model adjusting for "svi" only. Deviances are larger compared to cycle 1 because the five variables ("pgg45", "age", "weight", "bph" and "cp") no longer belong to the 'adjustment' model. However, the FSP still selects "log(cavol)".

The "svi" is evaluated in a model that includes only "log(cavol)", and the test of linear vs null confirms that the variable should be included in the model. Pgg45 is evaluated in a model with "svi" and "log(cavol)" as adjustment variables, but the test of inclusion (FP2 vs null) is not significant and the variable is again eliminated.

The variable "weight" is considered next in a model with "svi" and "log(cavol)". In contrast to the first cycle, the test for inclusion (FP2 vs null) is significant (p = 0.0052), indicating that "weight" needs to be re-included into the model. The second test of the FSP (FP2 vs linearity) is non-significant and a linear function is chosen for weight. The inclusion of weight in the second cycle is a result of eliminating "bph", "cp" and "age" in the first cycle, and of the correlation with these variables

In the subsequent steps, "bph", "cp", and "age" are investigated in models adjusting for "log(cavol)", "svi", and "weight" (linear). Compared to the first cycle, the p-values change, but all of them are much larger

than 0.05. Therefore, none of these variables are included in the model.

The second cycle ends with the model consisting of "log(cavol)", "svi" (binary), and "weight" (linear).

This model serves as the starting point for the third cycle (not shown), where all seven variables are investigated again. However, there are no further change in the selected variables or functions occurs, so MFP terminates with the three variables as the second cycle. For more detailed information, please visit the MFP website at For more detailed information, please visit the MFP website.

2.4.8 Graphical presentation of FP functions

The regression estimates ($\hat{\beta}$) for FP terms provide incomplete information as a report of results since they tell us little about the fitted function for the variable of interest. A more informative approach is to visualize the function plot.

Suppose an x is modeled as an FP with power(s) p and parameter estimate β. The partial predictor for x, denoted as η_x, is defined as $\beta_0 + x^p\beta$ (see Section 4.3). Its standard error (conditional on the selection of p) is given by the square root of Equation (4.1). In a multivariable model, the interpretation of η_x is an adjusted estimate of the functional form for the effect of x. With the recommended centering of FP functions around the mean x of x (see Section 4.11), η_x equals the intercept β_0 at x. In a model without an intercept (e.g., the Cox model), η_x and its standard error equal zero at x.

A function plot, sometimes referred to as a component-plus-residual plot when residuals are plotted, illustrates the relationship between x and η_x. It plots η_x plus or minus twice its standard error against x, overlaid with the sum of η_x and the residual for each individual observation.

Function plots for cavol and weight are shown in Figure 6.2. In a model with a single predictor, the points would exactly equal the y values. They show the amount of residual variation at each x and may indicate lack of fit and outliers in y. Additionally, they provide an understanding of the positions of the covariate observations. In this example, there are no obviously disturbing features to note about the model. Further discussion on the role of function plots in model criticism can be found in Section 6.5.1.

For generating function plots, the 'fracplot' function in Stata can be used. It produces a component-plus-residual plot, which, for normal-error models with constant weights and one covariate, amounts to a plot of the observations with the fitted line inscribed. For other normal-error models, weighted residuals are calculated and added to the fitted values. In models with additional covariates, the line represents the partial linear predictor for the variable in question (xvar1 or a covariate) and includes the intercept.

For generalized linear and Cox models, the fitted values are plotted on the scale of the 'index' (linear predictor). Deviance residuals are added to the (partial) linear predictor to give component-plus-residual values, which are represented as small circles in the plot.

2.4.9 Prediction

2.5 Logistic Regression

2.6 Poisson regression

2.7 Survival data

We illustrate two of the analyses performed by Sauerbrei and Royston (1999). We use brcancer.dta, which contains prognostic factors data from the German Breast Cancer Study Group of patients with node-positive breast cancer. The response variable is recurrence-free survival time (rectime), and the censoring variable is censrec. There are 686 patients with 299 events. We use Cox regression to predict the log hazard of recurrence from prognostic factors of which five are continuous (x1, x3, x5, x6, x7) and three are binary (x2,

x4a, x4b). Hormonal therapy (hormon) is known to reduce recurrence rates and is forced into the model. We use mfp to build a model from the initial set of eight predictors by using the backfitting model-selection algorithm. We set the nominal p-value for variable and FP selection to 0.05 for all variables except hormon, which it is set to 1: