Regression Model-building with continuous variables – multivariable fractional polynomials, with extension for interactions

Willi Sauerbrei[1] and **Patrick Royston**[2]

[1]**IMBI, University Medical Center Freiburg, Germany**
[2]**MRC Clinical Trials Unit, London, UK**

wfs@imbi.uni-freiburg.de

Thanks for inviting me again.

		Main topic	*One message*
1993	BBS	Variable selection	BE is good
1995	ROES	Resampling/Model stability	Stability investigations => simpler models
2004	BBS	MFP, MFPI	Useful for variable and function selection; treatment interactions
2011	BBS	MFPIgen	

Overview

Part 1 General issues in regression models

Part 2 Fractional polynomial models

- *univariate*
- *multivariate*

Part 3 Interactions

- *two continuous variables*
- *with treatment*
- *with time*

Observational Studies

Several variables, mix of continuous and (ordered) categorical variables

Different situations:

- prediction
- explanation
- confounders only

Explanation is the main interest here:

- Identify variables with (strong) influence on the outcome
- Determine functional form (roughly) for continuous variables

The issues are very similar in different types of regression models (linear regression model, GLM, survival models ...)

Use subject-matter knowledge for modelling ...
... but for some variables, data-driven choice inevitable

Regression models

$X=(X_1, \ldots, X_p)$ covariate, prognostic factors

g(x) $= ß_1 X_1 + ß_2 X_2 + \ldots + ß_p X_p$ (assuming effects are linear)

normal errors (linear) regression model

Y normally distributed

$\mathbf{E\ (Y|X) = ß_0 + g(X)}$

$\mathrm{Var}\ (Y|X) = \sigma^2 I$

logistic regression model

Y binary

$$\text{Logit } P\ (Y|X) = \mathbf{ln}\ \frac{\mathbf{P(Y=1|X)}}{\mathbf{P(Y=0|X)}} = \boldsymbol{\beta_0} + \mathbf{g(X)}$$

survival times

T survival time (partly censored)

Incorporation of covariates

$$\boldsymbol{\lambda(t|X) = \lambda_0(t)\exp(g(X))}$$

Implicit assumptions

- Subject matter knowledge (if available) determines (parts) of the model
- About 5 to 30 candidate variables
- No ‚small sample size‛ situation
- No missing data problem

Central issues

To select or not to select (full model)?

Which variables to include?

How to model continuous variables?

Continuous variables – The problem

“Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge”

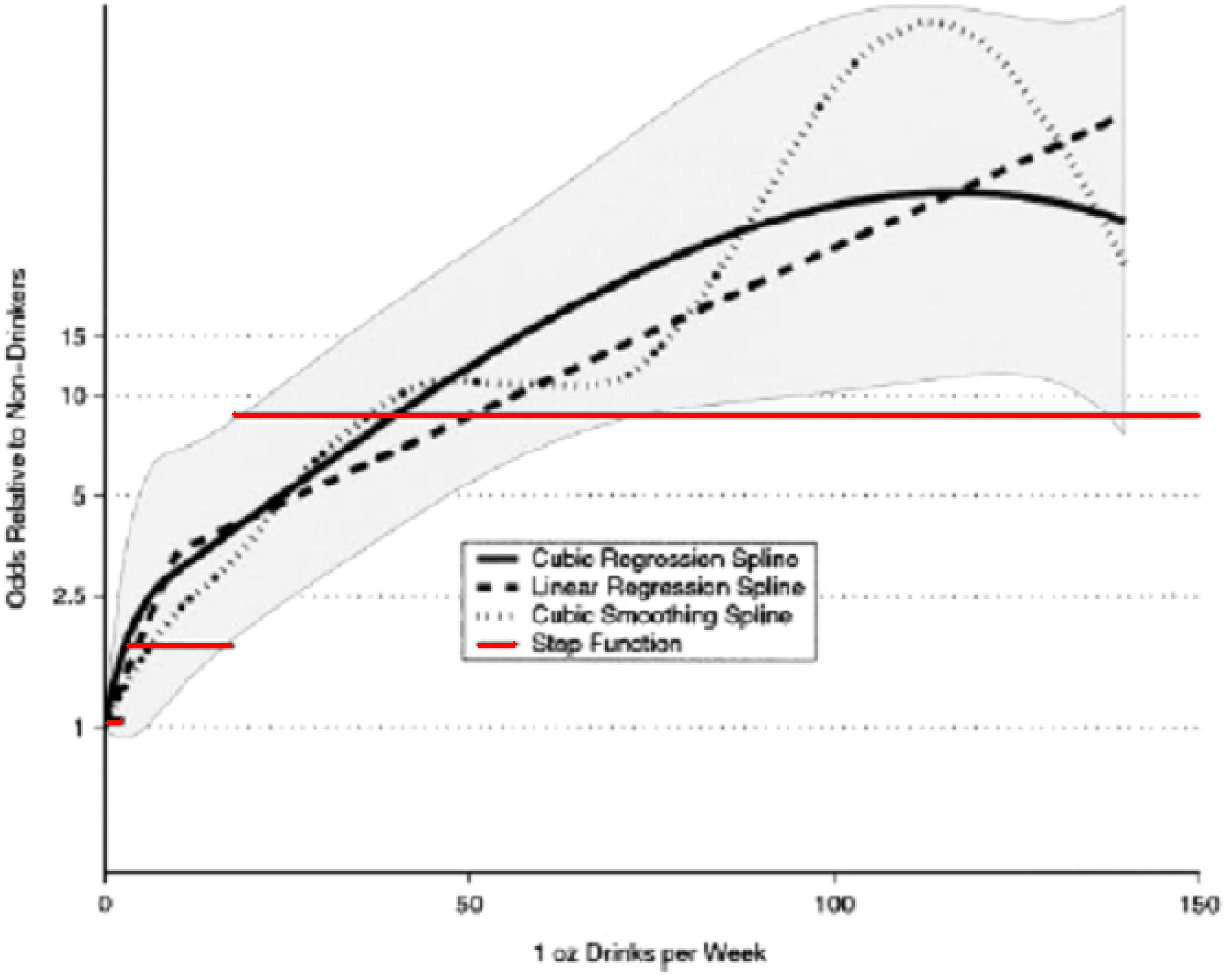
Rosenberg PS et al, Statistics in Medicine 2003; 22:3369-3381

Discussion of issues in (univariate) modelling with splines

Trivial nowadays to *fit* almost any model

To *choose* a good model is much harder

Alcohol consumption as risk factor for oral cancer



Rosenberg et al, StatMed 2003

Multivariable models –
methods for variable selection

Full model

- variance inflation in the case of multicollinearity

Stepwise procedures ⇒ prespecified (α_{in}, α_{out}) and actual significance level?

- forward selection (FS)
- stepwise selection (StS)
- backward elimination (BE)

All subset selection ⇒ which criteria?

•	C_p	Mallows	$= (SSE / \hat{\sigma}^2) - n$	$+ p\ 2$
•	AIC	Akaike Information Criterion	$= n \ln (SSE / n)$	$+ p\ 2$
•	BIC	Bayes Information Criterion	$= n \ln (SSE / n)$	$+ p \ln(n)$
			fit	penalty

Combining selection with Shrinkage

Bayes variable selection

Recommendations???

Central issue: MORE OR LESS COMPLEX MODELS?

Backward elimination is a sensible approach

- Significance level can be chosen
- Reduces overfitting

Of course required

- Checks
- Sensitivity analysis
- Stability analysis

Continuous variables – what functional form?

Traditional approaches

a) Linear function
   - may be inadequate functional form
   - misspecification of functional form may lead to wrong conclusions

b) ‘best‘ ‘standard‘ transformation

c) Step function (categorial data)
   - Loss of information
   - How many cutpoints?
   - Which cutpoints?
   - Bias introduced by outcome-dependent choice

Example 1: Prognostic factors

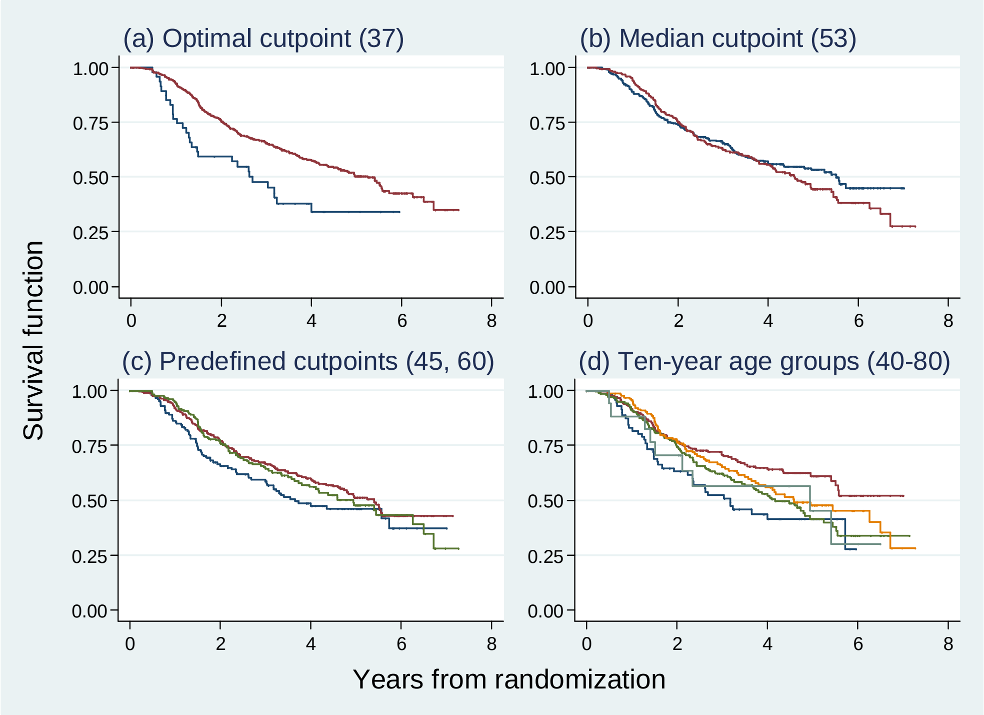
GBSG-study in node-positive breast cancer

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

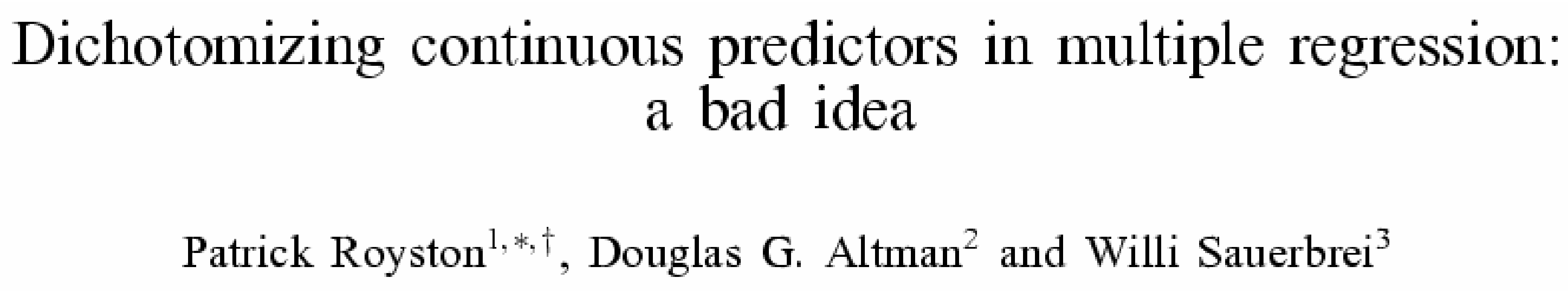
7 prognostic factors, of which **5** are continuous

Tamoxifen yes/no

Age as prognostic factor – cutpoint analyses



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);
(d) popular (10-year groups)



Dichotomizing continuous predictors in multiple regression: a bad idea

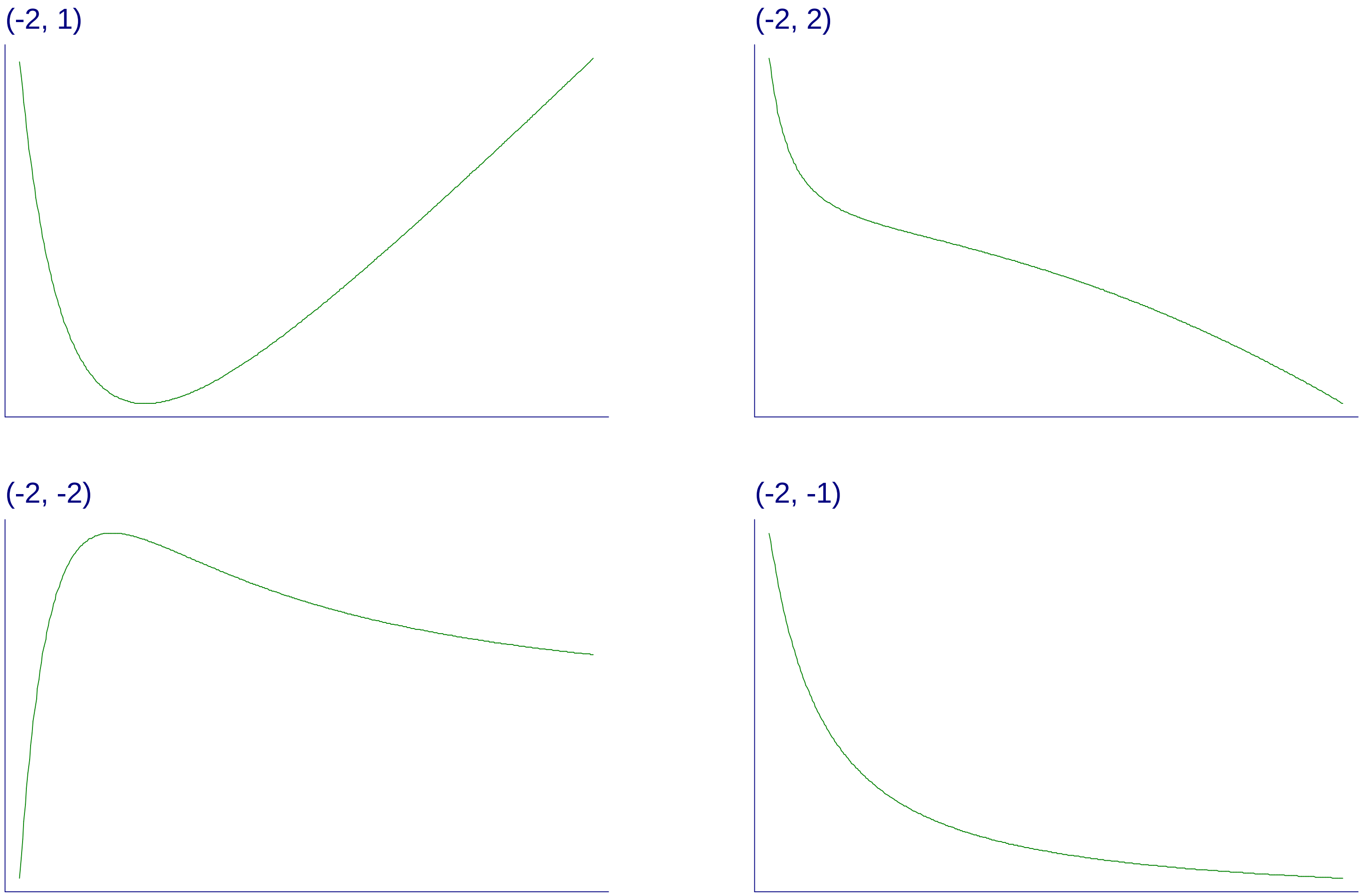
Patrick Royston[1,*,†], Douglas G. Altman[2] and Willi Sauerbrei[3]

