UNIVERSITÄTS KLINIKUM FREIBURG

Lecture 3
Continuous variables – to categorize or to model with fractional polynomials?

Willi Sauerbrei, Edwin Kipruto

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center -University of Freiburg, Germany

WS 2025/26

Overview

- Motivation to select a suitable function
- Short introduction to multivariable regression models
- Approaches to handle one continuous variables
  - Categorization
  - Fractional polynomials (FPs)
  - FP function selection procedure (FSP)
- Comparison of approaches in an example

Learning objectives:

Using cutpoints to categorize a continous variable can introduce severe problems.

The linearity assumption is often acceptable but can also be seriously wrong.

Fractional polynomials are a simple approach to estimate the functional form.

Preliminaries

- No small sample size
- Consider only the univariate situation, extension for multivariable models in a follower lecture
- Often we refer to the book

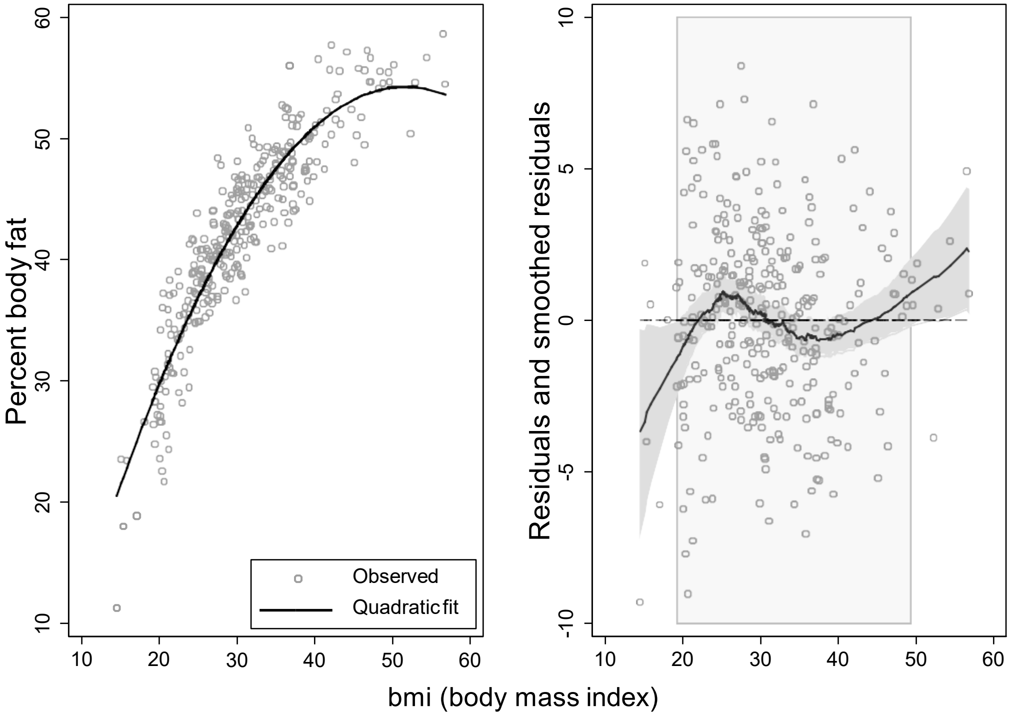
Royston and Sauerbrei (2008), Multivariable Model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables

- https://mfp.imbi.uni-freiburg.de/ all data sets are available

Relationship of percent body fat and BMI (N=326*)

The linearity assumptions is often acceptable but can also be seriously wrong.
Fractional polynomials are a simple approach to derive the functional relationship.

Quadratic model fits the data poorly



* Percentage body fat, Black People, Origin from three different countries [Luke et al (1997), see R+S(2008, A.2.1) for more information]

R&S 2008, Fig. 1.1

Drawback of the quadratic model

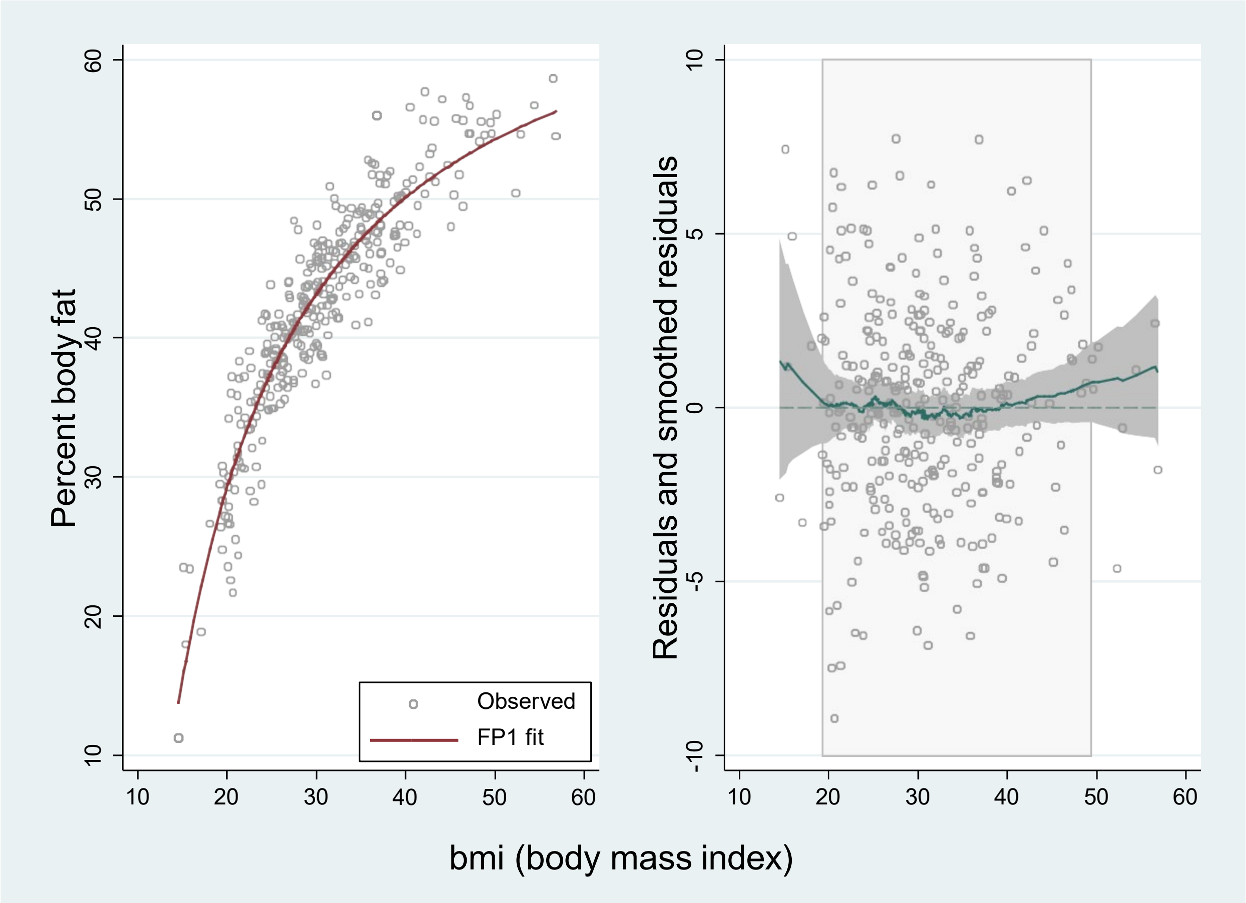
Quadratic model ($y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$) has at least two critical issues

- Turning point around 50 for bmi-makes no sense
- Lack of fit for very low and large values

→ Good illustrated with smoothed residuals

A possible solution

Fractional polynomial does better



Regression models

- Statistical method to investigate the association between a response variable Y and one or more explanatory variables $X_1, \ldots, X_k$
- Response variable Y may be continuous, binary or a survival time (partly censored).
- Typical handling of a continuous variable: categorization or assume linearity
- $X_1, \ldots, X_k$ may be risk factors, prognostic factors, diagnostic criteria etc.

Regression models

Usually several factors have an influence on the outcome.

Regression models are the **key tool** for the analysis of most of these studies. However, several alternatives such as **trees** and **neural nets** are available.

Outcome and regression models

Different types of regression models for different outcomes

Linear Model: Y continuous

- Simple linear regression: $y = \beta_0 + \beta_1 X_1 + \epsilon$
- Multiple linear regression: $y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$

Logistic Regression: Y binary response

Specification of a meaningful association between Y and X via the logit link function

Model: $$logit(\mathrm{p}) = \log\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

$$\Leftrightarrow \mathrm{p} = \left.\frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}\quad\right\}\text{“logistic“ function}$$

β_i is logistic regression coefficient (adjusted for other variables)

$\exp(\beta_i) =$ Odds Ratio (adjusted)

Cox-Model – Survival Data

Extension of the logistic regression model for survival analysis

Consider intensity of the occurrence of an event (hazard rate) at time t:

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} P(t < T \leq t + h | T > t)$$

Assumption:

For two groups of patients A and B the hazard rates are proportional to each other:

$$\frac{\lambda_B(t)}{\lambda_A(t)} = constant = HR\ (Hazard\ Ratio)$$

For k variables: $X_1, \ldots, X_k$

$$\lambda(t) = \lambda_0(t) \cdot \exp(\beta_1 X_1 + \cdots + \beta_k X_k)$$

Strategies for model building

- Strategies for model building are often developed for the multiple linear regression model and extended to other types of models.
- Some adaptations may be needed but key principles can be used in generalized linear models (GLM).

In multivariable modeling the key questions are:

  - Which of $X_1, \ldots, X_k$ to include?
  - Which functional form for continuous variables?

- More about multivariable model building in the **MFP lecture**.
- Rest of the talk on univariate investigation of the functional form

Some approaches to modelling a continuous variable

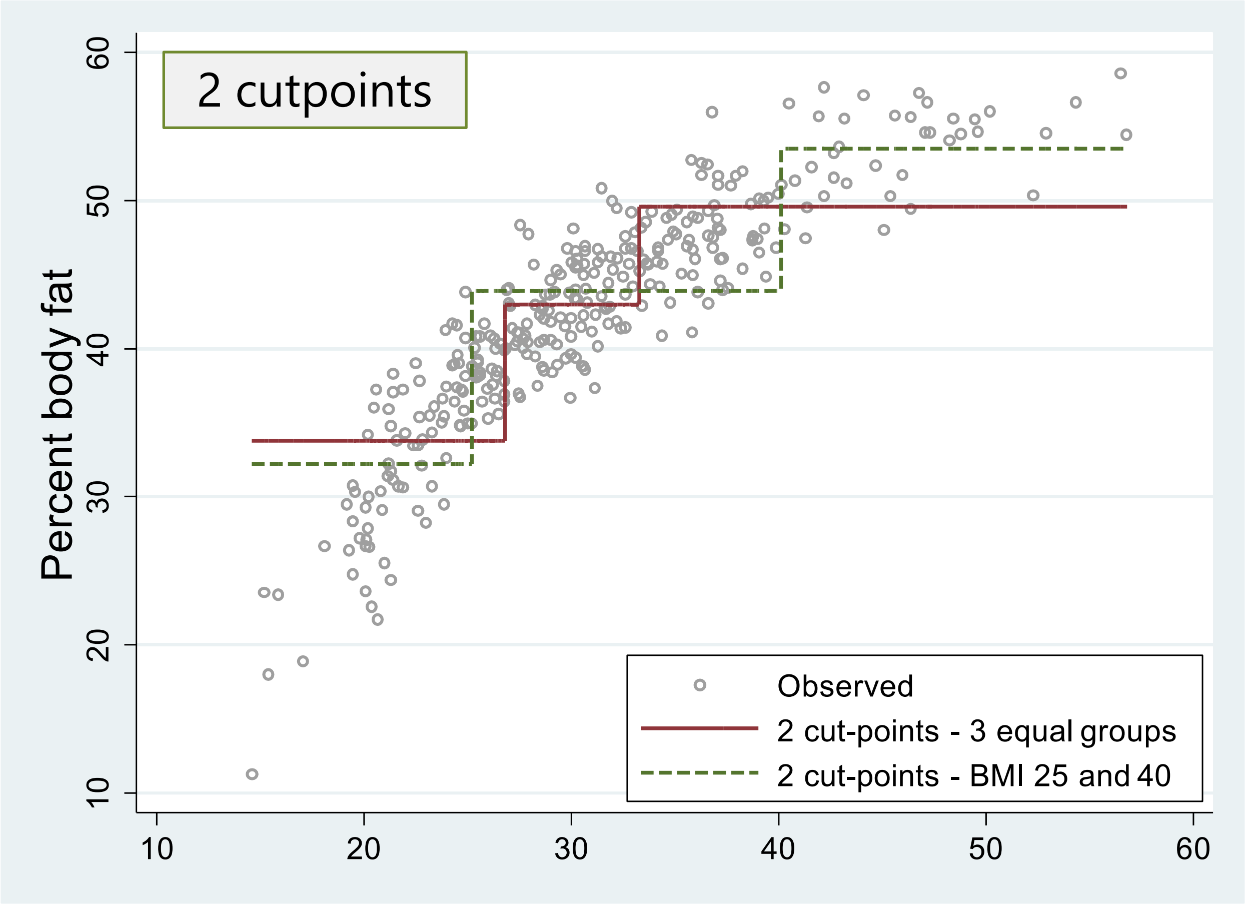
- Assume linearity
- Cut-points
- ‘Optimal’ cut-points
- Fractional polynomials
- Splines

