Int J Public Health (2013) 58:157–160
DOI 10.1007/s00038-012-0362-0

HINTS & KINKS

Linear, nonlinear or categorical: how to treat complex associations in regression analyses? Polynomial transformations and fractional polynomials

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: 9 May 2012

Keywords Regression analysis · Nonlinear modelling · Linear models · Categorization · Polynomials · Fractional polynomials

Introduction

Nonlinear approaches to assess exposure-outcome relations are still fairly uncommon in public health research. The predominant reliance on linear associations and categorized continuous predictors is surprising, given the availability of powerful alternatives with sophisticated and user friendly software implementations. This simplicity threatens one of the major aims in regression analyses: to obtain an unbiased mean estimate of the dependent variable conditional on the predictor variables.

In the first part of this non-technical series, we will briefly discuss problems of linear models and categorized continuous predictor variables. Polynomial and fractional polynomial approaches will be introduced, as well as information on selected statistical procedures in main software packages (Table 1). In the second part, splines and non-parametric approaches will be illustrated (Schmidt et al. 2012).

Linear modeling and categorization: the common approaches

Two approaches are commonly used to handle exposure-outcome relations: first, the association is assumed to be linear. This implies that the effect on the outcome is the same across the entire exposure range. If deviations from linearity are strong, relevant associations may be missed altogether, for example with threshold, U- or J-shaped associations (May and Bigelow 2005). The high acceptance of untested linear associations reflects the technical simplicity of this approach along a lack of elaborated theory development that allows for a priori assumptions regarding the shape of dose-response relations (Becher 2005; Beck and Jackman 1998). However, on the contrary to current practice, theoretical shortcomings should encourage a careful data driven model selection.

Second, continuous predictors are frequently categorized; either based on statistical considerations like the median, quartiles and reference limits or based on substantial considerations like biological plausibility. An example for the latter is alcohol intake per day below or above WHO thresholds of risk drinking. Categorizing conforms to the clinical practice of distinguishing the presence or absence of an attribute and seems to provide simple interpretation of the results. Yet, numerous shortcomings exist (Royston et al. 2006; Royston and Sauerbrei 2008):

1. Information is lost because all cases within one category are treated as being equal.
2. Because of this simplification and due to the inflated number of parameters with many categories, statistical power is lost.
3. The assessment of interactions is complicated and unreliable.

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,
University of Greifswald, Greifswald, Germany

Table 1 Selected commands to handle fractional polynomials in STATA, SAS, and R

Command (program)	Description
fracpoly (STATA)	Conducts nonlinear analyses based on fractional polynomials for a single continuous predictor variable while controlling for covariates
fracpred (STATA)	Creates variables containing the prediction, deviance residuals, or standard errors of fitted values
mfp (STATA) mfp (SAS) mfp (R)	Implements a multivariable model-building approach for fractional polynomials. Adequate powers for all continuous predictors in the regression model are selected based on deviance tests. Options to deal with predictors with a spike at zero are available (catzero)
mfpi (STATA)	Models interactions between categorical and continuous covariates using fractional polynomials (Royston and Sauerbrei 2004)
gdelta (STATA)	Performs pre-transformations on a continuous predictor variable to be used in an analysis with fractional polynomials, for example to avoid problems with non-positive values and with extreme values at both ends of the distribution (Royston and Sauerbrei 2007)

An extensive set of STATA commands of relevance for the calculation of fractional polynomials is available from: http://www.imbi.uni-freiburg.de/biom/Royston-Sauerbrei-book/

4. Cut-off points are more or less arbitrary; yet, associations may strongly depend on the cut-offs and may vary across study populations.
5. When dichotomizing, no information is available on potential nonlinearities at all.
6. The risk of chance findings is increased in case of sparse categories.
7. Categorization entails the risk of bias resulting from misclassification (Rothman et al. 2008). This is particularly threatening when forming subgroups of substantially different size in the presence of measurement error.
8. A test for trends is no adequate substitute for an appropriate nonlinear model (Maclure and Greenland 1992).

Despite these shortcomings, a categorization may offer valuable explorative insights and should not be dismissed altogether (May and Bigelow 2005).

Option 1: polynomial transformations of the dependent variable

One basic option is to conduct a simple nonlinear transformation of the predictor variable x. Quadratic (x^2) and cubic polynomials (x^3) are frequently used for this purpose. The transformed parameter is most commonly entered into the regression model in addition to the linear parameter. If the nonlinear parameter fails to improve the prediction, this may be determined based on a deviance test (Royston and Sauerbrei 2008), a linear approach might be sufficient. Yet, some caution is indicated. A simple transformation provides limited flexibility in capturing nonlinearities. Particularly towards the end of the exposure range, a poor und implausible performance is common (May and Bigelow 2005; Royston et al. 1999).

Option 2: fractional polynomials (FPs)

An important and powerful extension to the first option is the combination of polynomial and logarithmic functions, as offered by FPs (Royston et al. 1999, 2006; Royston and Sauerbrei 2008). One main advantage is a much wider set of functional forms. Again, a single function covers the whole exposure range. For many applications, first degree FPs suffice. These take on the form $\beta_1 x^p$, with x being the predictor of interest, β_1 the beta coefficient and p the power of the polynomial, which may be fractional. The power is commonly chosen from -2, -1, -0.5, 0, 0.5, 1, 2, and 3. For example, if $p = 2$, the term resembles a quadratic polynomial, whereas $p = 0.5$ refers to the square root and $p = 1$ to a linear function. By convention, x^0 denotes log x. Depending on theoretical considerations, higher degree FPs may be advisable. On contrary to the monotonous first degree FPs, second degree FPs allow for one maximum or minimum. Second degree FPs take one of two forms: (1) $\beta_1 x^p + \beta_2 x^q$ or (2) $\beta_1 x^p + \beta_2 x^p \times \log(x)$. The idea is to find a combination of powers p and q that optimize model fit. The most common option to accomplish this is to compare different models based on deviance tests, using pre-specified p values. Information criteria are another but less frequently applied option (Burnham and Anderson 2004; Royston and Sauerbrei 2008).

Powerful commands (mfp in Stata/SAS/R) are available to set up and select multivariable models comprising several nonlinear associations (Sauerbrei et al. 2006). The user specifies two significance levels for an automatic selection procedure: one for the selection of variables and one for the selection of powers. A backwards elimination strategy is used to drop variables from the model yielding a non-significant increase in deviance. A detailed description of the test procedure is available elsewhere (Royston and Sauerbrei 2007, 2008).

A disadvantage of FPs is the inability to handle non-positive predictor values, but suitable transformations are available (Royston and Sauerbrei 2008), as well as options to deal with predictor variables with a spike at zero (Royston and Sauerbrei 2008; Royston et al. 2010) and interactions (Royston and Sauerbrei 2004).

Example

All approaches have been applied to cross-sectional data from a general population survey, SHIP-Legende, in West Pomerania, Germany. Data comprise 2025 adults aged 29–85 (Volzke et al. 2010). We predicted an affirmative response to the questions “When I was growing up, I did not have enough to eat”, and “People in my family called me things like stupid, lazy, or ugly”, both items of the Childhood Traumata Questionnaire (Wingenfeld et al. 2010).

As we applied logistic regression, it is important to emphasize the distinction between nonlinear associations and nonlinear statistical models. We used a nonlinear statistical model to assess linear and nonlinear associations between age and CTQ items.

Results illustrated strong disparities between the established exposure-outcome relations on the nutrition item. Categorical representation of age showed curvilinear relation when using deciles, but not when using quintiles (Fig. 1a). A quadratic power transformation of age suggested a particularly strong ascent of nutrition deficits among the eldest (Fig. 1b).

