D. E. Knuth. *The Art of Computer Programming. Vol. 2: Seminumerical Algorithms*. Addison-Wesley, 3rd edition, 1998. 16

A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, 3rd edition, 2000. 16

P. L'Ecuyer. Random number generation. In J. E. Gentle, W. Haerdle, and Y. Mori, editors, *Handbook of Computational Statistics*, chapter II.2, pages 35–70. Springer-Verlag, Berrlin, 2004. 16

P. L'Ecuyer, R. Simard, E. J. Chen, and W. D. Kelton. An object-oriented random-number package with many long streams and substreams. *Operations Research*, 50(6):1073–1075, 2002. 16, 17

Pierre L'Ecuyer
Université de Montréal, Canada
lecuyer@iro.umontreal.ca

Josef Leydold
Wirtschaftsuniversität Wien, Austria
leydold@statistik.wu-wien.ac.at

mfp

Multivariable Fractional Polynomials

by Axel Benner

Introduction

The `mfp` package is targeted at the use of multivariable fractional polynomials for modelling the influence of continuous, categorical and binary covariates on the outcome in regression models, as introduced by Royston and Altman (1994) and modified by Sauerbrei and Royston (1999). It combines backward elimination with a systematic search for a 'suitable' transformation to represent the influence of each continuous covariate on the outcome. The stability of the models selected was investigated in Royston and Sauerbrei (2003). 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 covariate has to be eliminated and the functional forms of the continuous covariates do not change anymore.

In regression models often decisions on the selection and the functional form of covariates must be taken. Although other choices are possible, we propose to use P-values to select models. With many covariates in the model, a large nominal P-value will lead to substantial overfitting. Most often a nominal P-value of 5% is used for variable selection.

The choice of a 'selection' P-value depends on the number of covariates and whether hypothesis generation is an important aim of the study. The `mfp` procedure therefore allows to distinguish between covariates of main interest and confounders by specifying different selection levels for different covariates.

Fractional Polynomials

Suppose that we have an outcome variable, a single continuous covariate, Z, and a suitable regression model relating them. Our starting point is the straight line model, $\beta_1 Z$ (for easier notation we ignore the intercept, β_0). Often this assumption is an adequate description of the functional relationship, but other functional forms must be investigated for possible improvements in model fit. A simple extension of the straight line is a power transformation model, $\beta_1 Z^p$. This model has often been used by practitioners in an *ad hoc* way, utilising different choices of p. We formalise the model slightly by calling it a first-degree fractional polynomial or FP1 function where the powers p are chosen from a restricted set, S (Royston and Altman, 1994). For pragmatic reasons, Royston and Altman (1994) suggested $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where Z^0 denotes $\log(Z)$. No subsequent changes to S have been proposed since then. The power transformation set includes the 'identiy' transformation ($p = 1$), as well as the reciprocal, logarithmic, square root and square transformations. To fit an FP1 model each of the eight values of p is tried. The best-fitting model is defined as the one with the maximum likelihood among these models.

Extension from one-term FP1 functions to the more complex and flexible two-term FP2 functions follows immediately. FP2 functions with powers (p_1, p_2) include the formulations $\beta_1 Z^{p_1} + \beta_2 Z^{p_2}$ and $\beta_1 Z^{p_1} + \beta_2 Z^{p_1} \log(Z)$ if $p_2 = p_1$, the latter being the so-called repeated-powers functions. The best fit among the 36 (28 with $p_2 \neq p_1$ and 8 with $p_2 = p_1$) combinations of powers from S is defined as that with the highest likelihood.

An FPm model of degree m, $m = 1, 2$, is considered to have $2m$ degrees of freedom (d.f.), 1 d.f. for each parameter and 1 d.f. for each power. Because

a restricted number of powers (only those from the set S) are used, the value $2m$ is a slight over-estimate of the effective number of d.f. (Royston and Altman, 1994). Defining the 'deviance' of a FP model as minus twice the maximised log likelihood, statistical significance between FP models is assessed by using the χ^2 distribution.

Model selection

For choosing among FP models at a given significance level of α, two procedures have been proposed in the literature, a sequential selection procedure and a closed testing selection procedure.