Models based on cut-points…

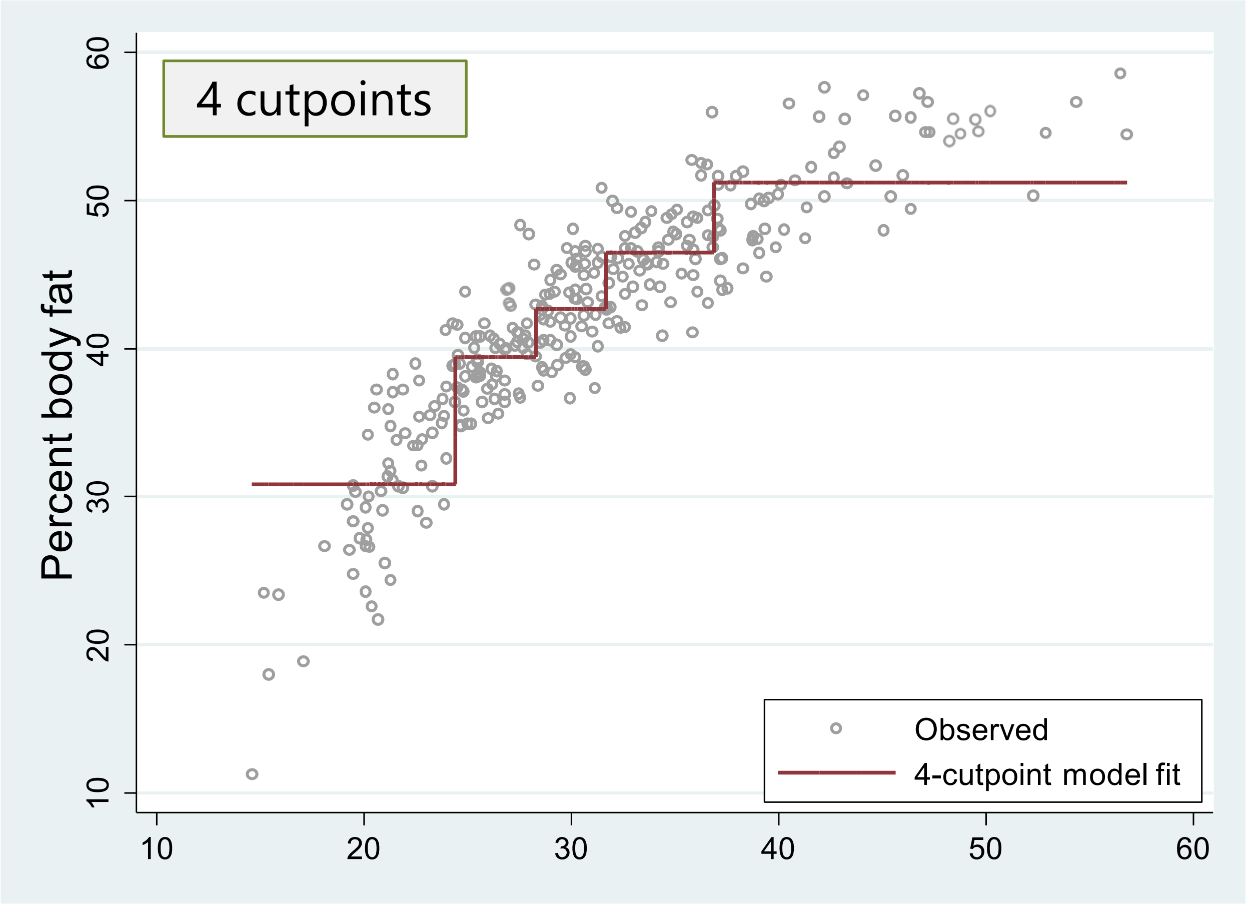
- Cut-points are still popular in clinical and epidemiological research
- Use of cut-points in a model gives a **step function**
- How many cut-points?
- Where should the cut-points be put?
- Poor approximation to the true relationship
- Almost always fits the data less well than a suitable continuous function

… have severe problems

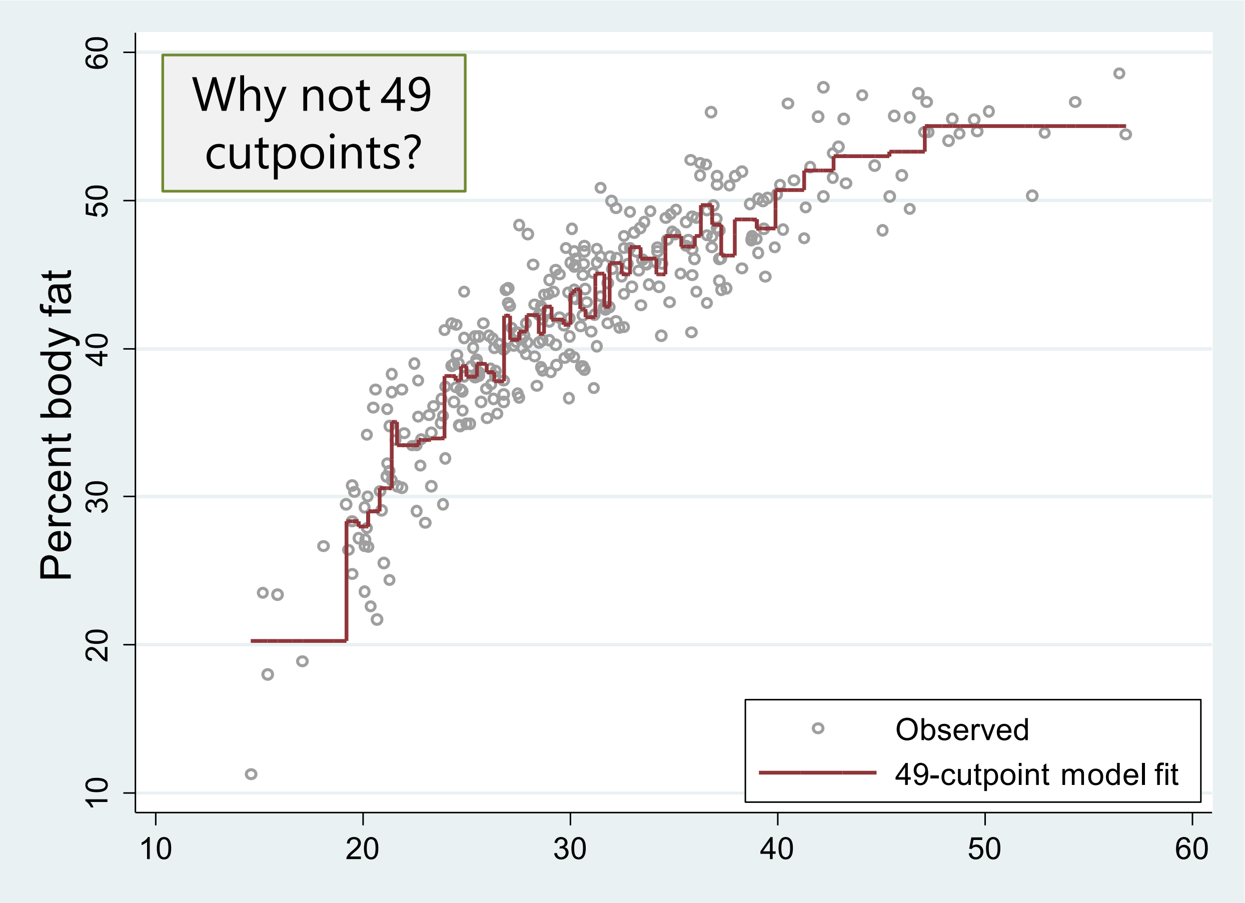
Cut-points and the body fat data (1)



