On 'state-of-the-art' for selection of variables and functional forms in multivariable analysis

Willi Sauerbrei[1],
Patrick Royston, Michal Abrahamowicz,
Georg Heinze, Aris Perperoglou
for TG2 of the STRATOS Initiative

[1]Medical Center – University of Freiburg, Germany

STRATOS INITIATIVE

UNIVERSITÄTS KLINIKUM FREIBURG

State-of-the-art

State-of-the-art *refers to the* ***highest level of general development****, as of a device, technique, or scientific field achieved at a particular time.*

Wikipedia, 12 June 2017

Conclusion

We are far away from 'state-of-the-art' on selection of variables and functional forms.

Much research urgently needed!

Content

1. General issues
2. Some approaches for the selection of variables
3. Typical approaches for the selection of functional forms
4. Flexible modelling (Perperoglou)
5. Combining variable and function selection
   - MFP
   - Splines
6. ‘State-of-the-art‘ – required research

General issue in observational studies

Several variables, mix of continuous and (ordered) categorical variables, pairwise- and multicollinearity present

Model selection required

Use subject-matter knowledge for modelling ...

... but for some variables, data-driven choice inevitable

Regression models

$X=(X_1, ...,X_k)$ covariate, prognostic factors

g(x) = $ß_1 X_1 + ß_2 X_2 +...+ ß_k X_k$ (assuming effects are linear)

normal errors (linear) regression model

Y normally distributed

E (Y|X) = $ß_0$ + g(X)

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

logistic regression model

Y binary

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

survival times

T survival time (partly censored)

Incorporation of covariates

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

Aims of multivariable models

- Prediction of an outcome of interest
- Identification of ‘important’ predictors
- Adjustment for predictors uncontrollable by experimental design
- Stratification by risk
- ... and many more

Building multivariable regression models – some preliminaries

- ‚Reasonable' model class was chosen
- Comparison of strategies
  - Theory

    only for limited questions, unrealistic assumptions
  - Examples or simulation
    - Examples based on published data
      - oversimplifies the problem
      - data clean
      - ‚relevant' predictors given

    → rigorous pre-selection → what is a full model?

… preliminaries continued

More problems are available,

see discussion on initial data analysis in Chatfield (2002) section ‚*Tackling real life statistical problems*‘

See also Mallows (1998)

2. Some approaches for the selection of variables

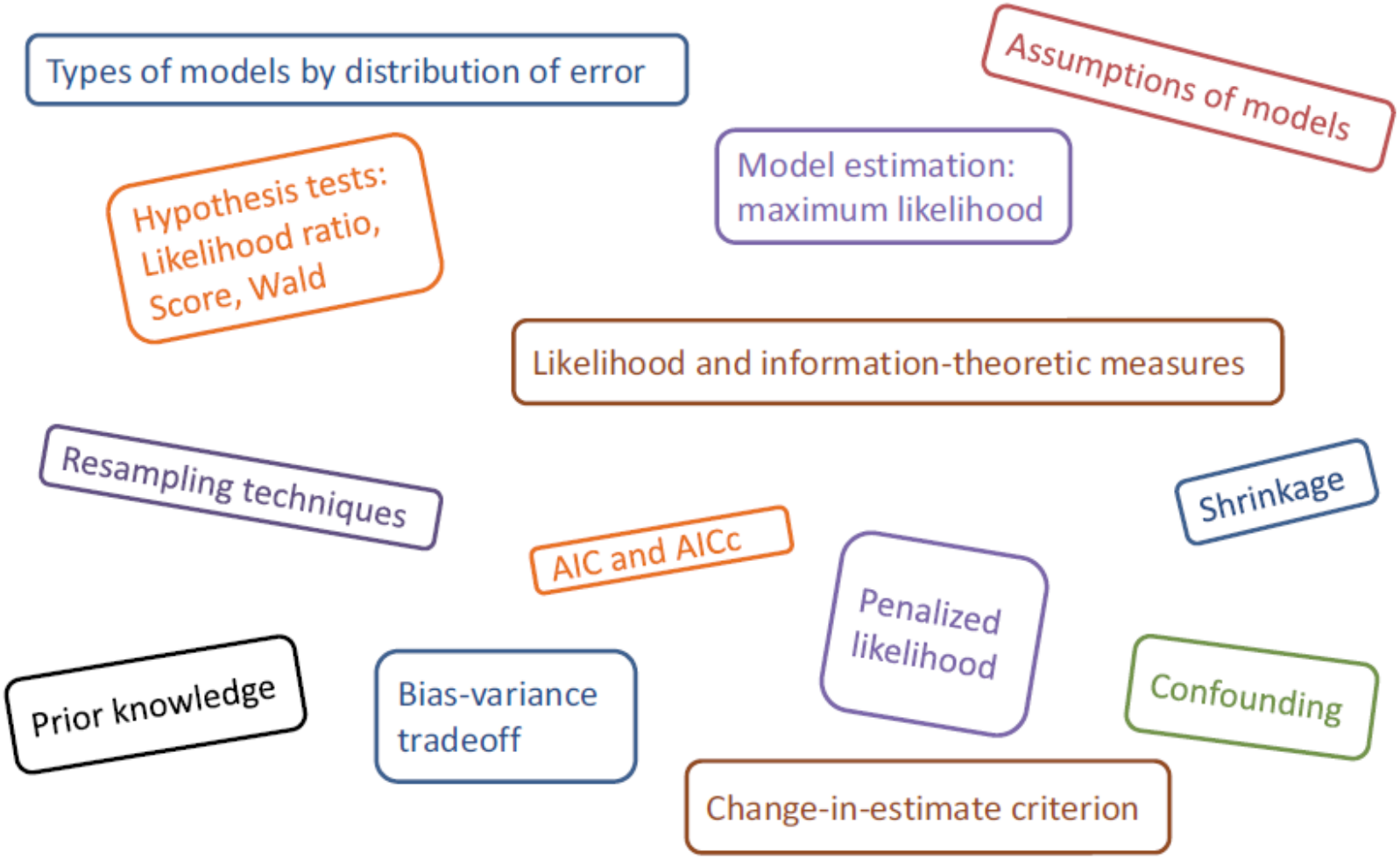
Central issues:

- Model with focus on prediction or explanation?
- To select or not to select (full model)?
- Which variables to include?

Selection of variables

- A large number of methods proposed for many decades
- High-dimensional data triggered the development of further proposals
- Many issues

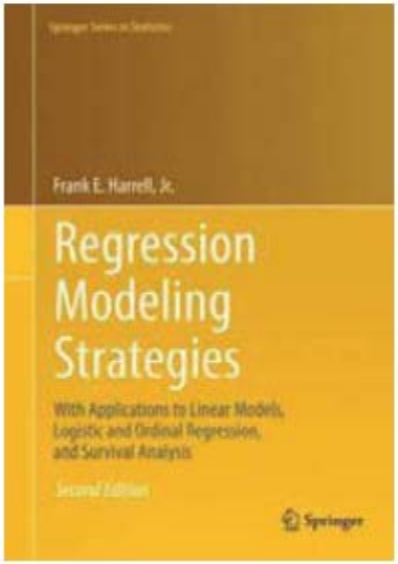
Selection of variables: Statistical prerequisites

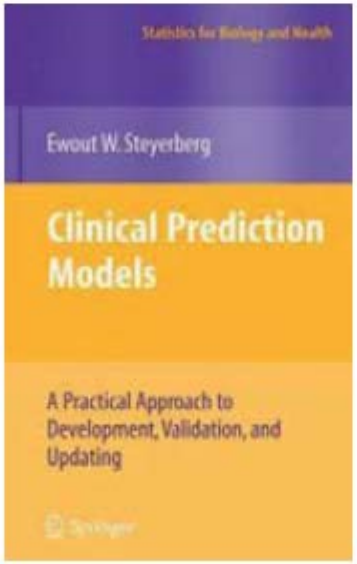


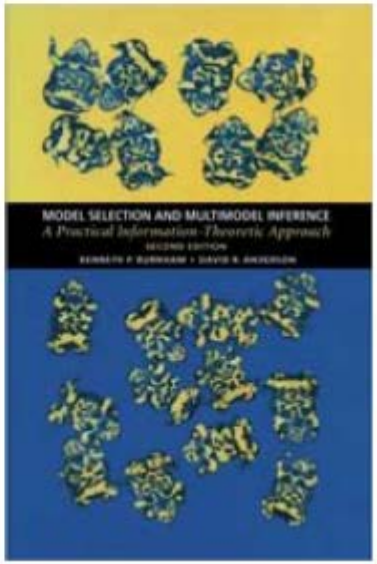
Opinions on variable selection

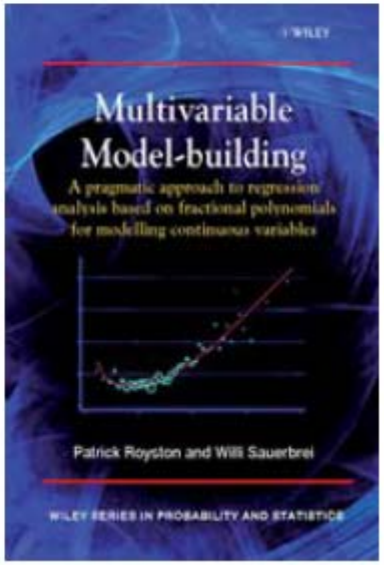
for models with focus on prediction and explanation.

