Int J Public Health (2013) 58:161–165
DOI 10.1007/s00038-012-0363-z

HINTS & KINKS

Linear, nonlinear or categorical: how to treat complex associations? Splines and nonparametric approaches

Carsten Oliver Schmidt · Till Ittermann · Andrea Schulz · Hans J. Grabe · Sebastian E. Baumeister

Received: 31 January 2011 / Revised: 16 April 2012 / Accepted: 16 April 2012 / Published online: 16 May 2012

Keywords Regression models · Nonlinear modelling · Splines · Nonparametric models · Locally weighted regression

Introduction

In the first part of this methods series on nonlinear exposure-outcome relations, polynomial transformations and fractional polynomials (FP) have been introduced (Schmidt et al. 2012). They share in common a single function that covers the whole exposure range. This limits their sensitivity to capture local characteristics of the data. This article will introduce two additional approaches to handle complex nonlinear associations. A brief overview of statistical procedures in Stata, SAS and R is provided in Table 1.

Option 3: Splines

Like FPs, spline functions are common in curve fitting. However, contrary to FPs, splines are a "piece-wise" curve fitting approach. Subsequent intervals along the predictor range are modelled by different polynomials (Desquilbet and Mariotti 2010; Royston and Sauerbrei 2007; Royston and Sauerbrei 2008). Polynomials in subsequent intervals are connected by so-called "knots". The number of knots typically ranges from 3 to 8. The degree of the polynomials strongly influences the shape of the curve: linear, quadratic, and cubic spline functions are a sum of polynomials of degree 1, 2, 3, respectively. Cubic splines are most common. The slope of a linear spline function may change abruptly with each knot. Piecewise polynomials of higher than linear order allow for a smooth transition between intervals (Fig. 1a). When using 2nd or higher order polynomials the behaviour of the spline function may behave poorly in the tails. To overcome this problem restricted cubic splines (RCS, alternative term: natural splines) have been developed. This approach restricts polynomials to be linear in the tails (Desquilbet and Mariotti 2010; Royston and Sauerbrei 2007).

Adequate fine tuning is a critical issue in handling splines because this flexible approach often renders results that are hard to interpret. Selecting the appropriate number of knots and to a lesser degree their position is essential. Overfitting easily occurs in models with many knots. Deviance tests are useful to determine the number of required knots and can also be used to test splines against fractional polynomials; information criteria may be used as well (Royston and Sauerbrei 2007; Royston and Sauerbrei 2008). Contrary to FPs, zero or negative values of the exposure variable pose no need for a transformation.

One potentially confusing issue when dealing with splines is the availability of many different types and combinations of them. So far, we have dealt with cubic regression splines and natural regression splines; other types are denoted by the terms B-splines, smoothing splines, penalized splines and thin-plate splines (Keele 2008), some of which extend to the realm of nonparametric options introduced below. Compared with FPs there is considerably less agreement regarding the appropriate model selection. Splines may be

C. O. Schmidt (✉) · T. Ittermann · S. E. Baumeister
Institute for Community Medicine, SHIP-KEF,
University Medicine of Greifswald, Walther Rathenau Str. 48,
17475 Greifswald, Germany
e-mail: carsten.schmidt@uni-greifswald.de

A. Schulz · H. J. Grabe
Department of Psychiatry and Psychotherapy, SHIP-LEGEND,
University of Greifswald, Greifswald, Germany

Table 1 Selected commands for splines and nonparametric methods in Stata, SAS, and R

Command (Program)	Description
mkspline/mkspline 2 (STATA)	These commands are used to create a basis for restricted cubic splines
rcspline (STATA)	
RCS (SAS), ns (R)	
bspline (STATA),	These commands generate a basis of B-splines
B-spline (SAS), bs (R)	
uvrs (STATA)	Implements univariate regression splines comparable to fracpoly for fractional polynomials of degree 0, 1, and 3
mvrs (STATA)	Extends uvrs to allow for a multivariable model-building approach for regression spline models comparable to mfp for fractional polynomials
lowess, mlowess (STATA)	Procedures to calculate locally weighted regression models. mlowess allows for multivariable models
lowess (R), loess (SAS)	
gam (STATA)	Procedures to calculate generalized additive models. The R package "mgcv" is the most versatile. The stata "gam" command relies on cubic smoothing splines as implemented in the gamfit Fortran program by Hastie and Tibshirani (1990)
Proc GAM (SAS)	
mgcv-package (R)	
mvss (STATA)	Implements a multivariable model-building approach for generalized additive models based on the Stata gam command comparable to mfp for fractional polynomials. mvss is available from http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/(Royston and Sauerbrei 2008)

combined with FPs to assess and improve model fit (Binder and Sauerbrei 2010; Royston and Sauerbrei 2008).