For the sequential selection procedure first a 2 d.f. test at level α of the best-fitting second-degree FP, FP2, against the best-fitting first-degree FP, FP1, is performed. If the test is significant, the final model is the best-fitting FP2. If the test is not significant, a 1 d.f. test at the α level of the best fitting FP1 against a straight line follows. If the test is significant, the final model is the best-fitting FP1. Otherwise, a 1 d.f. test at the α level of a straight line against the model omitting the covariate is performed. If the test is significant, the final model is a straight line, otherwise the covariate is omitted.

The closed testing selection procedure (denoted `RA2`) was originally described in Ambler and Royston (2001). It maintains approximately the correct Type I error rate for each covariate tested.

For a specific covariate X the `RA2` algorithm works as follows:

1. Perform a 4 d.f. test at the α level of the best-fitting FP2 against the null model. If the test is not significant, drop X and stop, otherwise continue.

2. Perform a 3 df test at the α level of the best-fitting FP2 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 α level of the best-fitting FP2 against the best-fitting FP1. If the test is significant, the final model is the best-fitting FP2, otherwise the best-fitting FP1.

For the sequential selection procedure the actual Type I error rate may exceed the nominal value when the true relationship is a straight line. The procedure tends to favour more complex models over simple ones. For this reason the R implementation provides only the `RA2` procedure.

Multivariable Fractional Polynomials

Modelling the joint effect of several covariates one may wish to simplify the model, either by dropping non-significant variables and/or by reducing the complexity of FP functions fitted to continuous covariates. A solution to finding a model including FP terms was proposed by Royston and Altman (1994) and refined by Sauerbrei and Royston (1999). Instead of an exhaustive search an iterative algorithm was recommended and termed the `MFP` procedure. It combines backward elimination of variables with a search for the best FP functions of continuous covariates.

The order of covariates for the variable selection procedure is determined by fitting a linear model including all variables and dropping each variable singly according to the first step of a conventional backward elimination procedure. The values of the corresponding test statistics are then used to order the variables, from the most to the least 'significant'. Having imposed an order, each covariate is considered in turn. If a covariate is categorical or binary, a standard likelihood ratio test is performed to determine whether it should be dropped or not. If a covariate is continuous, the FP model selection procedure `RA2` is applied to determine the best FP or to eliminate the covariate. The procedure cycles through the variables in the predetermined order until the selected variables and FP functions do not change any more.

Usage

A typical `mfp` model formula has the form `response` $\sim$ `terms` where `response` is the (numeric) response vector and `terms` is a series of terms, usually separated by $+$ operators, which specifies a linear predictor for `response` and provided by the `formula` argument of the function call. Fractional polynomial terms are indicated by `fp`.

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

The argument `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. Actually linear, logistic, Poisson and Cox regression models have been implemented.

The global argument `alpha` (default: 0.05) sets the FP selection level to be used for all covariates. The global variable selection level is set by argument `select` (default: 1, corresponding to 'no variable selection'). These values can be changed for individual covariates by using the `fp` function in the formula.

The fp function in the formula description of the model defines a fractional polynomial object for a single input variable using arguments df, select, alpha and scale for FP model selection. The df argument determines the maximum degree of the FP model to be tested for the corresponding covariate by its d.f. (default: 4, corresponding to an FP2 model).

The scale argument (default: TRUE) of the fp function denotes the use of pre-transformation scaling to avoid possible numerical problems.

Example

As an example we use the data from a clinical trial for patients with node positive breast cancer from the German Breast Cancer Study Group (GBSG). A total of 7 covariates were investigated. Complete data on covariates and the survival times is available for 686 patients of the study. During a median follow-up of about 5 years 299 events were observed for recurrence free survival time (rfst). For more details see Sauerbrei and Royston (1999) or references given there. The results are computed using the mfp package, version 1.3.1.

The dataset GBSG which contains the data for the set of 686 patients with node-positive breast cancer is provided as example data set in the mfp package.

```
> data(GBSG)
```

The response variable is recurrence free survival time (Surv(rfst, cens)). In the example data set seven possible prognostic factors were given, of which 5 are continuous, age of the patients in years (age), tumor size in mm (tumsize), number of positive lymph nodes (posnodal), progesterone receptor in fmol (prm), estrogen receptor in fmol (esm). One more covariate is binary, menopausal status (menostat), and one is ordered categorical with three levels, tumor grade (tumgrad). Finally, the model has to be adjusted for hormonal therapy (htreat).