Variable selection









(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

(Traditional) methods for variable selection

Full model

- variance inflation in the case of multicollinearity
  - Wald-statistic

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

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

All subset selection $\Rightarrow$ which criteria?

- C_p Mallows
- AIC Akaike Information Criterion
- BIC Bayes Information Criterion

Bayes variable selection

MORE OR LESS COMPLEX MODELS?

Stepwise procedures

Central Issue:

- significance level
  choice depends on aim of the study

Criticism

- FS and StS start with ‚bad' univariate models (underfitting)
- BE starts with the full model (overfitting),
  less critical
- Multiple testing, P-values incorrect

Nevertheless very popular in practice

Other procedures

- Bootstrap selection
- Change-in-estimate
- Variable clustering
- Incomplete principal components
- Penalized approaches (selection and shrinkage; Lasso, Garotte, SCAD, …)
- Directed acyclic graph (DAG-) based selections
-
-
-

"Recommendations" from the literature

We do not know any recommendation which is supported by good evidence from theory or meaningful simulation studies

3. Approaches for the selection of functional forms

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

Functional forms: the problem (1)

"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

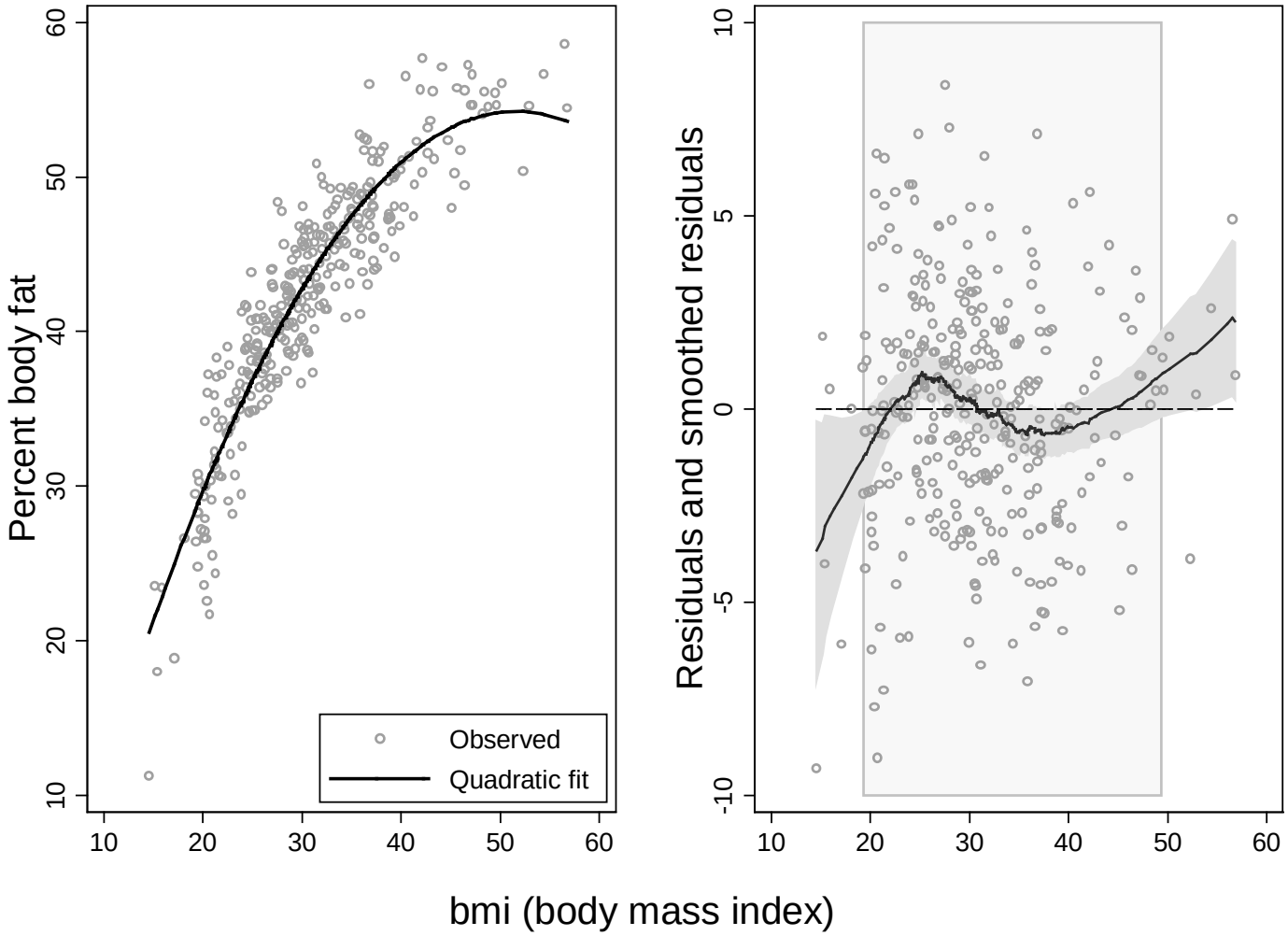
Discussion of issues in (univariate) modelling with splines

