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)**

## Part 2: Examples on data

* Here, we reproduce part of that analysis with the aim of showcasing practical applications and provide simple examples for applied research. We re-run their analysis on two datasets: the PIMA indians data and PBC survival data. We apply MFP and MVRS on natural splines but also extend their analysis to include other approaches based on thin plate (TP) regression splines and p-splines (PS). We compare MFP, to MVRS, TP and PS approaches and discuss similarities. We should briefly illustrate the importance of key parameters (MFP with 0.01, 0.05, 0.157)

## 

## PIMA indians data (some findings)

## Data set of 768 observations with 8 variables

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: <https://mfp.imbi.uni-freiburg.de/book#dataset_tables>

## Fit models (as in book page 216), MVRS(5), TS\_1, TS\_2, PS

MVRS(5) is a procedure described by Royston-Saubrei where degrees of freedom of natural splines are set to 5 and then an MFP-like method chooses spline df vs linear vs non-inclusion of the variable.

TS\_1: Thin plate regression splines, fitted with mgcv and REML with a modification on the penalty to drop the effective degrees of freedom below one. When that happens, we assume that the smooth term and the variable should be dropped from the model.

TS\_2: Thin plate regression splines, fitted with mgcv and REML with an additional penalty on the penalty to drop the effective degrees of freedom below one. When that happens, we assume that the smooth term and the variable should be dropped from the model.

PS\_2: P-splines, fitted with mgcv and REML with an additional penalty on the penalty to drop the effective degrees of freedom below one. When that happens, we assume that the smooth term and the variable should be dropped from the model.

NS\_2: Natural splines, fitted with mgcv and REML with an additional penalty on the penalty to drop the effective degrees of freedom below one. When that happens, we assume that the smooth term and the variable should be dropped from the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Term edf | TPS | TP\_1 | TP\_2 | PS\_2 | NS\_2 |
|  | No var selection | Modified penalty | Extra penalty | Extra penalty | Extra penalty |
| s(Pregnancies) | 1.00 | 0.61 | 0.57 | 0.525 | 0.62 |
| s(Glucose) | 1.00 | 1.28 | 0.99 | 0.99 | 2.16 |
| s(BloodPressure) | 1.00 | 0.01 | 0.10 | 0.07 | 0.09 |
| s(SkinThickness) | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| s(Insulin) | 2.10 | 0.00 | 0.00 | 0.54 | 0.00 |
| s(BMI) | 4.18 | 3.70 | 3.91 | 3.71 | 3.65 |
| s(DiabetesPedigreeFunction) | 2.22 | 0.89 | 1.79 | 1.38 | 1.58 |
| s(Age) | 3.41 | 2.99 | 2.88 | 2.70 | 3.01 |

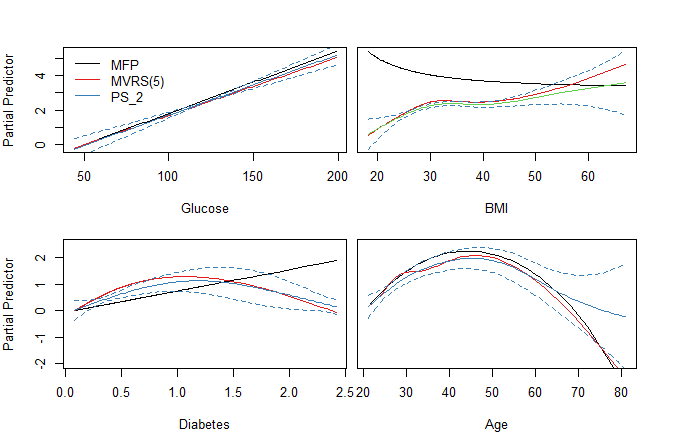
Table 1: Effective degrees of freedom from all mgcv approaches. Methods mostly agree on which variables to select. TP selection 1 excludes diabetes with is included with all other models.

Similar findings can be seen with MFP and MVRS. See Table 2 below. Methods seem to agree on the included variables. MVRS(5) selects one additional variable (Pregnancies) which is not in any other model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **MFP(2)** | **MVRS(5)** | **TS\_1** | **TS\_2** | **PS\_2** |
|  | power | df | edf | edf | edf |
| **Glucose** | lin | 1 | 1.3 | 1.0 | 1.0 |
| **BMI** | -2 | 5 | 3.7 | 3.9 | 3.7 |
| **Pregnancies** | - | 1 | - | - | - |
| **Diabetes** | lin | 2 | 0.9 | 1.8 | 1.4 |
| **Age** | (0, 3) | 5 | 3.0 | 2.9 | 2.7 |
| **Systolic** | - | - | - | - | - |
| **Biceps** | - | - | - | - | - |
| **Insulin** | - | - | - | 0-0 | - |

Table 2: FP powers and effective degrees of freedom from applied models.

## Plot effects



**Results on PBC data**

This part can reproduce the analysis in the book. Advantages: well-known and available to all data. It is also a survival model so not just reproducing exactly the same model in different data.