Variable selection with MFP and Splines

## Part 1: Introduction

* Fractional Polynomials are a great and easy tool for many practical applications, with simplicity being one of the greatest merits of the approach. In their work Sauerbrei W, Royston P (1999) introduced their Multivariable Fractional Polynomials (MFP) procedure to select variables and their functional forms in a regression model setting. P-values (or Information Criteria) for each of the two parts a) variable selection with BE and b) selection of a FP function with FSP are the key parameters to determine the complexity of a model selected. P-values may be different for the two parts.
* When working with splines such a procedure is not yet established. Spline users focus on identifying the functional form of a covariate and do not seem to emphasize on the variable selection part. In their book, Royston and Sauerbrei (chapter 9) discuss an informal comparison of MFP with two spline approaches which adhere to the MFP philosophy. BE and a spline procedure (restricted cubic spline or smooting spline) replaces FSP. They focus on restricted cubic splines (or natural splines) and discuss MVRS, a procedure on multivariable variable selection with splines, very similar in nature to MFP. They then present examples on four (Boston, GBSG, PIMA, PBC) different datasets and informally compare results. They consider 4 or 8 df for spline functions and illustrate it’s effect on the model and the spline functions selected.
* Wood (2001) and Marra and Wood (2011) discussed the issue of variable selection by adding extra penalty terms in an additive model. They introduced two approached, both implemented in variable mgcv in R. **The first approach is to modify the smoothing penalty with an additional shrinkage term… so that for large enough smoothing parameters the smooth becomes identically zero. This allows automatic smoothing parameter selection methods to effectively remove the term from the model altogether. The shrinkage component of the penalty is set at a level that usually makes negligable contribution to the penalization of the model, only becoming effective when the term is effectively ‘completely smooth’ according to the conventional penalty. (R help file)**
* **The second approach leaves the original smoothing penalty unchanged, but constructs an additional penalty for each smooth, which penalizes only functions in the null space of the original penalty (the ‘completely smooth’ functions). Hence, if all the smoothing parameters for a term tend to infinity, the term will be selected out of the model. This latter approach is more expensive computationally, but has the advantage that it can be applied automatically to any smooth term. The select argument to gam turns on this method.(R help file)**
* At the moment, there is no specific recommendation on how to fit multivariable models with splines. Although for fractional polynomial the MFP algorithm is well explored and accepted, when working with splines issues of variable and functional form selection are seldomly discussed. The aim of this work would be to illustrate model fitting with splines and fractional polynomials in a set of real datasets and discuss model performance, compare approaches and showcase how to present results in practical applications. The methods that will be considered are:
  + Fractional polynomials with MFP
  + P-splines and thin-plate regression splines in mgcv with variable selection as proposed by Marra and Wood
  + Restricted cubic splines with MVRS (Multivariable Regression Splines -Royston, Saurbrei chapter 9)

## Part 2: Examples on data

## 

## PIMA indians data

## Data set of 768 observations with 8 variables. Binary outcome, logistic regression setting







## Download from: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download>. Or get the same data as in Sauerbrei, Royston book from: PBC data

Survival data available in R with survival package (including only the original 312 patients)

## Bacteremia data

The data set consists of 14,691 observations from different patients with the clinical suspicion to suffer from bacteremia, for whom a a blood culture analysis was performed at the Vienna General Hospital, Austria, between January 2006 and December 2010. It contains the results of the blood culture analysis for bacteremia and the values of 51 potential predictors of bacteremia.

In the ISCB presentation, PIMA and PBC data have already been used to produce some first findings. For each of these datasets we expect to fit models as described in slide 13. Outputs from each model should include:

* Plots of functional forms (example page 216 Royston-Sauerbrei book and slide 16 ISCB presentation)
* Table of variable inclusion (example page 216 Royston-Sauerbrei book and slide 15 ISCB presentation)
* Prediction error of each model (slide 19)
* Other metric of comparison?