Statistics in Medicine, 2006, 25:127-141

Fractional polynomial models

- Describe for one covariate, X
- Fractional polynomial of degree m for X with powers $p_1, \ldots , p_m$ is given by
  $$\mathrm{FP}m(X) = \beta_1 X^{p1} + \ldots + \beta_m X^{pm}$$
- Powers $p_1,\ldots, p_m$ are taken from a special set
  $$\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$
- Usually $m = 1$ or $m = 2$ is sufficient for a good fit
- Repeated powers (p_1=p_2)
  $$\beta_1 X^{p1} + \beta_2 X^{p1}\log X$$
- 8 FP1, 36 FP2 models

Examples of FP2 curves - varying powers



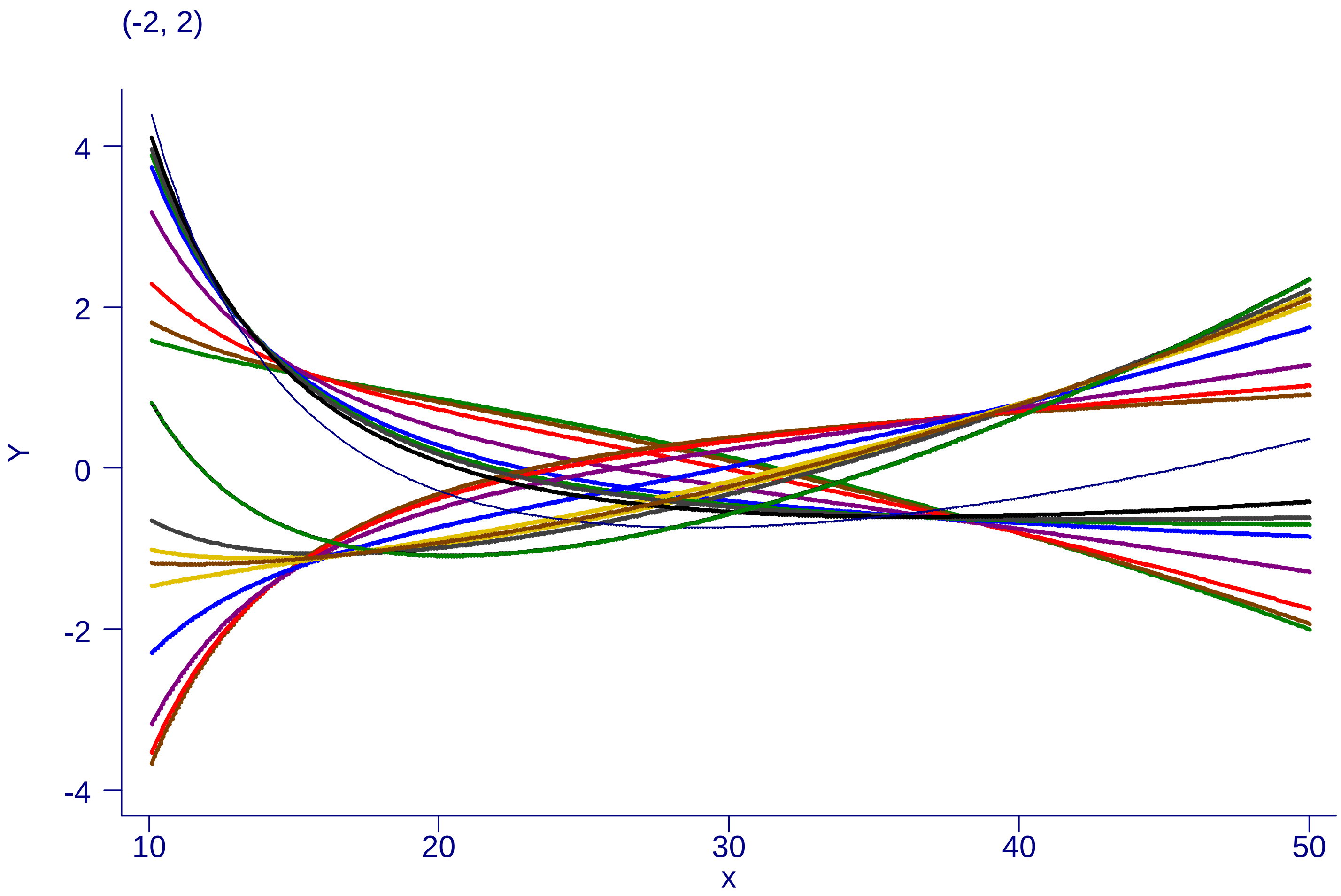
(-2, 1)

(-2, 2)

(-2, -2)

(-2, -1)

Examples of FP2 curves
- single power, different coefficients



Our philosophy of function selection

- Prefer simple (linear) model
- Use more complex (non-linear) FP1 or FP2 model if indicated by the data
- Contrasts to more local regression modelling (eg splines)
  - Already starts with a complex model

FP analysis for the effect of age

Degree 1		Degree 2								
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

Function selection procedure (FSP)

Effect of age at 5% level?