Fractional polynomials yielded different results, depending on their degree (Fig. 1b). First degree FPs describe a monotonous increase. The second and third degree FPs indicated a curvilinear association, with the third degree FP having a steeper ascent and a better fit according to information criteria (Fig. 1b). Deviance residuals of second and to a lesser extent third degree FP revealed substantial misfit (Fig. 2).

Regarding the second item on being called stupid, the results across methods were more similar. Little is gained by applying higher degree FPs, which is reflected in the selected powers for FPs based on deviance tests and by the larger Akaike information criteria for models requiring more degrees of freedom compared to a first degree FP.

The behaviour of the different approaches addressing nonlinearities in the latter example is typical for public health applications, whereas the former indicates the need to address strong local changes to the shape of the curve in some situations. We know that only subjects within a narrowly defined age range were subject to the exposure “nutrition shortages during childhood” after World War 2. The second article in this series will therefore introduce

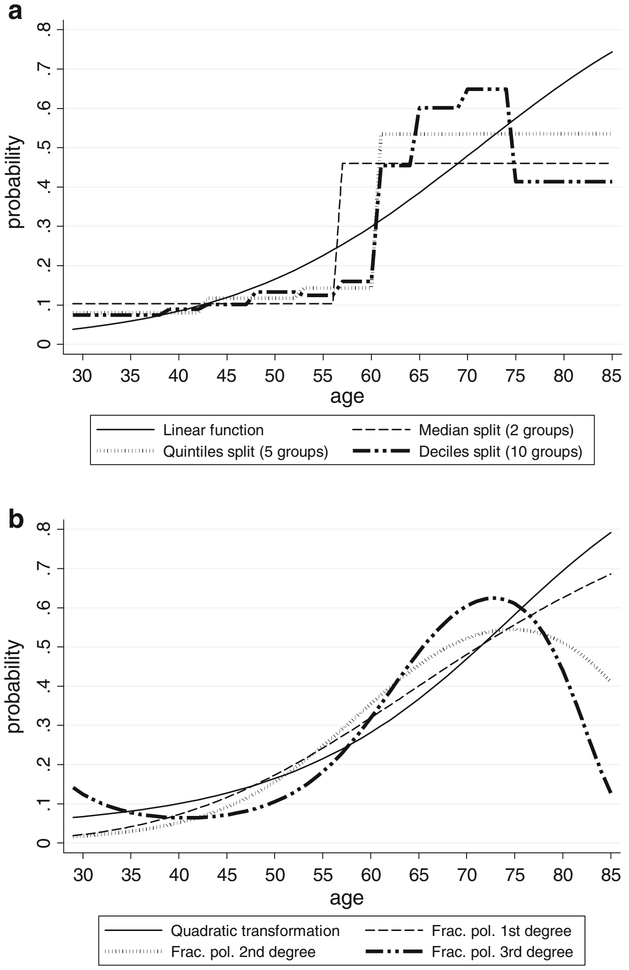

Fig. 1 **a** Age as a linear and categorized predictor for “not having enough to eat”, Study of Health in Pomerania, West Pomerania, 2007–2010. Probabilities were derived from logistic regression models. Akaike Information Criteria (AIC): (1) linear function: 2060.63; (2) Median split: 2070.5; (3) Quintiles split: 1974.10; (4) Deciles split: 1951.56. **b** Using different approaches to model age effects on “not having enough to eat” during childhood, Study of Health in Pomerania, West Pomerania, 2007–2010. Probabilities were derived from logistic regression models. Calculations for the fractional polynomials (FPs) were first performed with the STATA fracpoly command. The selection of fractional polynomials was based on the maximum deviance difference compared to the fit of a straight line. The powers were selected from the default setting (−2, −1, −0.5, 0, 0.5, 1, 2, 3), resulting in: first degree: 0.5; second degree: 3,3; third degree: 3,3,3. When selecting fractional polynomials based on the implemented closed test procedure in the STATA mfp command (Royston and Sauerbrei 2008; Sauerbrei et al. 2006) with α set to 0.1 for the selection of powers, the same powers emerged as described before. Akaike Information Criteria (AIC): (1) Quadratic transformation: 2079.74; (2) first degree FP: 2054.54; (3) second degree FP: 2015.20; (4) third degree FP: 1961.47