Cut-points and the body fat data (2)



Cut-points and the body fat data (2)

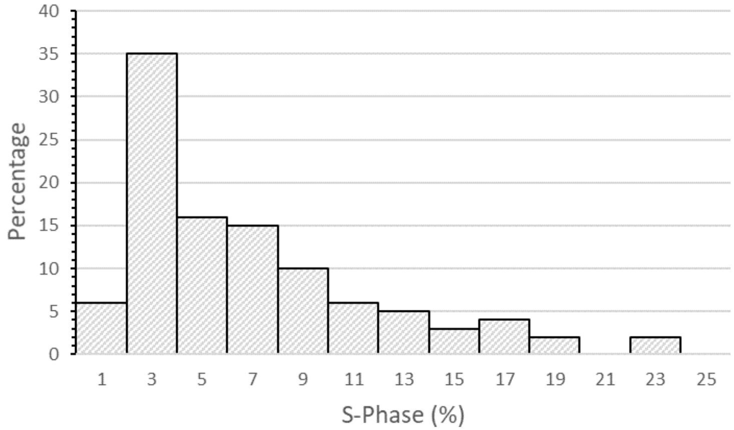


Why not 49 cutpoints?

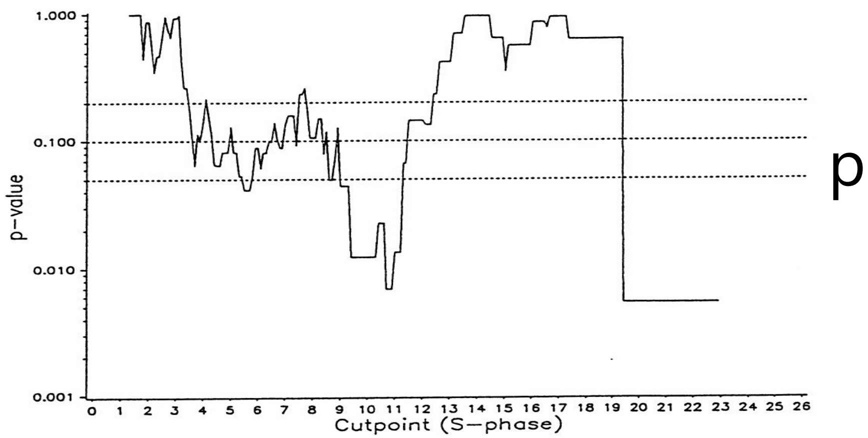
‘Optimal’ cut-point - minimal P-value approach

Search for the cutpoint which best separates two groups

Altman *et al.* (1994)



Is SPF a prognostic factor in breat cancer?



p-value as a function of the SPF cutpoint

‘Example: Prognostic factors

GBSG-study in node-positive breast cancer

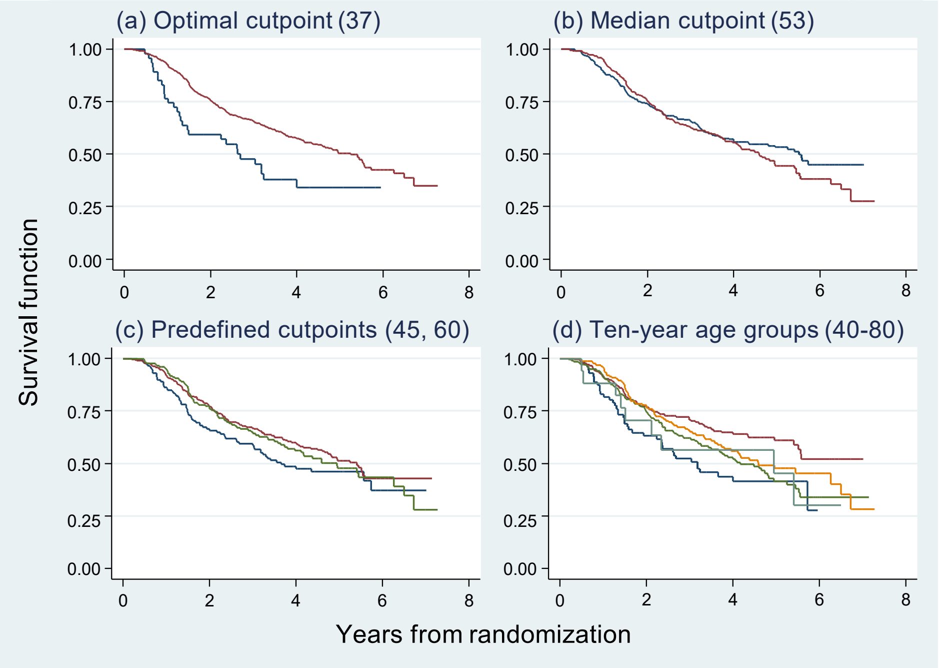
299 events for recurrence-free survival time (RFS) in
686 patients with complete data

7 prognostic factors, of which **5** are continuous

Treatment with Tamoxifen (yes/no)

Is age a prognostic factor?

Age as prognostic factor – cutpoint analyses



Sauerbrei, W., & Royston, P. (2010). Continuous variables: to categorize or to model.

The youngest group is always in blue.