Trivial nowadays to *fit* almost any model

To *choose* a good model is much harder

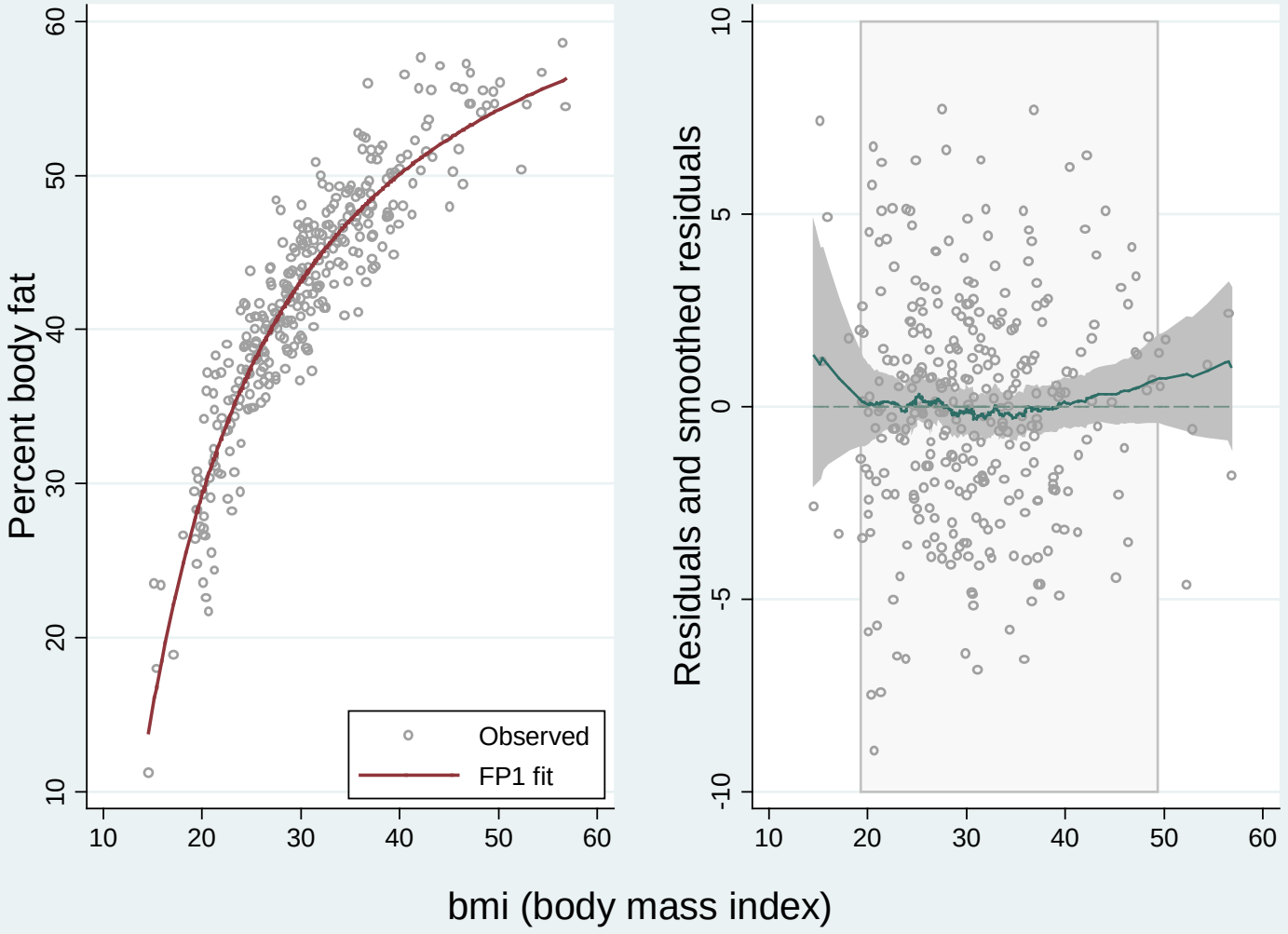
Functional forms: the problem (2)