Option 4: nonparametric functions

All previously introduced options described parametric approaches to treat nonlinearities. This means that an explicit regression function is defined which determines the flexibility of the curve. To overcome this limitation, nonparametric approaches have been developed. No a priori assumption is made about the shape of the curve. One basic nonparametric approach is the moving linear regression smoother (LOWESS or LOESS) (Beck and Jackman 1998; Cleveland and Devlin 1988):

Basically, in LOWESS, a separate regression analysis is conducted for each single data point across the exposure range. Low-degree polynomials are used to describe associations between the continuous predictor and the dependent variable. A user-specified number of the nearest neighbours of a reference point on the exposure variable is included in the regression model. To achieve this, commonly a fixed proportion of the sample to be included is specified. This proportion is denoted "bandwidth". Each data point within the bandwidth is assigned a weight that depends on its distance from the reference point. The smaller the distance, the higher the weight. The bandwidth is critical in obtaining a good fit. If it is too wide, the smoother lacks flexibility in capturing nonlinearities. If the bandwidth is too small, the curve may display an excess variability and overfitting is likely.

Generalized additive models are another important nonparametric approach (Beck and Jackman 1998; Hastie and Tibshirani 1990; Wood 2006). Their presentation and use as smoothing splines (Keele 2008; Royston and Sauerbrei 2008) are beyond the scope of this paper; however, some commands are provided in Table 1.

Example

We applied spline functions and LOWESS to our item of interest on not having had enough to eat during childhood, an item of the Childhood Trauma Questionnaire (Wingenfeld et al. 2010). The data base is SHIP-Legende, a population-based study comprising 2,025 adults aged 29–85 (Volzke et al. 2010).

Figure 1a displays the results for linear, quadratic and cubic splines. All splines follow the expected curvilinear relationship with a peak of self-reported insufficient nutrition among those experiencing their childhood in the late or post World-War II period while differences among models remained regarding the steepness of the ascent. The example illustrates that the placement of knots may have some important consequences. This becomes obvious when comparing RCS with default and user-specified knots. The latter were based on theoretical expectations regarding the shape of the curve and allowed for a much steeper ascent. Inspection of smoothed residuals revealed substantial misfit for the 3rd degree FP but not for the RCS, indicating a superior performance of the latter (Fig. 2).