(a) 'Optimal' (37 years); HR (older vs younger) 0.54, p= 0.004

(b) median (53 years); HR (older vs younger) 1.1, p= 0.4

(c) predefined from earlier analyses (45, 60years) p= 0.20

(d) popular (10-year groups) p>0.5

Optimal cutpoints introduce severe problems!

- Multiple testing → inflation of significance level
  - about 40% instead of nominal 5%
- Inflated significance level does not disappear with increased sample size
- Large bias in estimate of difference between groups
- Results depend on chance
- Different cutpoints in each data set-impossible to summarize across studies

Dichotomizing continuous predictors in multiple regression:
a bad idea

Patrick Royston[1,*,†], Douglas G. Altman[2] and Willi Sauerbrei[3]

StatMed 2006, 25:127-141

Fractional polynomial models – FP1

- FP1 models include power transformation of X
  - Model is $\beta_1 X^{p_1}$
  - Powers (p_1) are chosen from a restricted set
  - 1/X^2, 1/X, 1/√X, log X, √ X, X, X^2, X^3
  - For example β_1(1/X) has power $p_1 = -1$
  - FP1 'powers' are −2, −1, −½, 0 (= log), ½, 1 (= linear), 2, 3
- 8 models defined by 8 power terms
- X has to be positive (use simple transformation otherwise)

Fractional polynomial models – FP2

- FP2 models have 2 powers p_1, p_2
  - $\beta_1 X^{p1} + \beta_2 X^{p2}$
  - For example $\beta_1(1/X) + \beta_2(X^2)$ has powers $p_1 = -1, p_2 = 2$
  - 28 models with combinations of p_1 and p_2, where $p_1 \neq p_2$
- Also 'repeated powers' models, $p_1 = p_2$
  - $\beta_1 X^{p1} + \beta_2 X^{p1} \log(X)$
  - E.g. $\beta_1(1/X) + \beta_2(1/X)\log X$
  - This model has powers −1, −1
  - 8 models with repeated powers
- FP1 and FP2 together = 44 combinations of powers

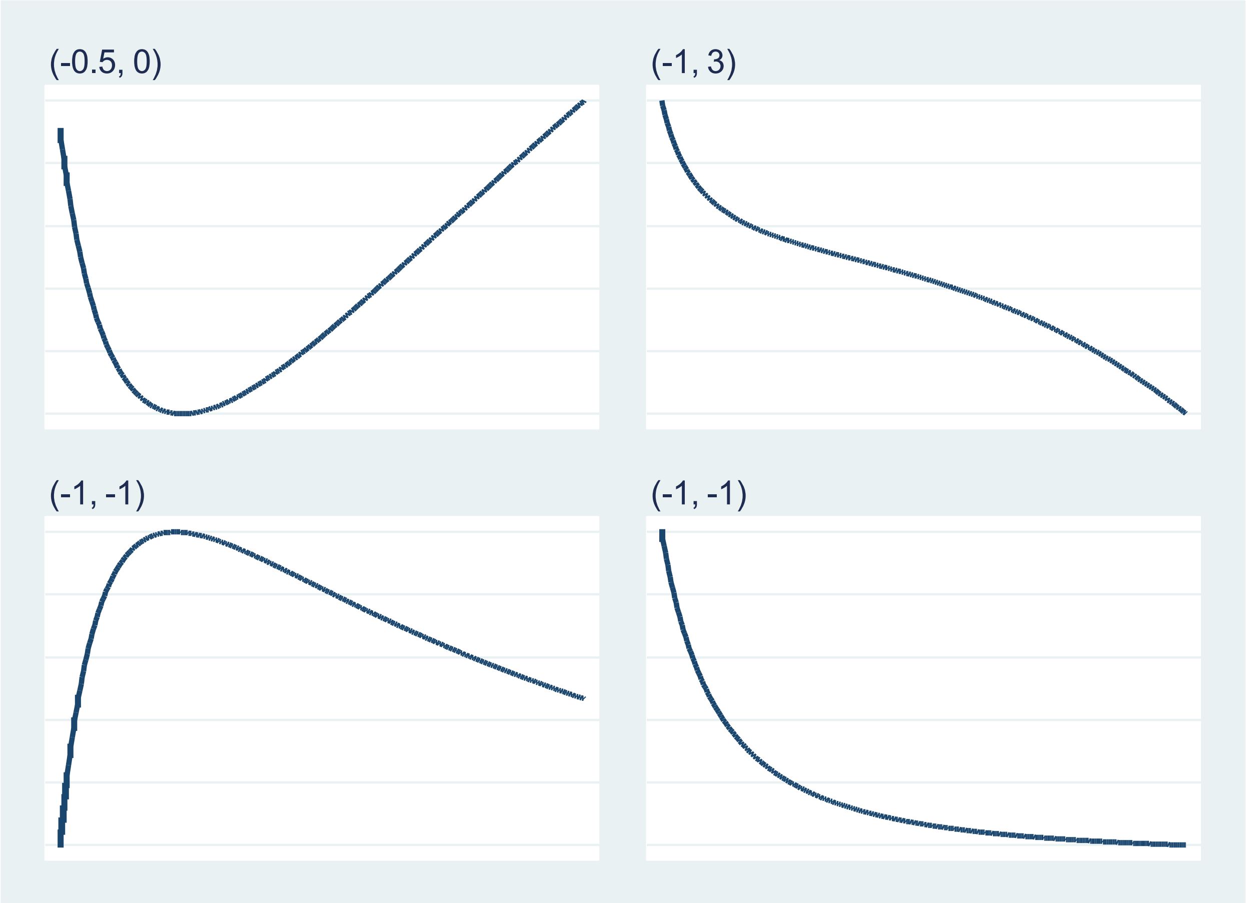
Fractional polynomial models – more complex?