Body fat data: quadratic model fits the data badly



Functional forms: a possible solution

Fractional polynomial does better

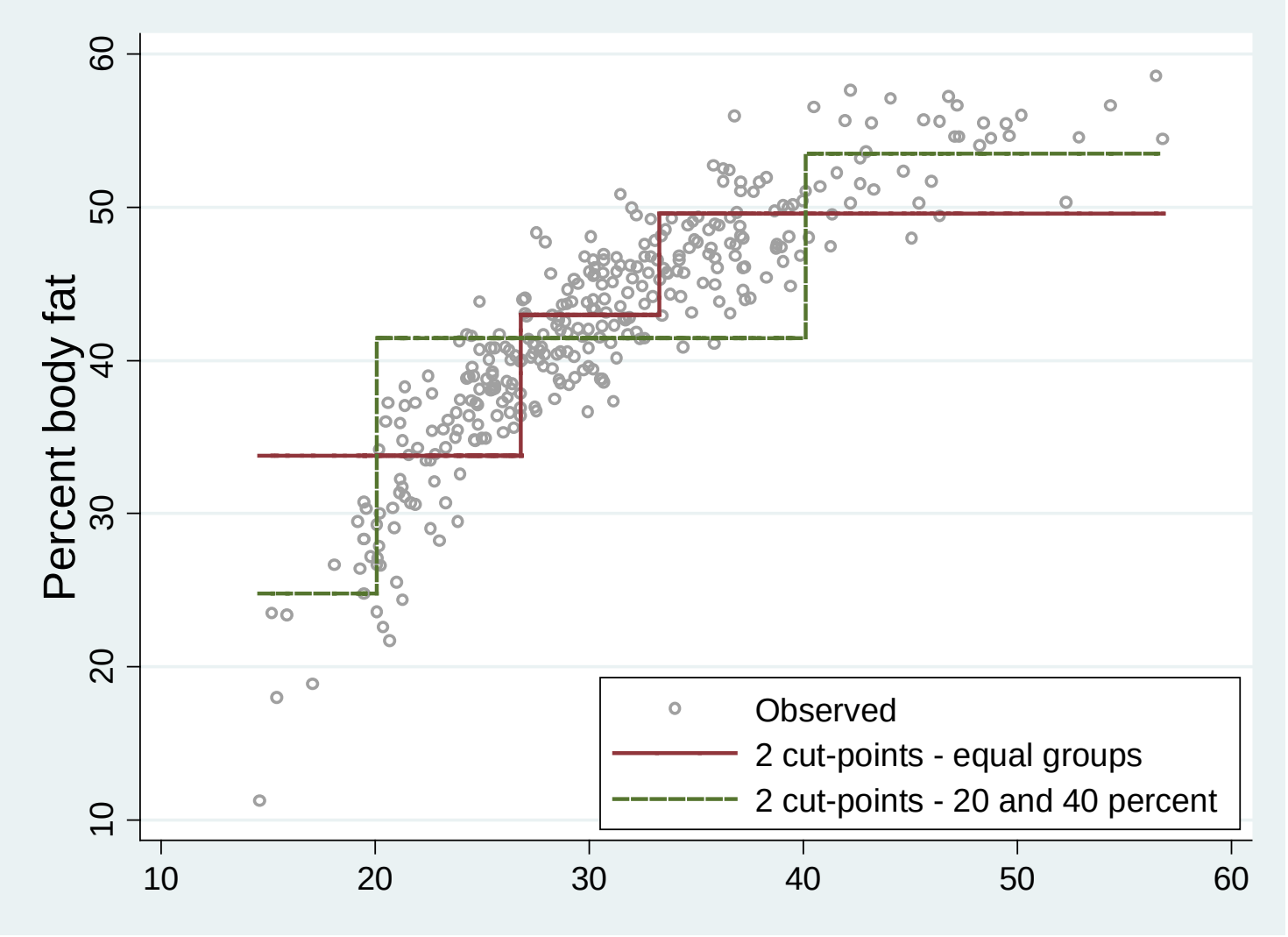


Functional forms: Models based on cut-points: problems!