	χ^2	df	p-value
Any effect?			
Best FP2 versus null	**17.61**	**4**	**0.0015**
Linear function suitable?			
Best FP2 versus linear	**17.03**	**3**	**0.0007**
FP1 sufficient?			
Best FP2 vs. best FP1	**11.20**	**2**	**0.0037**

Many predictors – MFP

With many continuous predictors selection of best FP for each becomes more difficult → MFP algorithm as a standardized way to variable and function selection

(usually binary and categorical variables are also available)

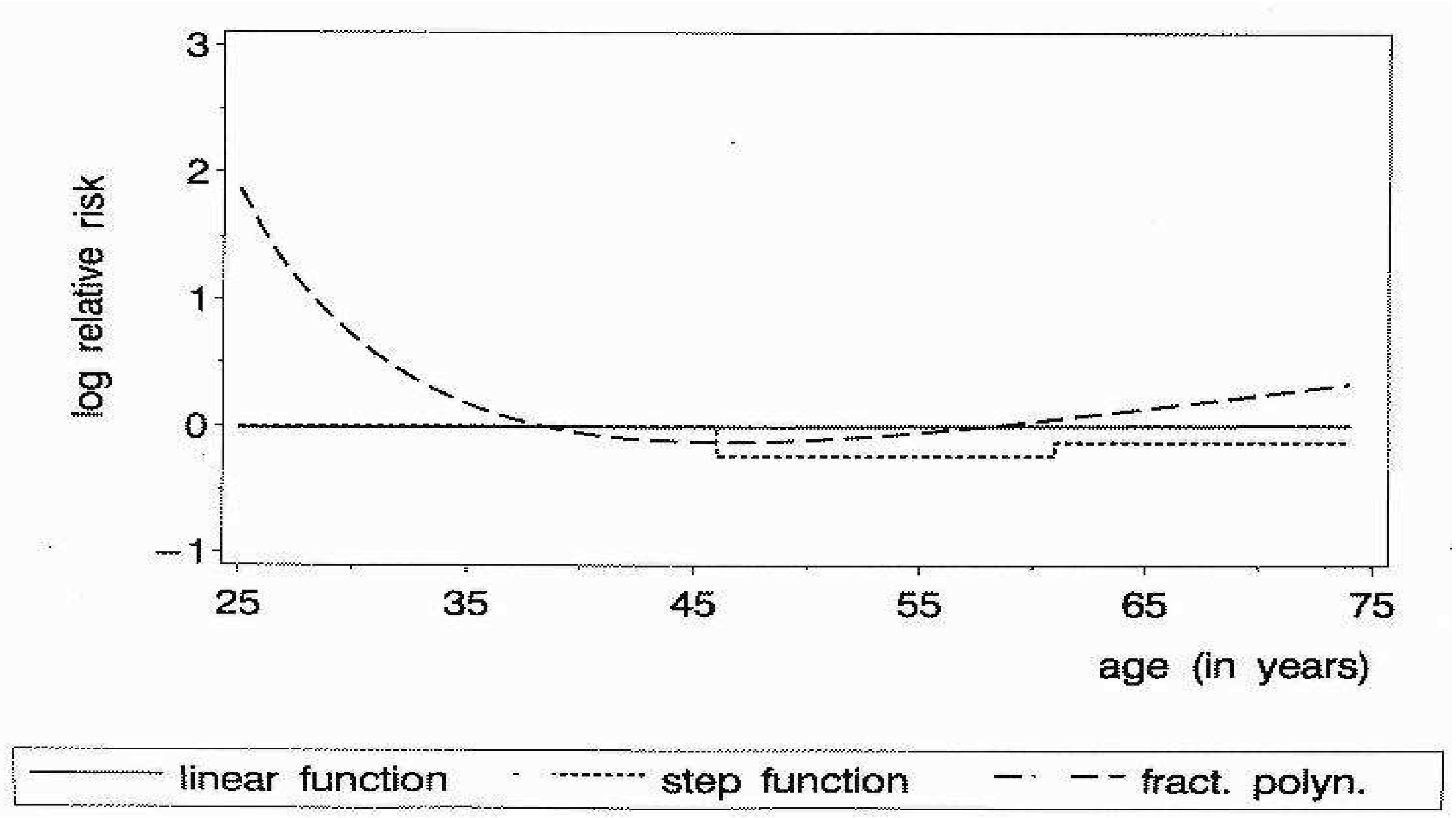
MFP algorithm combines

backward elimination with
FP function selection procedures

Continuous factors

Different results with different analyses

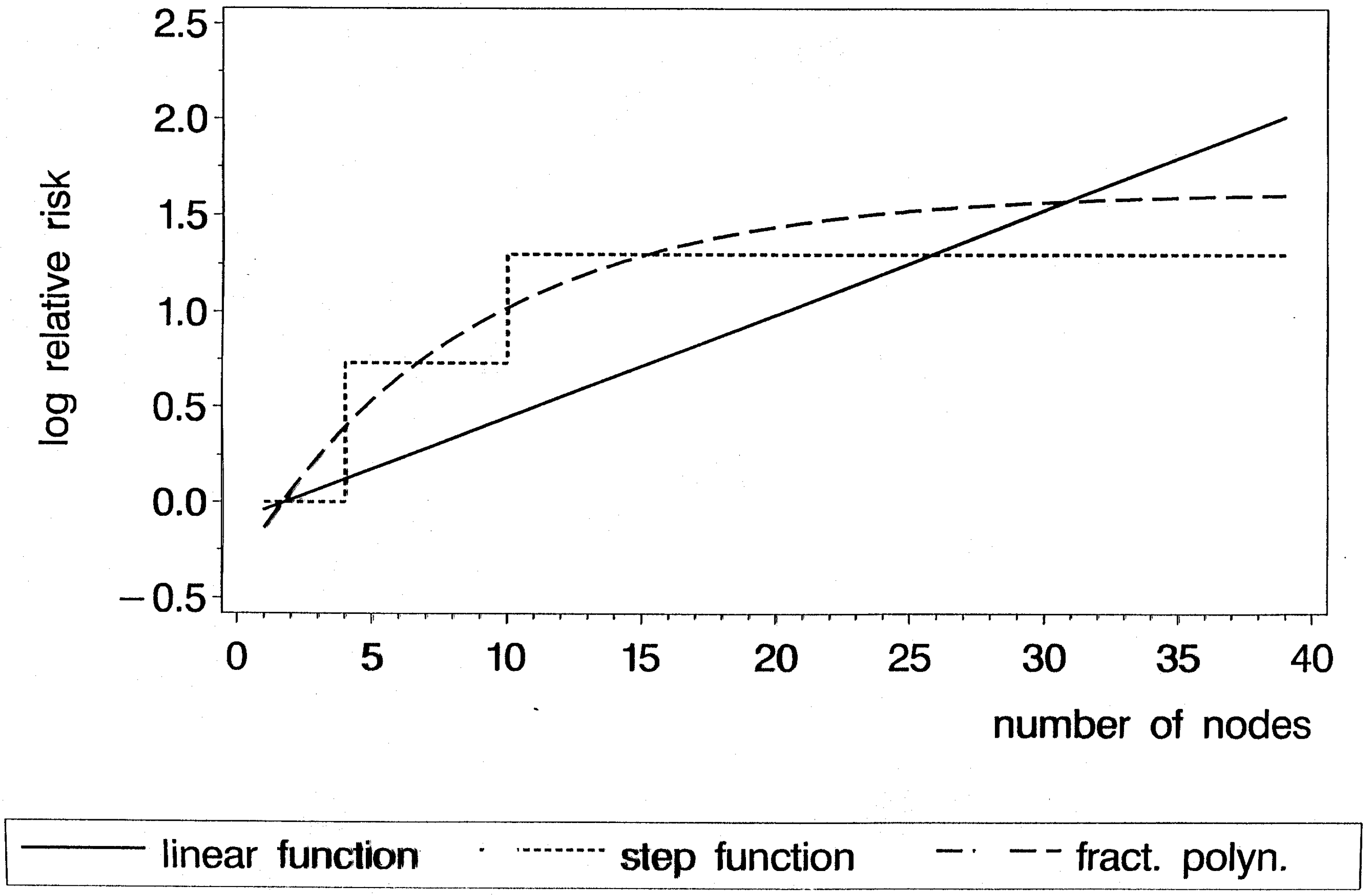
Age as prognostic factor in breast cancer (adjusted)