According to Sauerbrei and Royston (1999) a pre-transformation of the number of positive nodes was applied,

```
> GBSG$nodetrans <- exp(-0.12 * GBSG$posnodal)
```

In addition a value of 1 was added to prm and esm in the example data set GBSG to obtain positive values for these variables. To be compatible with other applications of mfp on this data set (cp. Sauerbrei et al., 2005) the data for age were divided by 50 and tumor grade levels 2 and 3 were combined.

```
> GBSG$age50 <- GBSG$age/50
> levels(GBSG$tumgrad) <-
    list("1"=1,"2 or 3"=c(2,3))
```

A Cox proportional hazards regression is the standard approach to model the hazard of tumour recurrence. Therefore the mfp model of recurrence free survival on the initial set of 7 covariates, stratified for hormone therapy, is fitted according to a Cox regression model by a call to mfp() using family argument cox. The global FP selection level alpha as well as the global variable selection level select are set to 5%.

```
> f <- mfp(Surv(rfst, cens) ~  fp(age50)
    +fp(nodetrans)+fp(prm)+fp(esm)+fp(tumsize)
    +menostat+tumgrad+strata(htreat),
    family = cox, method="breslow", alpha=0.05,
    select=0.05, data = GBSG)
```

To be compatible with other implementations of multivariable fractional polynomials in SAS and Stata we use the Breslow method for tie handling.

Pre-transformation scaling was used for esm, prm and tumsize.

```
> f$scale
```

	shift	scale
nodetrans	0	1
prm	0	100
tumgrad2 or 3	0	1
tumsize	0	10
menostat2	0	1
age50	0	1
esm	0	100

The final model is given as

```
> f
```

```
Call:
mfp(formula = Surv(rfst, cens) ~ fp(age50) + fp(nodetrans) +
    + fp(prm) + fp(esm) + fp(tumsize) + menostat + tumgrad +
    strata(htreat), data = GBSG,
    family = cox, method = "breslow",
    alpha = 0.05, select = 0.05)


Fractional polynomials:
              df.initial select alpha df.final power1 power2
nodetrans              4   0.05  0.05        1      1      .
prm                    4   0.05  0.05        2    0.5      .
tumgrad2 or 3          1   0.05  0.05        1      1      .
tumsize                4   0.05  0.05        0      .      .
menostat2              1   0.05  0.05        0      .      .
age50                  4   0.05  0.05        4     -2     -1
esm                    4   0.05  0.05        0      .      .

                    coef exp(coef) se(coef)     z       p
nodetrans.1     -1.9777   0.13839   0.2272 -8.70 0.0e+00
prm.1           -0.0572   0.94442   0.0111 -5.16 2.5e-07
tumgrad2 or 3.1  0.5134   1.67092   0.2495  2.06 4.0e-02
age50.1          2.4230  11.27964   0.4753  5.10 3.4e-07
age50.2         -5.3060   0.00496   1.1807 -4.49 7.0e-06

Likelihood ratio test=142  on 5 df, p=0 n= 686
```

Of the possible prognostic covariates esm, menostat and tumsize were excluded from the model (df.final=0). For covariates age and prm nonlinear transformations were chosen (df.final>1). Because of the pre-transformation the function used for the number of positive lymph nodes is also non-linear. For prm an FP1 was selected with $p = 0.5$ corresponding to a square-root transformation. For age an FP2 model was found with $p_1 = -2$, providing a new artificial variable age.1, and $p_2 = -1$, resulting in the new artificial variable age.2.

If one had used a standard Cox regression model with linear functional forms for the continuous covariates the number of positive lymph nodes

(nodetrans), progesterone receptor (prm) and tumor grade (tumgrad) would have a statistical significant effect at the 5% level.