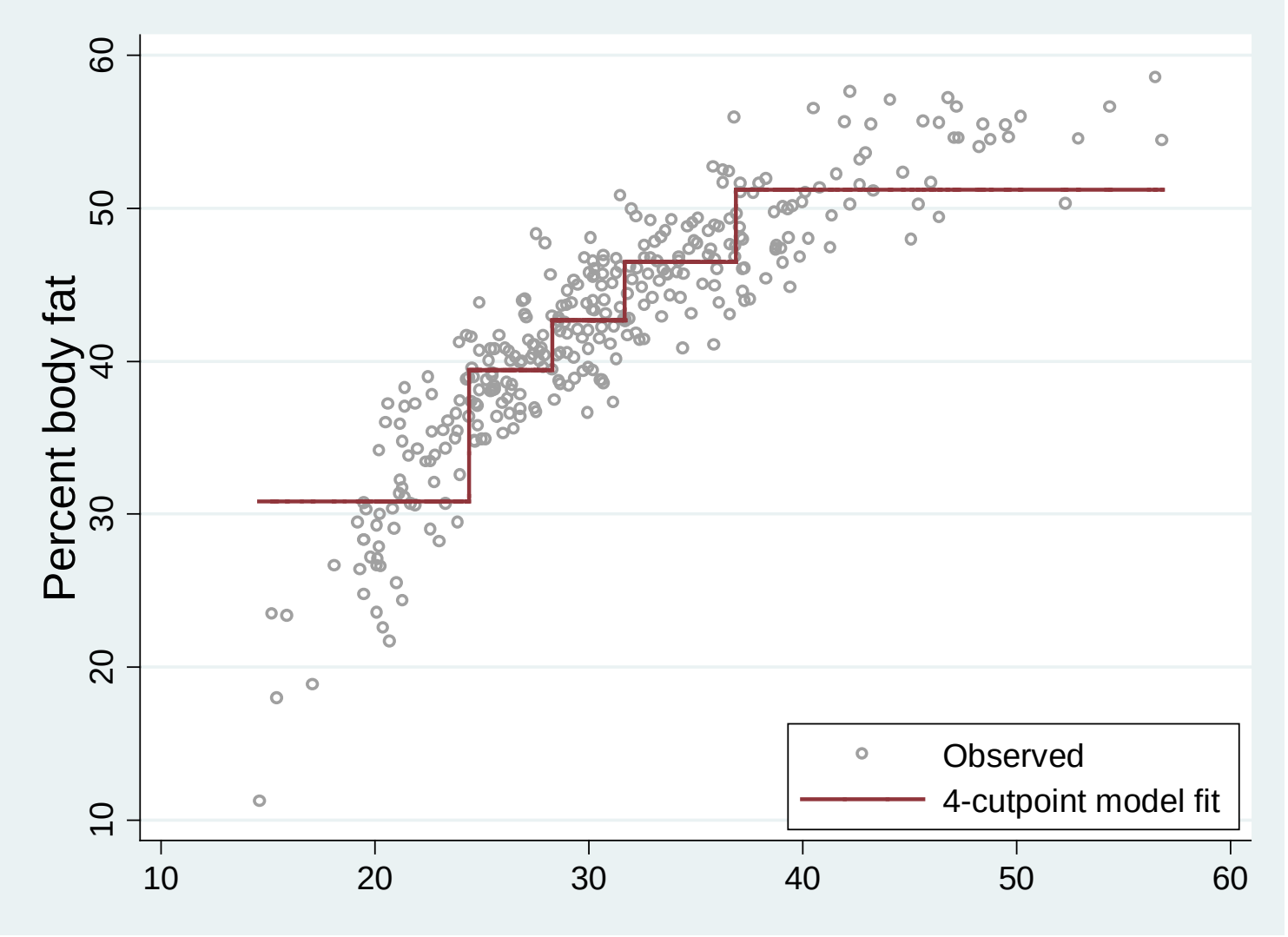
- Cut-points are still popular in clinical and epidemiological research
- Use of cut-points in a model gives a step function
- How many cut-points?
- Where should the cut-points be put?
- Biologically implausible step functions are a poor approximation to the true relationship
- Almost always fits the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

- Nevertheless, in many areas still the preferred approach!