P-value 0.9 0.2 0.001

Results similar?

Nodes as prognostic factor in breast cancer (adjusted)



P-value 0.001 0.001 0.001

Example 2: Risk factors

- Whitehall 1
  - 17,370 male Civil Servants aged 40-64 years, 1670 (9.7%) died
  - Measurements include: age, cigarette smoking, BP, cholesterol, height, weight, job grade
  - Outcomes of interest: all-cause mortality at 10 years ⇒ logistic regression

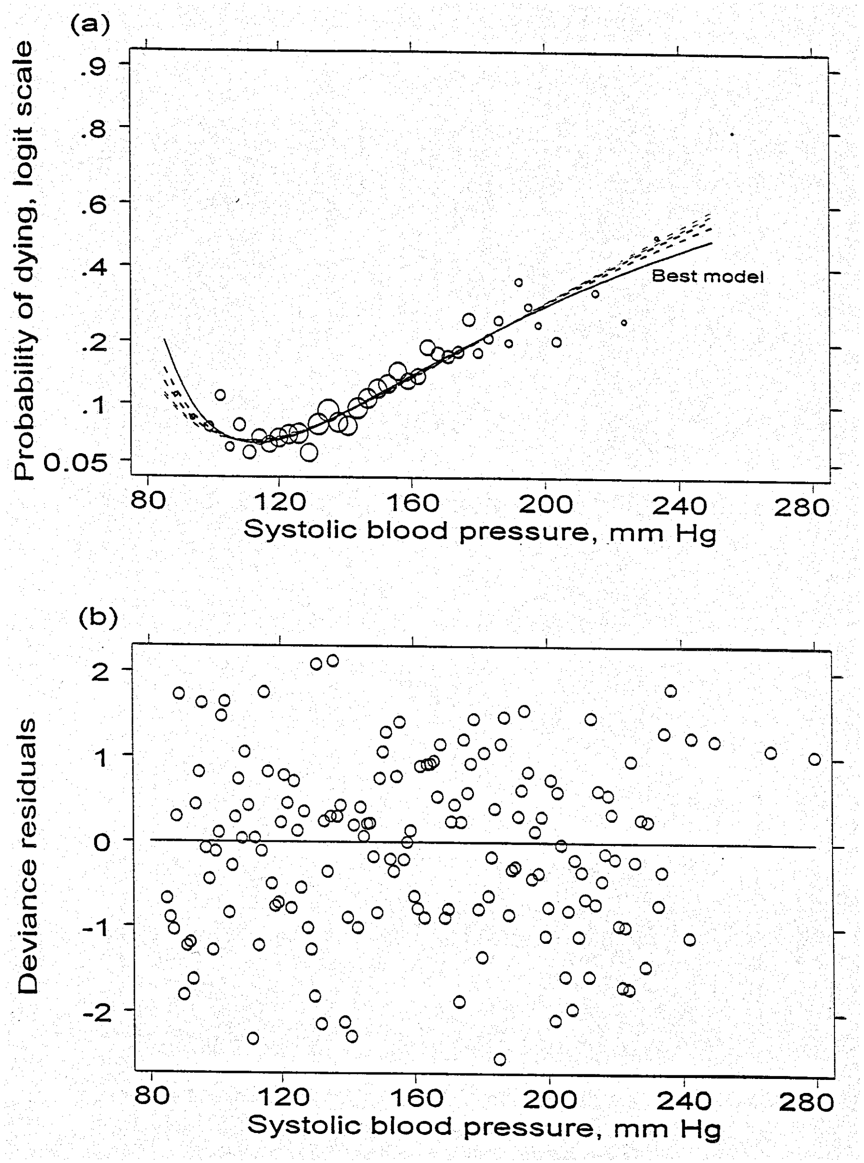
Whitehall 1

Systolic blood pressure

Deviance difference in comparison to a straight line for FP(1) and FP(2) models

Fractional polynomials										
First degree		Second degree								
Power	Deviance	Powers		Deviance	Powers		Deviance	Powers		Deviance
p	Difference	p	q	Difference	p	q	Difference	p	q	difference
-2	-74.19	-2	-2	26.22*	-1	1	12.97	0	2	7.05
-1	-43.15	-2	-1	24.43	-1	2	7.80	0	3	3.74
-0.5	-29.40	-2	-0.5	22.80	-1	3	2.53	0.5	0.5	10.94
0	-17.37	-2	0	20.72	-0.5	-0.5	17.97	0.5	1	9.51
0.5	-7.45	-2	0.5	18.23	-0.5	0	16.00	0.5	2	6.80
1	0.00	-2	1	15.38	-0.5	0.5	13.93	0.5	3	4.41
2	6.43*	-2	2	8.85	-0.5	1	11.77	1	1	8.46
3	0.98	-2	3	1.63	-0.5	2	7.39	1	2	6.61
		-1	-1	21.62	-0.5	3	3.10	1	3	5.11
		-1	-0.5	19.78	0	0	14.24	2	2	6.44
		-1	0	17.69	0	0.5	12.43	2	3	6.45
		-1	0.5	15.41	0	1	10.61	3	3	7.59

Similar fit of several functions



Presentation of models for continuous covariates

- The function + 95% CI gives the whole story
- Functions for important covariates should always be plotted
- In epidemiology, sometimes useful to give a more conventional table of results in categories
- This can be done from the fitted function

Whitehall 1

Systolic blood pressure

Odds ratio from final FP(2) model

$\text{LogOR} = 2.92 - 5.43X^{-2} - 14.30 * X^{-2} \log X$

Presented in categories

Systolic blood pressure (mm Hg)		Number of men		OR (model-based)	
Range	ref. point	at risk	dying	Estimate	95%CI
≤ 90	88	27	3	2.47	1.75, 3.49
91-100	95	283	22	1.42	1.21, 1.67
101-110	105	1079	84	1.00	-
111-120	115	2668	164	0.94	0.86, 1.03
121-130	125	3456	289	1.04	0.91, 1.19
131-140	135	4197	470	1.25	1.07, 1.46
141-160	150	2775	344	1.77	1.50, 2.08
161-180	170	1437	252	2.87	2.42, 3.41
181-200	190	438	108	4.54	3.78, 5.46
201-240	220	154	41	8.24	6.60, 10.28
241-280	250	5	4	15.42	11.64, 20.43