- FPm models, m > 2

$$\text{FP}m(X) = \beta_1 X^{p1} + \beta_2 X^{p2} + \ldots + \beta_m X^{pm}$$

→ hardly needed in medical applications

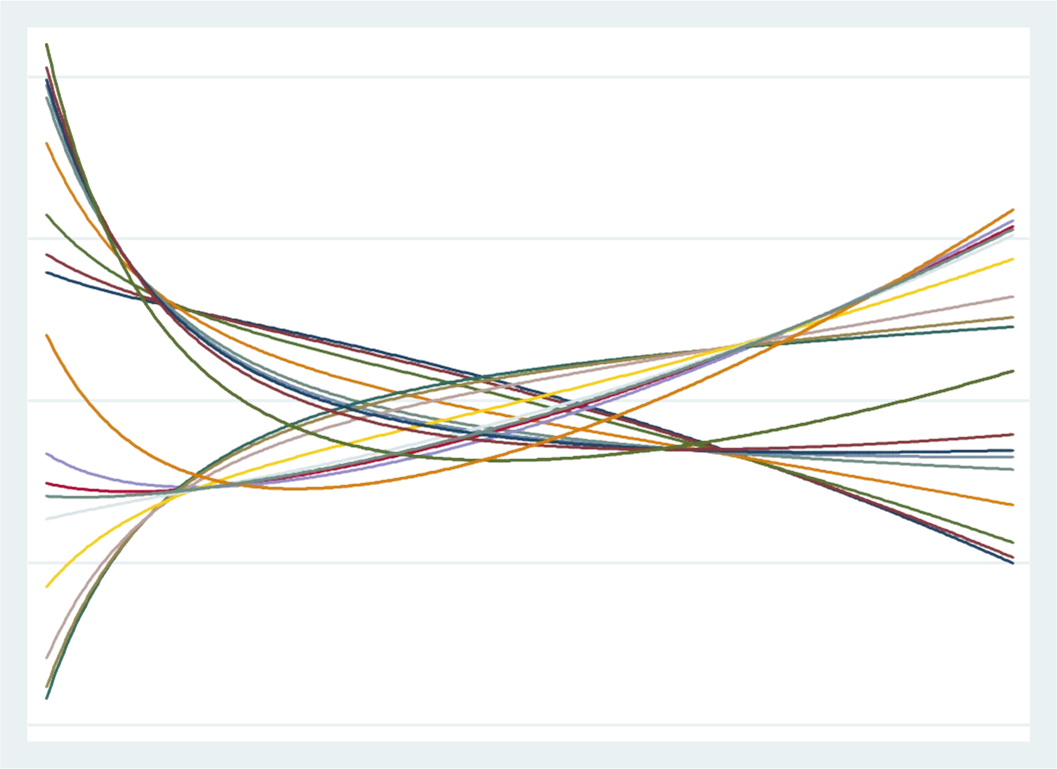
Some FP2 curve shapes (with varying powers)



R&S 2008, Fig.4.4

Some more FP2 curve shapes

fixed powers, different β's



R&S 2008, Fig.4.5

Function Selection Procedure (FSP) for FP2

The FP function is determined for the variable x using the following closed test procedure:

Any effect of X?

1. Test the **best FP2** model for x at the α significance level against the **null model** using **4 d.f**. If the test is not significant, stop, concluding that the effect of x is "not significant" at the α level. Otherwise continue.

Linear function sufficient?

2. Test the **best FP2** for x against the **default** (usually a linear function) at the α level using **3 d.f**. If the test is not significant, stop, the final model being the default. Otherwise continue.

FP1 sufficient or FP2 needed?

3. Test the **best FP2** for x against the **best FP1** at the α level using **2 d.f**. FP2 selects two power terms and estimates two corresponding parameters, therefore 4 d.f.; correspondingly FP1 has 2 d.f., giving a difference of two d.f. - If the test is significant, the final model is the best FP2, otherwise the final model is the best FP1. End of procedure.

R&S 2008, Fig.4.5

An example of FP model fitting

Is age a prognostic factor in breast cancer?

FP1		FP2								
Power	Model chi-square	Powers		Model chi-square	Powers		Model chi-square	Powers		Model chi-square
-2	6.41	-2	-2	17.09	-1	1	15.56	0	2	11.45
-1	3.39	-2	-1	17.57	-1	2	13.99	0	3	9.61
-0.5	2.32	-2	-0.5	17.61	-1	3	12.37	0.5	0.5	13.37
0	1.53	-2	0	17.52	-0.5	-0.5	16.82	0.5	1	12.29
0.5	0.97	-2	0.5	17.30	-0.5	0	16.18	0.5	2	10.19
1	0.58	-2	1	16.97	-0.5	0.5	15.41	0.5	3	8.32
2	0.17	-2	2	16.04	-0.5	1	14.55	1	1	11.14
3	0.03	-2	3	14.91	-0.5	2	12.74	1	2	8.99
		-1	-1	17.58	-0.5	3	10.98	1	3	7.15
		-1	-0.5	17.30	0	0	15.36	2	2	6.87
		-1	0	16.85	0	0.5	14.43	2	3	5.17
		-1	0.5	16.25	0	1	13.44	3	3	3.67

7

- deviance difference to null model: best FP1(-2), best FP2(-2, -0.5)
- Fit of five FP2 models not much worse, power terms can be very different

FSP: age as prognostic factor?

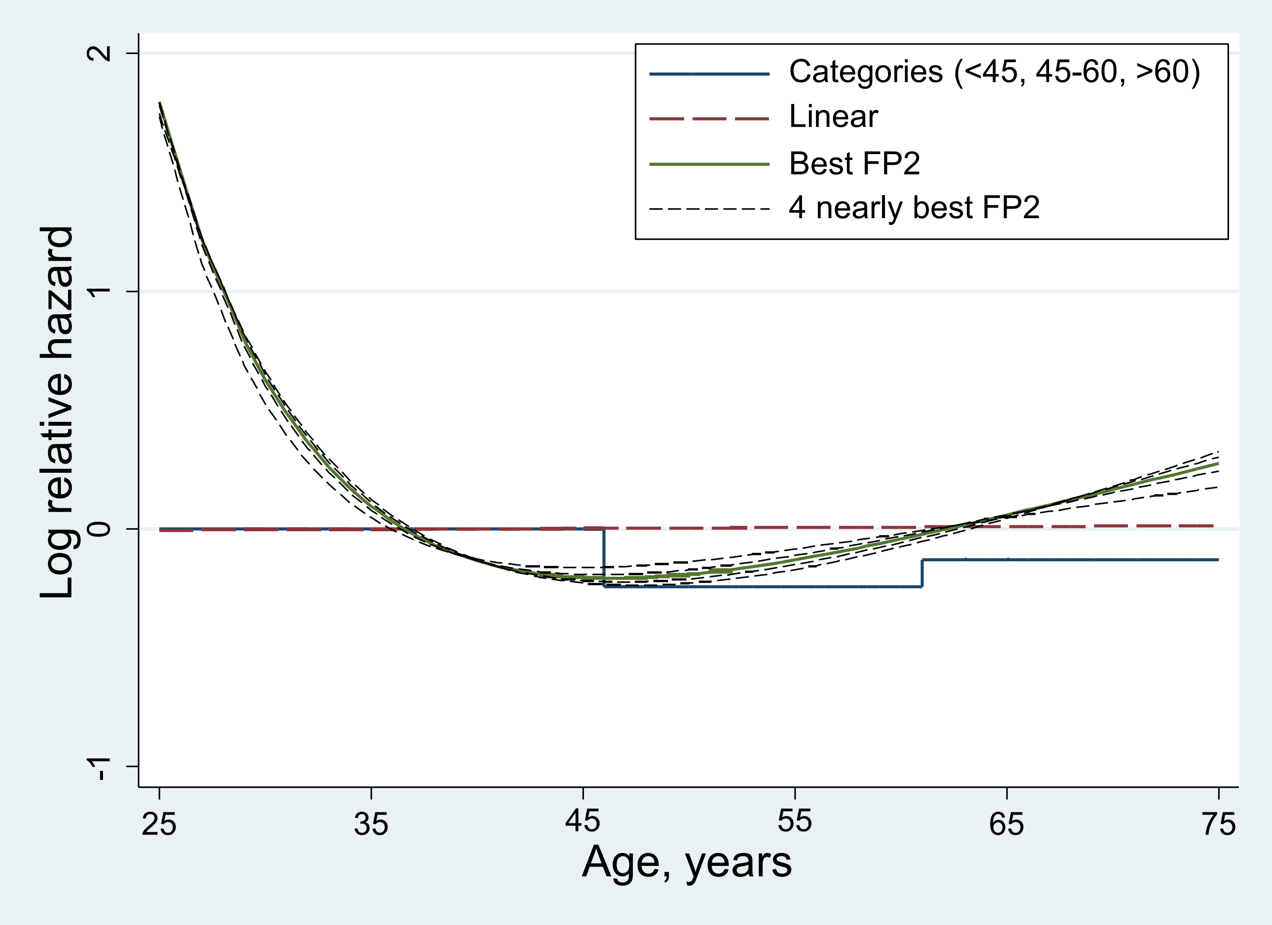
- Significance level $\alpha = 0.05$
- Deviance difference (Dev diff):
    - Linear vs. null model = 0.58
    - Best FP1(-2) vs. null model = 6.41
    - Best FP2(-2, -0.5) vs. null model = 17.61