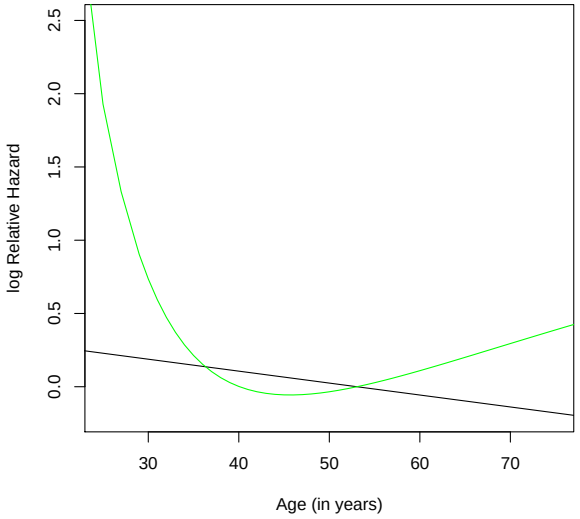

Figure 1: Estimated functional relationship between age and the log relative hazard of recurrence-free survival. The linear function is derived from a standard Cox regression model and is shown as solid black line. The green line shows the functional form as postulated by the mfp model. For both functions a constant was subtracted from the estimated function before plotting so that the median value of age yields 'y=0' in both cases.

The multivariable fractional polynomial Cox regression model additionally selected the FP2 transformations of variable age50 (age50.1 and age50.2). Instead of the untransformed (prm) variable an FP1 of prm was chosen and three variables were deleted.

In Figure 1 the estimated functional forms for age resulting from the standard Cox regression model and the multivariable fractional polynomial regression model are compared.

The result of the R implementation of mfp differs from the one using Stata and SAS by selection of an FP2 transformation with powers -2 and -1 instead of -2 and -0.5 (Sauerbrei et al. (2005)). Using the actual implementation of the Cox regression model in R (function 'coxph' in package "survival", version 2.18) to compare the results of the models including the two different FP2 versions for age50 of R and Stata resulted in nearly the same computed log-likelihoods. The log-likelihood of the model including age50 with powers -2 and -1 is -1528.086 as compared to -1528.107 for the model including age50 with powers -2 and -0.5. The value fit of the resulting mfp object can be used for survival curve estimation of the final model fit (2).

```
pf <- survfit(f$fit)
plot(pf, col=c("red","green"), xscale=365.25,
  xlab="Time (years)",
  ylab="Recurrence free survival rate")
```

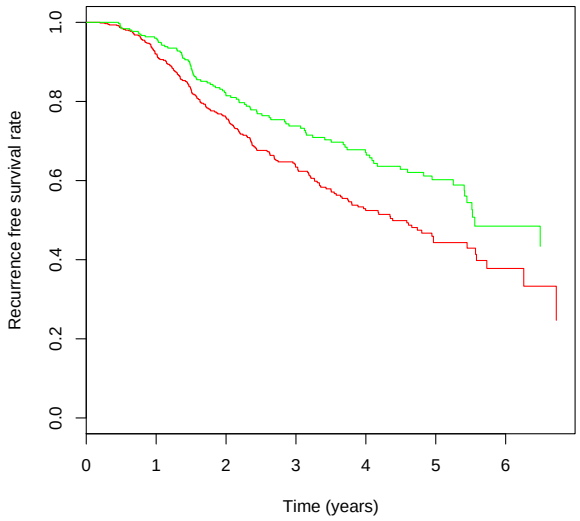

Figure 2: Predicted survival curves of the final mfp model for the two strata defined by hormonal treatment (red line: without hormonal treatment, green line: with hormonal treatment).

Bibliography

G. Ambler and P. Royston. Fractional polynomial model selection procedures: investigation of Type I error rate. *Journal of Statistical Simulation and Computation*, 69:89–108, 2001. 21

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, 43(3):429–467, 1994. 20, 21

P. Royston and W. Sauerbrei. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine*, 22:639–659, 2003. 20

W. Sauerbrei and P. Royston. 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, 1999. 20, 21, 22

W. Sauerbrei, C. Meier-Hirmer, A. Benner, and P. Royston. Multivariable regression model building by using fractional polynomials: Description of SAS, Stata and R programs. *Computational Statistics and Data Analysis*. Article in press. 22, 23

Axel Benner
German Cancer Research Center - Heidelberg, Germany.
benner@dkfz.de