Whitehall 1
MFP analysis

Covariate	FP etc.
Age	Linear
Cigarettes	0.5
Systolic BP	-1, -0.5
Total cholesterol	Linear
Height	Linear
Weight	-2, 3
Job grade	In

No variables were eliminated by the MFP algorithm

Assuming a linear function weight is eliminated by backward elimination

Interactions
Motivation – I

Detecting predictive factors (interaction with treatment)

- Don’t investigate effects in separate subgroups!
- Investigation of treatment/covariate interaction requires statistical tests
- Care is needed to avoid over-interpretation
- Distinguish two cases:
  - Hypothesis generation: searching several interactions
  - Specific predefined hypothesis

Motivation - II

Continuous by continuous interactions

- usually linear by linear product term
- not sensible if main effect is non-linear
- mismodelling the main effect may introduce spurious interactions

Continuous by continuous interactions MFPIgen

- Have Z_1, Z_2 continuous and X confounders
- Apply MFP to X, Z_1 and Z_2, forcing Z_1 and Z_2 into the model.
  - FP functions $f_1(Z_1)$ and $f_2(Z_2)$ are selected for Z_1 and Z_2
- Often $f_1(Z_1)$ and/or $f_2(Z_2)$ are linear
- Add term $f_1(Z_1)* f_2(Z_2)$ to the model chosen and use LRT for test of interaction
- Check (graphically) interactions for artefacts
- Check all pairs of continuous variables for an interaction
- Use forward stepwise if more than one interaction remains
- Low significance level for interactions

Interactions – continuous by continuous
Whitehall 1

Consider only age and weight

Main effects:
age – linear
weight – FP2 (-1,3)

Interaction?

Include age*weight^{-1} + age*weight^{3}
into the model

LRT: $\chi^2 = 5.27$ (2df, $p = 0.07$) $\Rightarrow$ no (strong) interaction

Erroneously assume that the effect of weight is linear

Interaction?

Include age*weight into the model

LRT: $\chi^2 = 8.74$ (1df, $p = 0.003$)

$\Rightarrow$ intercation appears higly significant

Interaction: checking the model

- Model check:

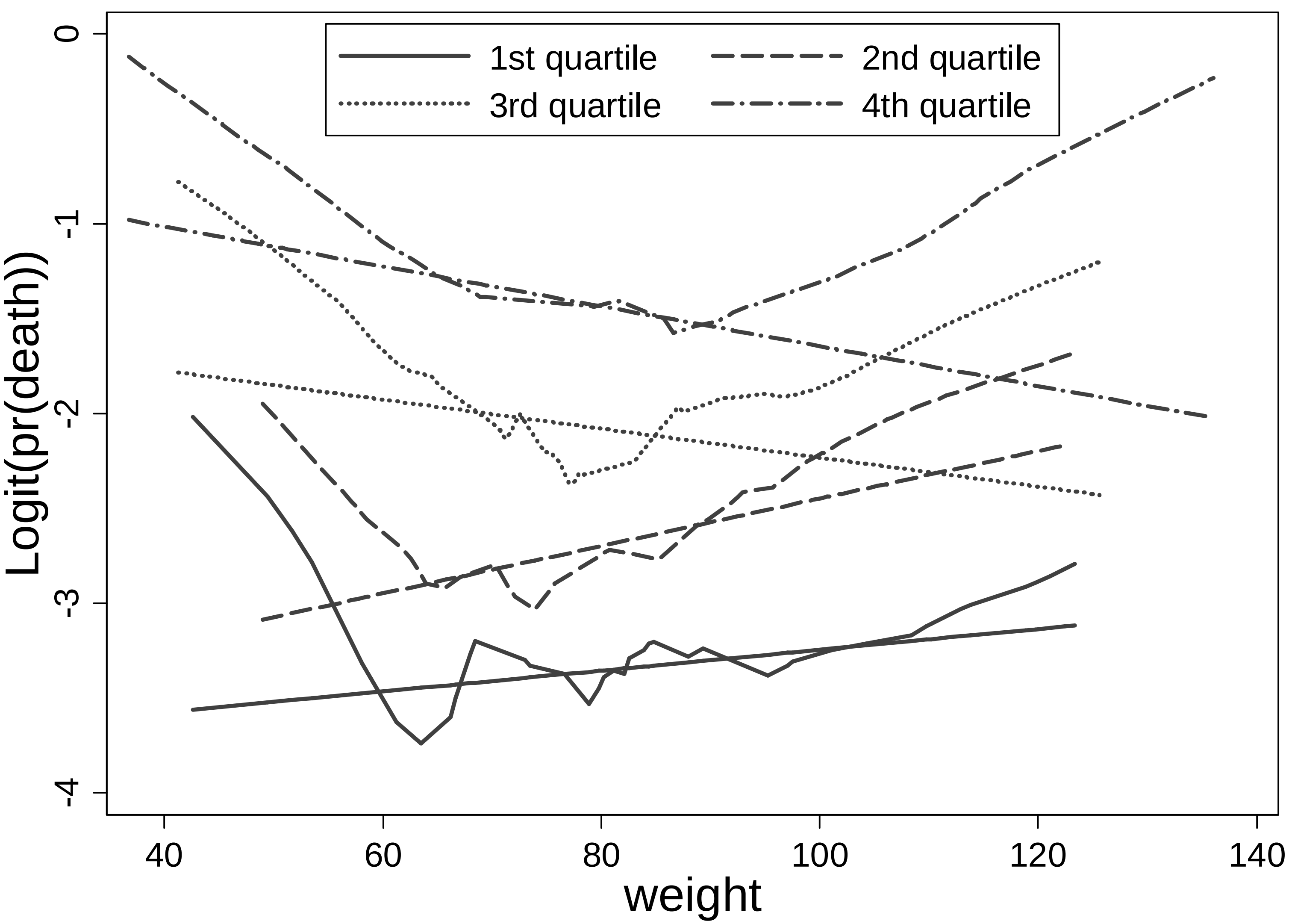
  categorize age in (equal sized) groups (e.g. 4 groups)