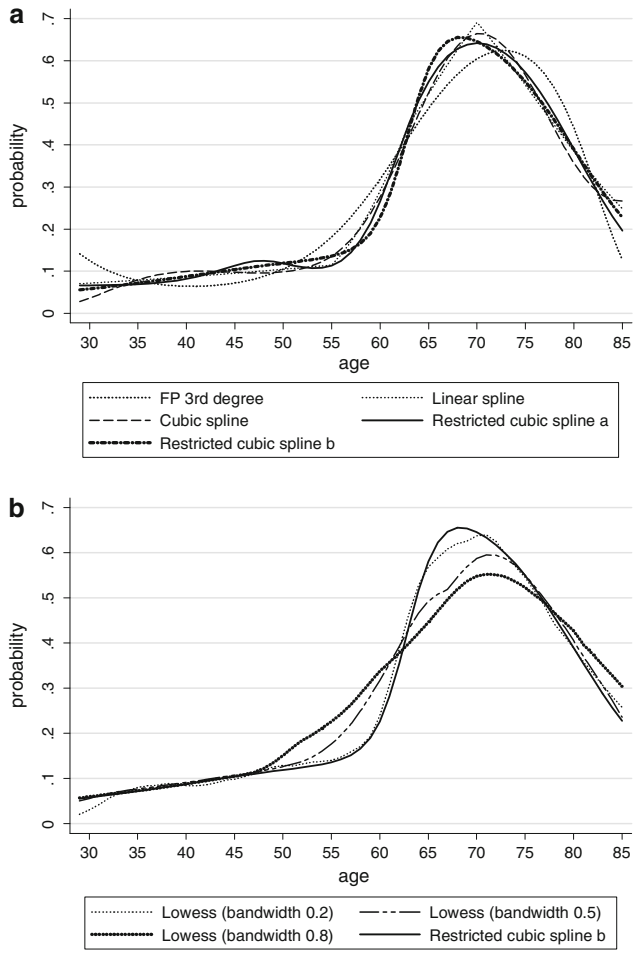

Fig. 1 **a** Using linear and cubic splines to model age effects on "not having enough to eat" during childhood, Study of Health in Pomerania, Pomerania, 2007–2010. Probabilities were derived from logistic regression models. The Stata "bspline" command was used for the modelling of the linear and cubic spline, using five knots (29, 55, 63, 70, 85). For the restricted cubic spline (**a**), an initial number of ten knots was allowed. These knots were by default placed at equally spaced centiles of the distribution of the predictor. Five knots (41, 48, 56, 63, 70) were retained based on deviance tests with an alpha level set to 0.2. The second restricted cubic spline (b) used five user-specified knots that were placed across the expected range of maximum change (50, 55, 60, 65, 70). This "fine tuning" provided a better fit and indicated that the lower maximum of the first restricted cubic spline around the age of 48 is likely to be spurious. Adding more knots did not improve fit any more. Results of the 3rd degree fractional polynomial with powers 3, 3, 3 are displayed as a reference. Akaike Information Criteria (AIC): a 3rd degree FP: 1961.47; b Linear Spline: 1943.97; c Cubic spline: 1941.89; d Restricted Cubic spline a: 1937.92; e Restricted Cubic spline b: 1934.37. **b** Using locally weighted regression to model age effects on "not having enough to eat" during childhood, Study of Health in Pomerania, Pomerania, 2007–2010. All calculations were based on locally weighted regression models as implemented in the Stata "lowess" command. *The bandwidth* refers to the proportion of sample that was used for calculating smoothed values for each point. The restricted cubic spline is displayed as a reference and shows strong similarities to the LOWESS graph with the lowest bandwidth

LOWESS indicated the expected curvilinear association as well (Fig. 1b). Smaller bandwidths entailed much steeper ascents compared with solutions with larger bandwidths. All curves remained smooth. This indicates a superior performance of small compared with large bandwidths.

While the correct solution is unknown, the RCS with five user-specified knots or the small bandwidth LOWESS regressions seem to yield the most plausible results. This is supported by the information criteria as displayed below the figures. The better performance of models that were particularly suitable to capture local features of the data seems conceptually plausible, given the fast onset of nutrition deficiencies during late or post World-War II. It is important to note that no general conclusion on the usability of the introduced methods is intended based on our example.

Conclusion

Powerful approaches to deal with nonlinearities have briefly been highlighted. Based on the introduced literature, some general considerations are

- Splines and nonparametric functions to model nonlinearities on the one hand can be contrasted with FPs on the other hand. The former are more flexible in handling local features of the data while the latter focus on global nonlinear effects. Under most applied circumstances, FPs provide stable results, and 1st or 2nd degree FPs suffice, as shown in the second example of the first paper. In case of our data set, which included more than 20 CTQ items, results from FPs and Splines were similar in most cases.
- Within each class of models, the challenge is to adequately select parameters that determine the shape of the curve. A typical choice regarding FPs is power $\in (-2, -1, -0.5, 0, 0.5, 1, 2, 3)$; degree = 2 (=4 *df*); $\alpha = 0.05$. Regarding splines, the degree of the splines has to be specified along the number of knots, their location and a criterion for the selection and comparison of models, e.g. an alpha level for deviance testing. Less agreement exists on these selection criteria. With "uvrs", a Stata command that implements RCS, the default is a 3rd degree spline, 3 interior knots (=4 *df*), $\alpha = 0.05$,++ and a placement of knots at equally spaced centiles of the exposure variable. It is important to note that many more options are available to build spline functions (Desquilbet and Mariotti 2010; Keele 2008).
- Generally, the type and flexibility of the models should correspond to the expected nonlinearities. Complex