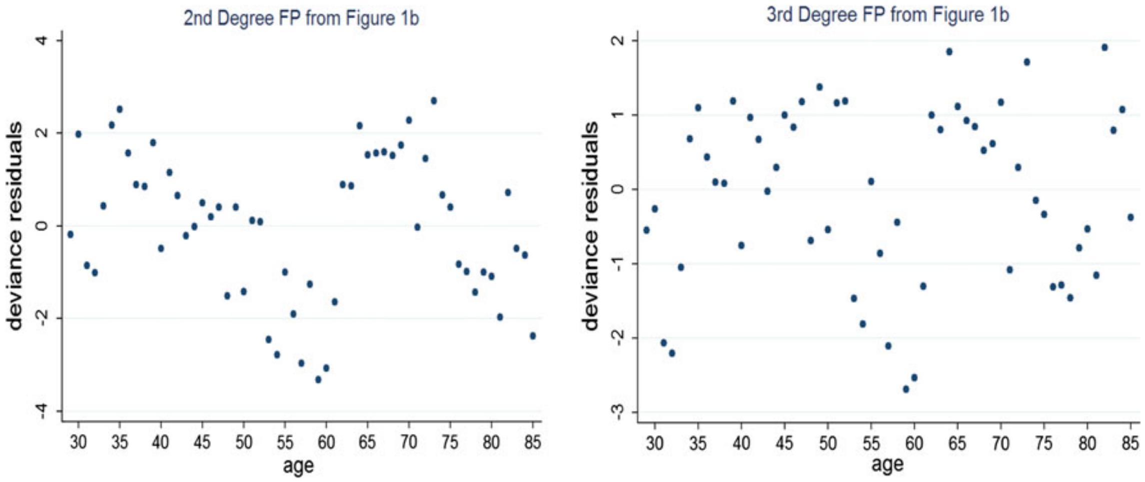

Fig. 2 Deviance residuals for the second degree fractional polynomials (3,3) and third degree fractional polynomials (3,3,3) for the age as predictors for "not having enough to eat", Study of Health in Pomerania, West Pomerania, 2007–2010. Fractional polynomials (FPs) deviances are plotted against age and correspond to the respective logistic regression models

nonlinear modelling approaches that are particularly suitable to capture local properties of the data.

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

Becher H (2005) General principles of data analysis: continuous covariables in epidemiological studies. In: Ahrens W, Pigeot I (eds) Handbook of epidemiology. Springer, Heidelberg, p 595–624

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

Burnham KP, Anderson DR (2004) Multimodel inference—understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304

Maclure M, Greenland S (1992) Tests for trend and dose response: misinterpretations and alternatives. AmJ Epidemiol 135(1):96–104

May S, Bigelow C (2005) Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges. Dose Response 3(4):474–490

Rothman KJ, Greenland S, Lash T (2008) Modern epidemiology, vol 4. Lippincott Williams & Wilkins, Philadelphia

Royston P, Sauerbrei W (2004) A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 23(16):2509–2525

Royston P, Sauerbrei W (2007) Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. Comput Stat Data Anal 51(9):4240–4253

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

Royston P, Ambler G, Sauerbrei W (1999) The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol 28(5):964–974

Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25(1):127–141

Royston P, Sauerbrei W, Becher H (2010) Modelling continuous exposures with a 'spike' at zero: a new procedure based on fractional polynomials. Stat Med 29(11):1219–1227

Sauerbrei W, Meier-Hirmer C, Benner A, Royston P (2006) Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. Comput Stat Data Anal 50(12):3464–3485

Schmidt CO, Ittermann T, Schulz A, Grabe HJ, Baumeister SE (2012) Linear, nonlinear or categorical: how to treat complex associations in regression analyses? Splines and nonparametric approaches. Int J Public Health (IJPH-11-41)

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