- Best FP2 vs null: dev diff= 17.61, P = 0.0015 (4 d.f.)
- Best FP2 vs linear: dev diff= 17.03, P = 0.0007 (3 d.f.)
- Best FP2 vs Best FP1: dev diff= 11.20, P = 0.0037 (2 d.f.)

All tests are significant, therefore select the best FP2 model for any
A > 0.0037

OF CAUSE: a priori specification of significance level

Age: comparing several models



Signif tests: categories, P = 0.9; linear, P = 0.2; FP2, P = 0.001

Comparison of approaches

Assuming linearity
- nothing points to an effect of age

Two predefined cutpoints
- minor effect possible, but far from being significant

FP function
- hardly any effect for age larger than about 40 years
- severe increase of risk for age below 40 years

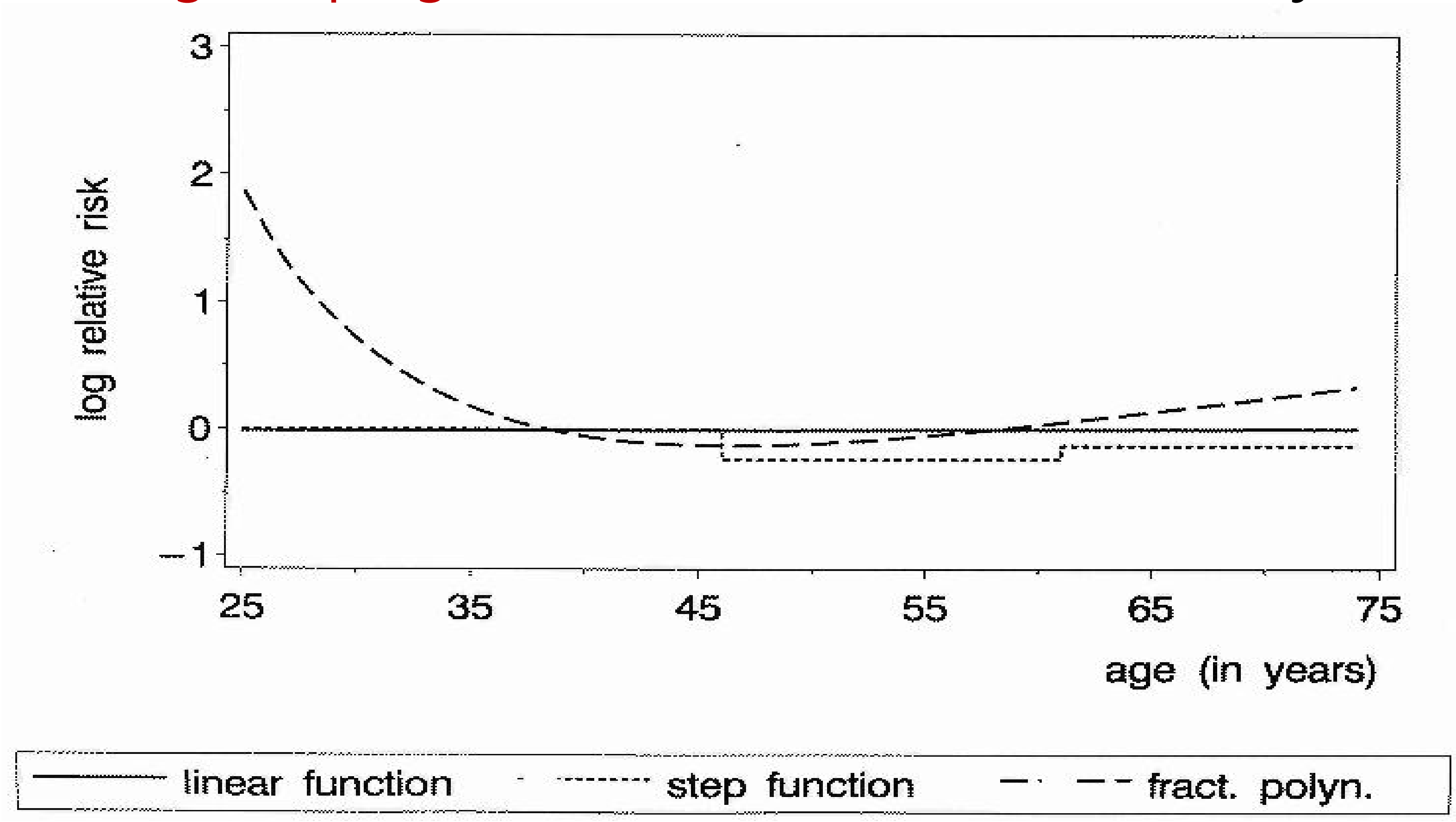
Interpretation of FP2 models

- Different powers may have similar X^2-values
- In general, *the regression coefficients are meaningless*
- Must *plot the function* to understand the model
- In the age example, the 5 best FP2 models have *almost identical curves* but *different powers* and hence *different regression coefficients*