- Computer running line smooth of the binary outcome on weight in each group

- Plot results for each group

Whitehall 1: check of age × weight interaction

– 4 subgroups for **age**

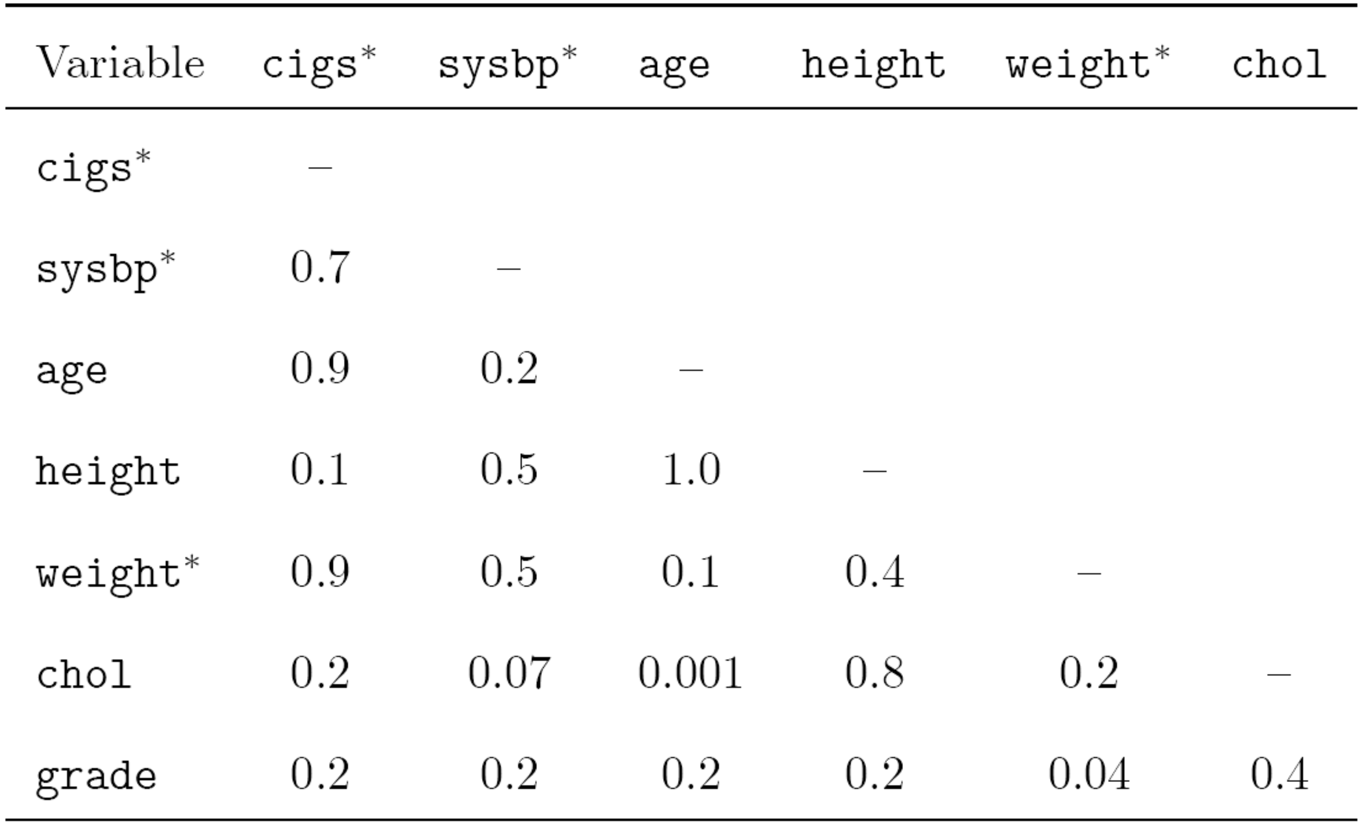


Interpreting the plot

- Running line smooth are roughly parallel across age groups ⇒ no (strong) interactions
- Erroneously assume that the effect of weight is linear ⇒ estimated slopes of weight in age-groups indicate strong qualitative interaction between age und weight

Whitehall 1: 7 variables – any interactions?

P-values for two-way interactions from MFPIgen



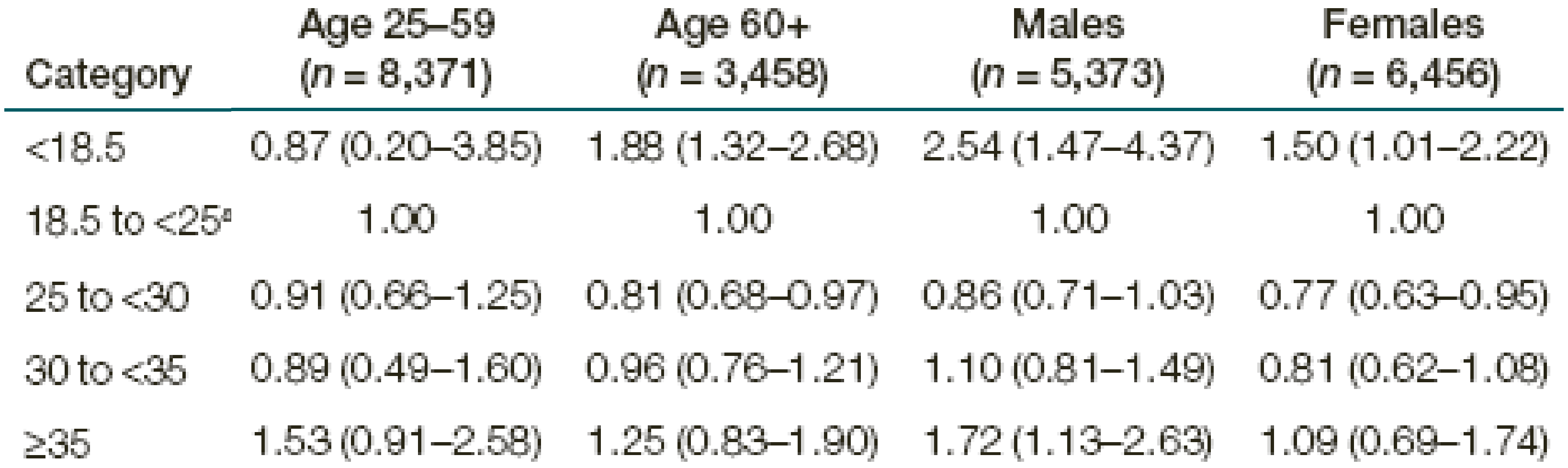
Variable	cigs*	sysbp*	age	height	weight*	chol
cigs*	–					
sysbp*	0.7	–				
age	0.9	0.2	–			
height	0.1	0.5	1.0	–		
weight*	0.9	0.5	0.1	0.4	–	
chol	0.2	0.07	0.001	0.8	0.2	–
grade	0.2	0.2	0.2	0.2	0.04	0.4

*FP transformations

=> chol∗age highly significant, but needs checking

State of the art??
Analyses in subgroups
Main effect categorized, age categorized

BMI RR (95% confidence intervall), adjusted



Category	Age 25–59 (*n* = 8,371)	Age 60+ (*n* = 3,458)	Males (*n* = 5,373)	Females (*n* = 6,456)
<18.5	0.87 (0.20–3.85)	1.88 (1.32–2.68)	2.54 (1.47–4.37)	1.50 (1.01–2.22)
18.5 to <25[a]	1.00	1.00	1.00	1.00
25 to <30	0.91 (0.66–1.25)	0.81 (0.68–0.97)	0.86 (0.71–1.03)	0.77 (0.63–0.95)
30 to <35	0.89 (0.49–1.60)	0.96 (0.76–1.21)	1.10 (0.81–1.49)	0.81 (0.62–1.08)
≥35	1.53 (0.91–2.58)	1.25 (0.83–1.90)	1.72 (1.13–2.63)	1.09 (0.69–1.74)

Orpana et al, Obesity 2009

BMI*age interaction?