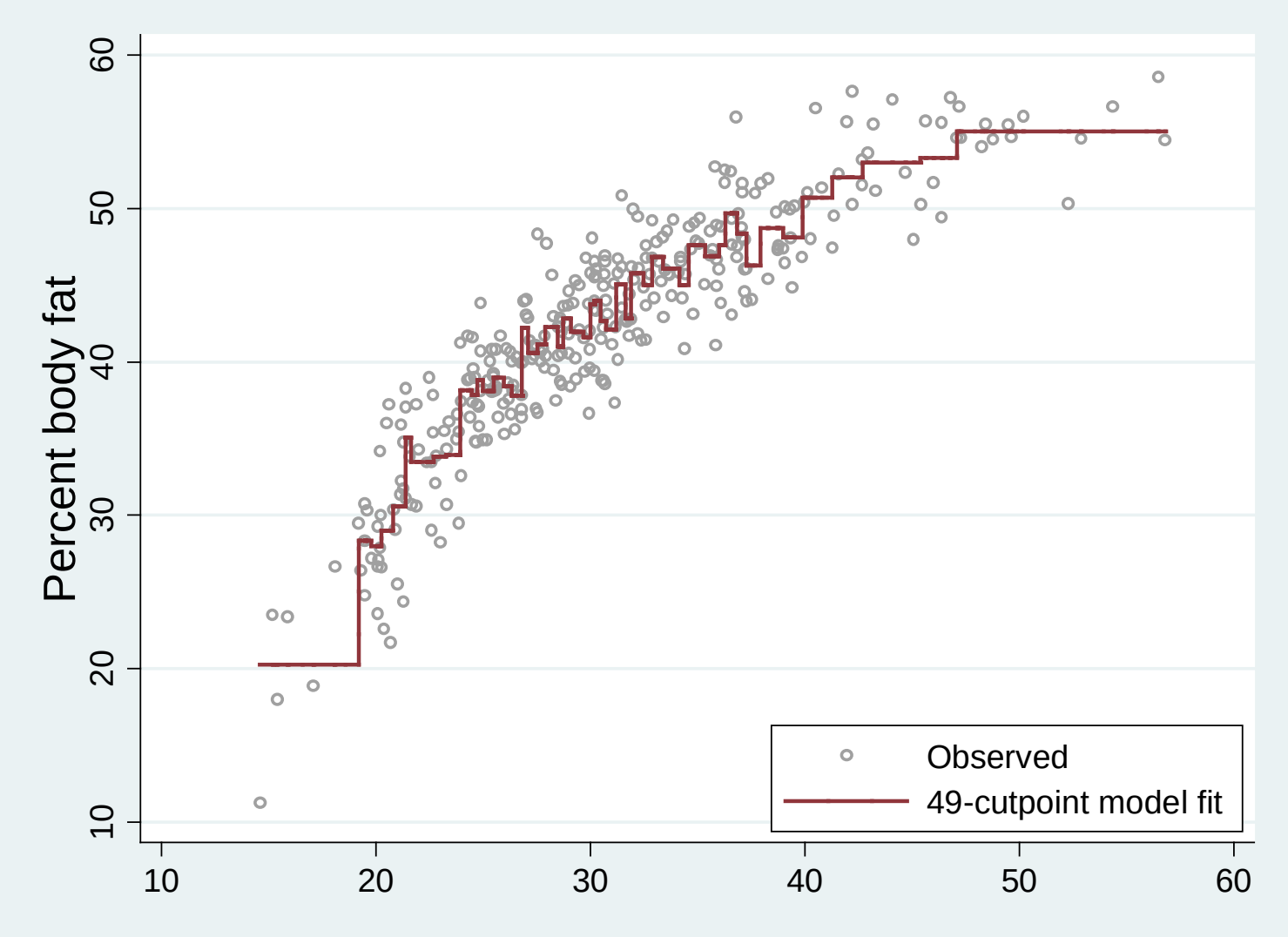
Body fat data (1) – two cutpoints



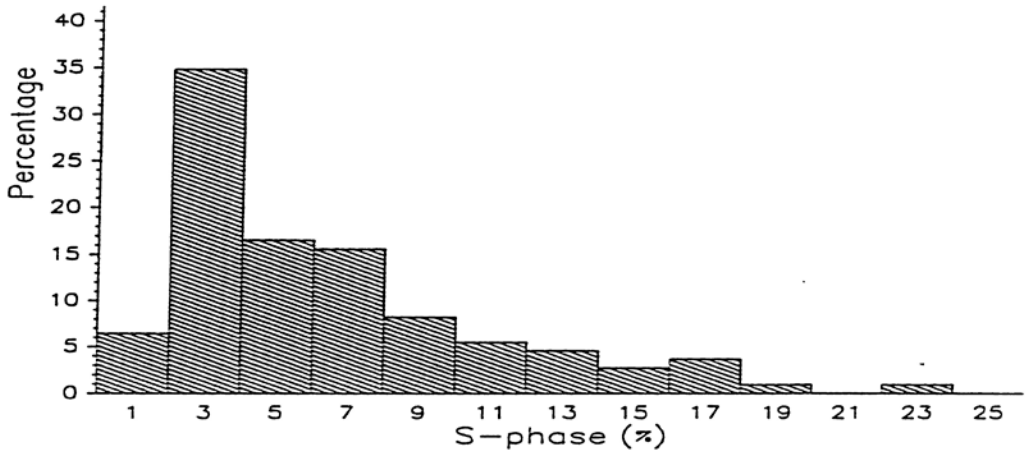
Body fat data (2) – four cutpoints

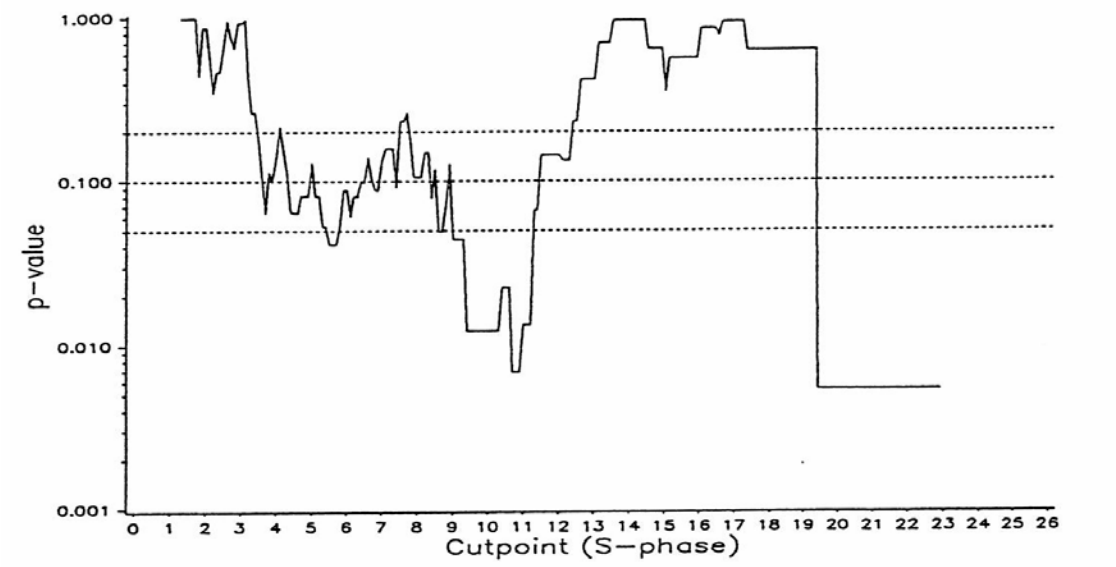


Body fat data (3) – 49 cutpoints



‘Optimal’ cutpoint (better: minimal P-value approach)





Optimal cutpoints: problems!

- Multiple testing $\Rightarrow$ inflation of significance level
  - 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
- Never reproducible – impossible to summarize across studies

4. Flexible modelling of the functional forms for continuous predictors

- Many approaches and many open issues

- Talk by Aris Perperoglou on spline based approaches

5. Combining variable and function selection

Two inter-related questions, common to many multivariable explanatory models

Results of

- Data-dependent selections of independent variables

may depend on

- decisions regarding functional forms of both
  1. the variable of interest (X)
  2. other variables, correlated with X

and *vice versa*

Combining variable and function selection

- Multivariable fractional polynomials (MFP)
- Various spline based approaches

Comparison in a large simulation study (Binder et al., 2013)

Nevertheless, much more research is needed!

6. State of the art – research required!

- Which strategies for variable selection exist?
  What about their properties?
- Data-dependent modeling introduces bias.
  What about the role of shrinkage approaches?
- Comparison of spline procedures in a univariate context.
  Which criteria are relevant? Can we derive guidance for practice?
- What about variables with a ‘spike-at-zero’?
- Multivariable procedures
  MFP well defined strategy
  Which of the spline based procedures?
  Comparison in large simulation studies needed
- Multivariable procedures and correction for selection bias
  How relevant? One step or two step approaches?
  E.g. selection of variables and forms followed by shrinkage
- Big Data
  Does it influence properties of procedures and their comparison?
- Role of model validation