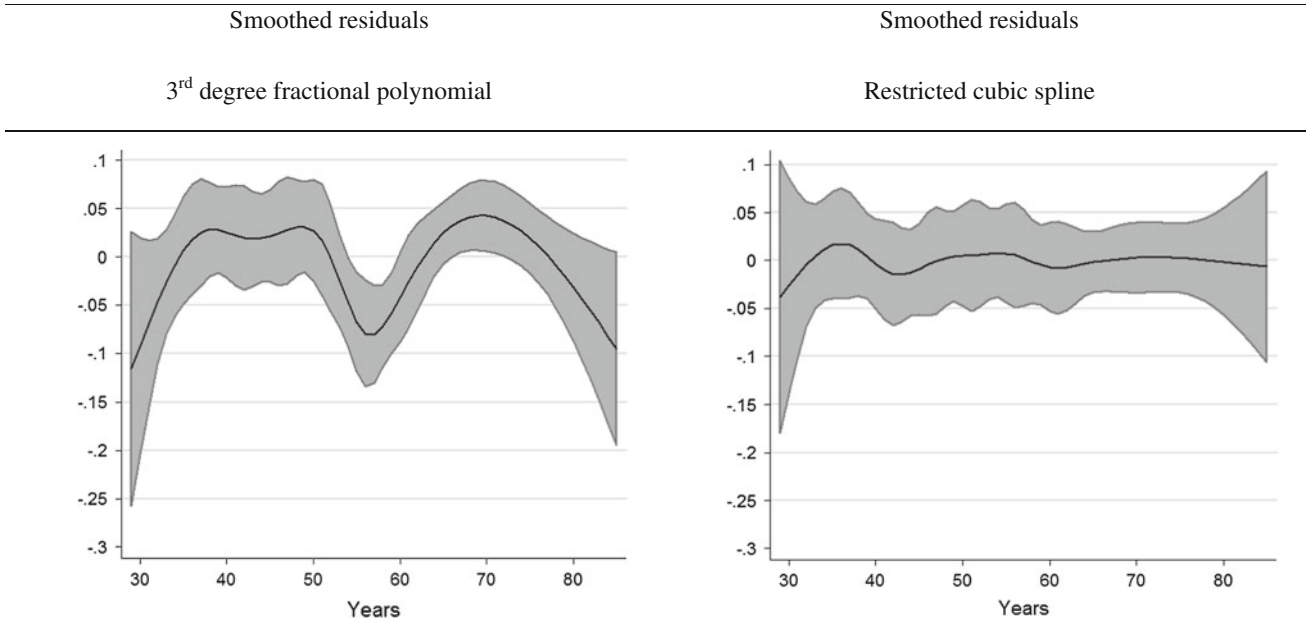

Fig. 2 Smoothed residuals from third degree fractional polynomial (FP) and restricted cubic spline (RCS), item "not having enough to eat", Study of Health in Pomerania, Pomerania, 2007–2010. Raw residuals were computed from logistic regression analyses. The FP and the RCS with user specified knots were taken from Fig. 1a. A RCS with 10 *df* was used to predict raw residuals. *The shaded area* around the line displays the 95 % CI. In case of the 3rd degree FP it illustrates systematic lack of fit especially in subjects between 50 and 65 years of age, whereas this was not the case with residuals computed from the RCS. Similar conclusions can be drawn based on nonparametric approaches to smooth residuals, but are not displayed

models, needing many degrees of freedom, may improve prediction but increase the risk of overfitting the data and should be avoided unless necessary to improve the transferability and interpretability of results. How to balance prediction and interpretability strongly depends on the research aims.

- The measured exposure range must contain the area with nonlinearities as well as sufficient cases within this area to reliably detect effects.
- Outliers may exert a substantial influence on associations but adequate transformations may reduce associated problems.
- Residual analysis is essential. A useful technique to obtain informative graphics on potential model deficiencies is to smooth residuals, see also Fig. 2.
- Bootstrap can be applied to assess replication stability of selected models (Royston and Sauerbrei 2008).
- Elaborated procedures for multivariable model building are available. One example is the Stata "mfp" commands for FPs, "mvrs" for RCS, and "mvss" for non-parametric smoothing splines (Royston and Sauerbrei 2008; Sauerbrei et al. 2007). All share in common a backward elimination strategy for the selection of variables based on pre-specified *p* values. Splines and FPs have recently been compared in a simulation study on multivariable model building indicating advantages of FPs over splines, yet neither approach performed best across all scenarios (Binder et al. 2011).

Acknowledgments SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs and the Social Ministry of the Federal State of Mecklenburg-West Pomerania. This work was also funded by the German Research Foundation (DFG: GR 1912/5-1).

References

Beck N, Jackman S (1998) Beyond linearity by default: generalized additive models. Am J Political Sci 42(2):596–627

Binder H, Sauerbrei W (2010) Adding local components to global functions for continuous covariates in multivariable regression modeling. Stat Med 29(7–8):808–817

Binder H, Sauerbrei W, Royston P (2011) Multivariable model-building with continuous covariates: 2. Comparison between splines and fractional polynomials, vol 106. Institut für Medizinische Biometrie und Medizinische Informatik, Universitätsklinikum Freiburg, Freiburg

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83(403):596–610

Desquilbet L, Mariotti F (2010) Dose-response analyses using restricted cubic spline functions in public health research. Stat Med 29(9):1037–1057

Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, London

Keele L (2008) Semiparametric regression for the social sciences. Wiley, Chichester

Royston P, Sauerbrei W (2007) Multivariable modeling with cubic regression splines: a principled approach. Stata J 7(1):45–70

Royston P, Sauerbrei W (2008) Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. Wiley, Chichester

Sauerbrei W, Royston P, Binder H (2007) Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 26(30): 5512–5528

Schmidt CO, Ittermann T, Schulz A, Grabe HJ, Baumeister SE (2012) Linear, nonlinear or categorical—how to treat complex associations in regression analyses? Polynomial transformations and fractional polynomials. Int J Public Health. doi:IJPH-11-40

Volzke H et al (2010) Cohort profile: the study of health in Pomerania. Int J Epidemiol 40(2):294–307

Wingenfeld K et al (2010) The German version of the Childhood Trauma Questionnaire (CTQ): preliminary psychometric properties. Psychother Psychosom Med Psychol 60(11):442–450

Wood S (2006) Generalized additive models: an introduction with R. Chapman & Hall, Boca Raton