Continuous factors

Different results with different analyses

Age as prognostic factor in breast cancer (adjusted)

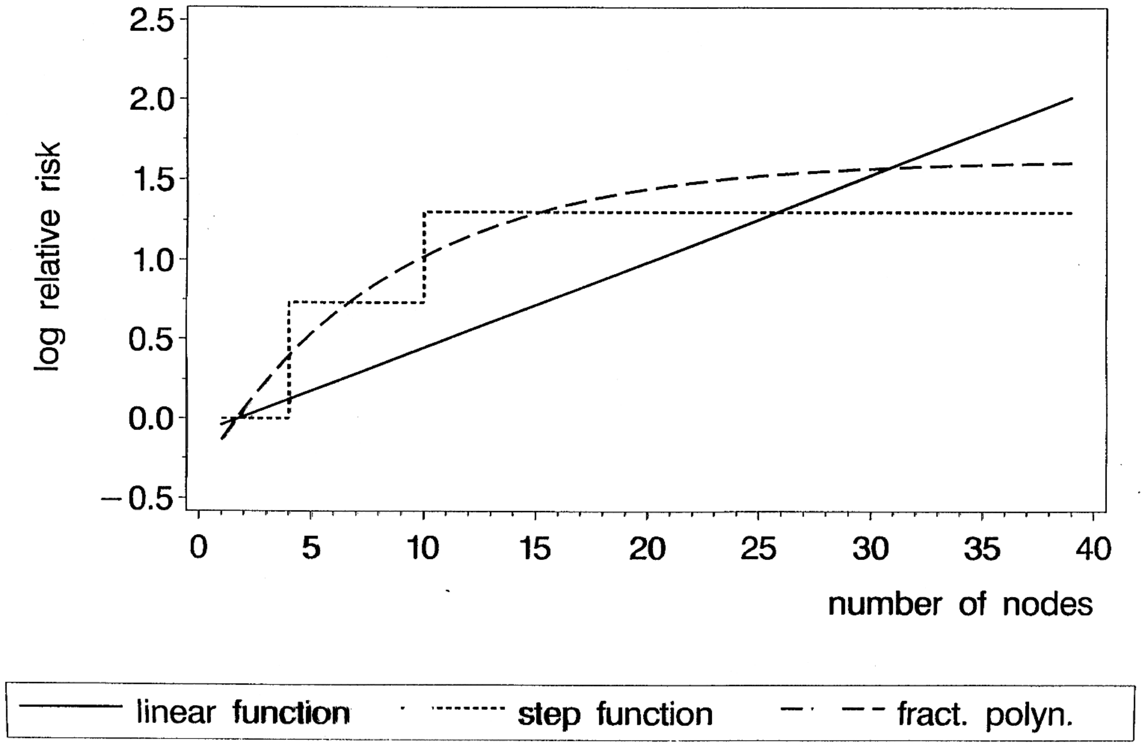


P-value 0.9 0.2 0.001

Different approaches: Results similar?

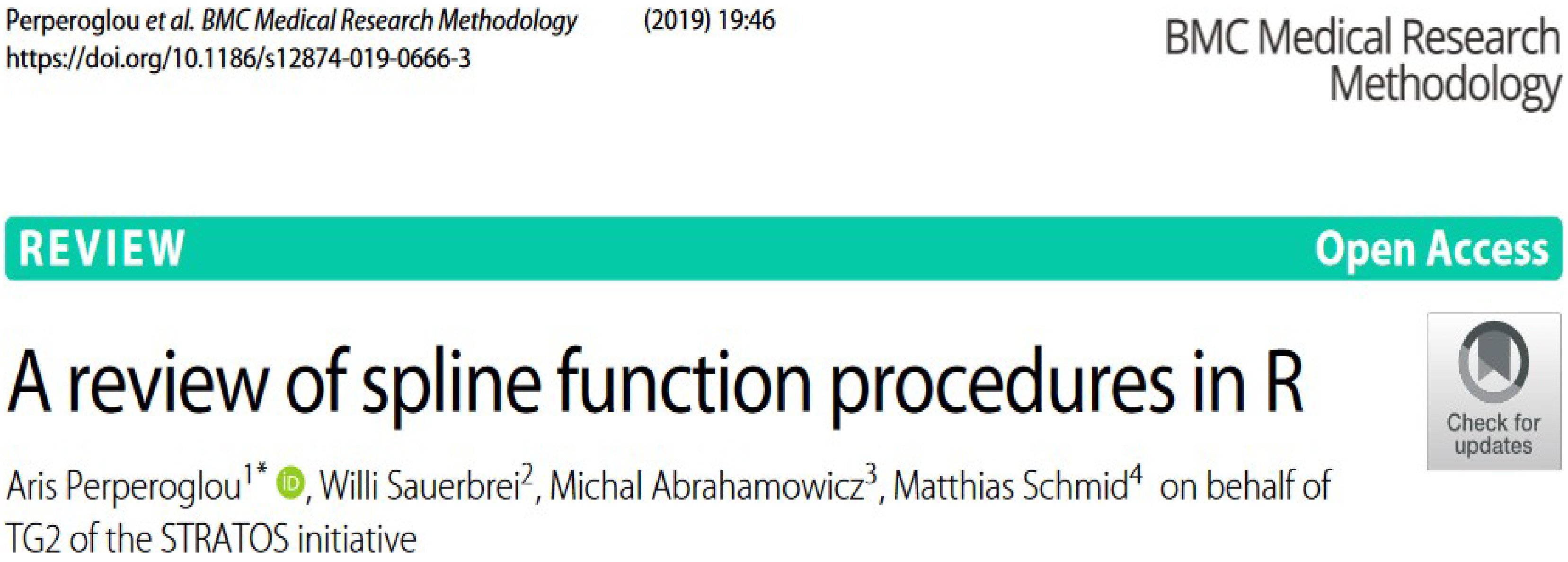
Nodes as prognostic factor in breast cancer (adjusted)

- Yes, if only p-values are considered (all <0.001)
- No, functions are very different



P-value 0.001 0.001 0.001

Splines as alternative



Perperoglou *et al. BMC Medical Research Methodology* (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

BMC Medical Research Methodology

REVIEW

Open Access

A review of spline function procedures in R

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

Splines are more flexible and more popular alternatives. However, they are also more complicated and guidance (which spline approach, how many knots, …) is missing.

Summary

- Categorization has severe problems
- FPs use the full information and generally fit the data well
- Closed test procedure for function selection
- Although the class of FP1 and FP2 functions is small, very different types of functions can be modelled
- For multivariable models MFP is a pragmatic procedure with the twin aims of selecting important variables and determining a suitable functional form for continuous variables.