Males: BMI effect interpretable?

Software sources MFP

- Most comprehensive implementation is in Stata
  - Command **mfp** is part since Stata 8 (now Stata 11)
- Versions for SAS and R are available
  - SAS
    www.imbi.uni-freiburg.de/biom/mfp
  - R version available on CRAN archive
    - **mfp** package
- Extensions to investigate interactions
    - So far only in Stata

Concluding comments – MFP

- FPs use full information - in contrast to a priori categorisation
- FPs search within flexible class of functions (FP1 and FP2 - 44 models)
- MFP is a well-defined multivariate model-building strategy – combines search for transformations with BE
- Important that model reflects medical knowledge,

  e.g. monotonic / asymptotic functional forms

Towards recommendations for model-building by selection of variables and functional forms for continuous predictors under several assumptions

Issue	**Recommendation**
Variable selection procedure	Backward elimination; significance level as key tuning parameter, choice depends on the aim of the study
Functional form for continuous covariates	Linear function as the 'default', check improvement in model fit by fractional polynomials. Check derived function for undetected local features
Extreme values or influential points	Check at least univariately for outliers and influential points in continuous variables. A preliminary transformation may improve the model selected. For a proposal see R & S 2007
Sensitivity analysis	Important assumptions should be checked by a sensitivity analysis. Highly context dependent
Check of model stability	The bootstrap is a suitable approach to check for model stability
Complexity of a predictor	A predictor should be 'as parsimonious as possible'

Sauerbrei et al. SiM 2007, Royston & Sauerbrei 2008

Interactions

- Interactions are often ignored by analysts
- Continuous × categorical has been studied in FP context because clinically very important
- Continuous × continuous is more complex
- Interaction with time important for long-term FU survival data

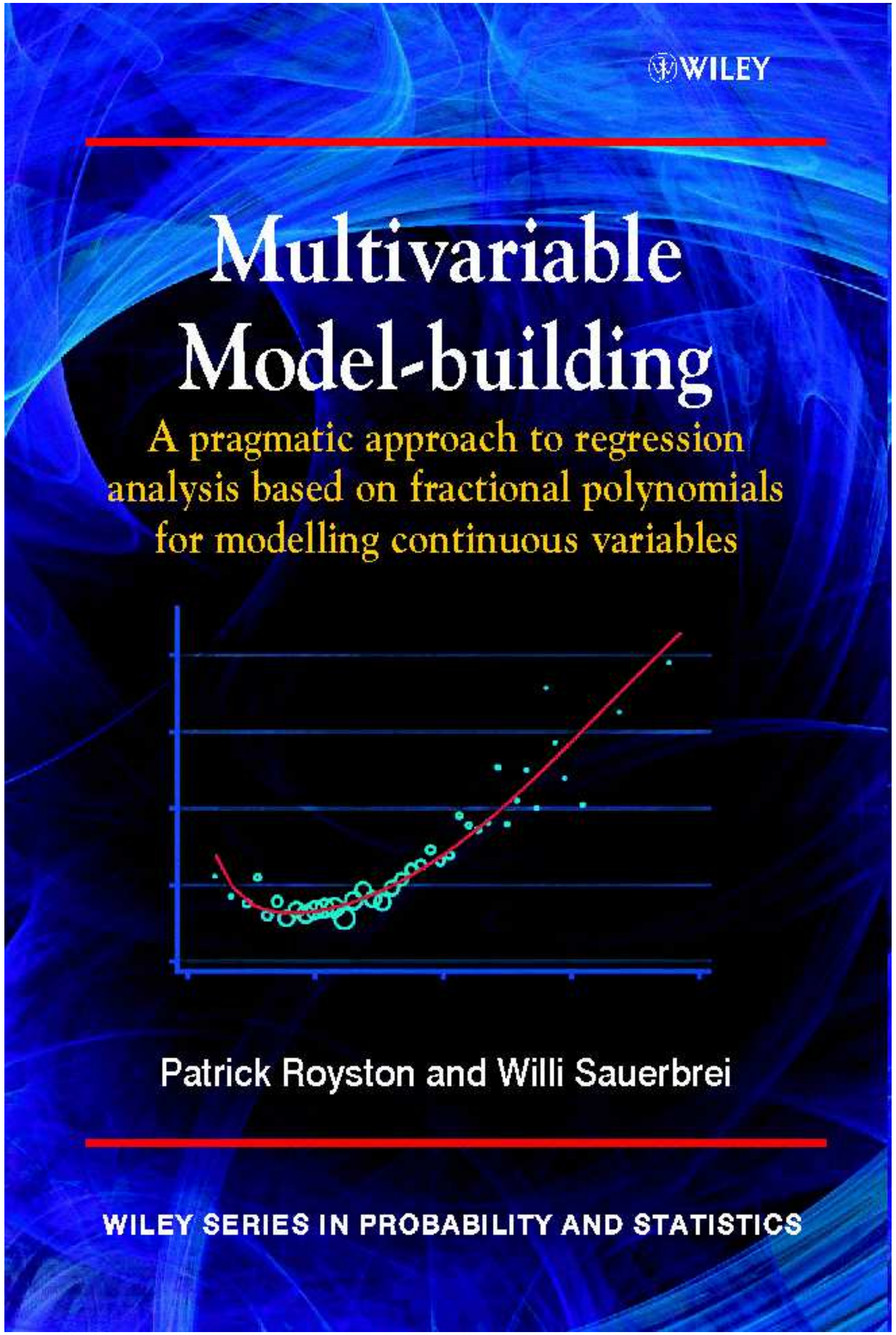
MFP extensions

- MFPI – treatment/covariate interactions
- MFPIgen – interaction between two continuous variables
- MFPT – time-varying effects in survival data

Summary

Getting the big picture right is more important than optimising aspects and ignoring others

- strong predictors
- strong non-linearity
- strong interactions
- strong non-PH in survival model



WILEY

Multivariable Model-building

A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables

Patrick Royston and Willi Sauerbrei

WILEY SERIES IN PROBABILITY AND STATISTICS

References

Royston P, Altman DG. (1994): Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Applied Statistics, 43:429-467.

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

Royston P, Sauerbrei W (2008): Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for continuous variables. Wiley.

Royston P, Sauerbrei W (2009): Two techniques for investigating interactions between treatment and continuous covariates in clinical trials. Stata Journal, 9: 1-22.

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. Computational Statistics & Data Analysis, 50:3464-3485.

Sauerbrei W, Royston P. (1999): Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society A, 162:71-94.

Sauerbrei W, Royston P, Binder H (2007): Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in Medicine, 26:5512-28.

Sauerbrei W, Royston P, Zapien K. (2007): Detecting an interaction between treatment and a continuous covariate: a comparison of two approaches. Computational Statistics and Data Analysis, 51:4054-4063.