

Steps to derive guidance for variable and function selection

Willi Sauerbrei*

Institute for Medical Biometry and Medical Informatics,
Medical Center - University of Freiburg, Germany

***for the STRATOS Topic Group TG2**

(Selection of variables and functional forms in multivariable analysis):

Michal Abrahamowicz, Heiko Becher, Harald Binder, Frank E. Harrell, Georg Heinze, Matthias Schmid, Aris Perperoglou, Patrick Royston, Willi Sauerbrei

Overview

- Background of the STRATOS initiative
- TG 2 – Variable and function selection
 - Issues in
 - Variable selection
 - Function selection for continuous variables
 - Requirements for evidence supported guidance

Statistical methodology - Current situation

- Statistical methodology has seen some substantial development
- Computer facilities can be viewed as the cornerstone
- Possible to assess properties and compare complex model building strategies using simulation studies
- Resampling and Bayesian methods allow investigations that were impossible two decades ago
- Wealth of new statistical software packages allow a rapid implementation and verification of new statistical ideas

Unfortunately, many sensible improvements are ignored in practical statistical analyses

Reasons why improved strategies are ignored

- Overwhelming concern with **theoretical aspects**
- Very **limited guidance** on key issues that are **vital in practice**, discourages analysts from utilizing more sophisticated and possibly more appropriate methods in their analyses

Improvement

At least two tasks are essential

- **Experts** in specific methodological areas have to work towards **developing guidance documents**
- An ever-increasing need for **continuing education** at all stages of the career
- For busy applied researchers it is often difficult to follow methodological progress even in their principal application area
 - Reasons are diverse
 - Consequence is that analyses are often deficient
- **Knowledge** gained through research on statistical methodology needs to be **transferred** to the broader community
- Many **analysts** would be **grateful for** an overview on the current **state of the art** and for **practical guidance documents**

Aims of the initiative

- **Provide guidance documents** for highly relevant issues in the design and analysis of observational studies
- As the statistical **knowledge** of the analyst **varies** substantially, guidance has to keep this background in mind. **Guidance** documents have to be provided **at several levels**
- For the **start** we will concentrate on **state-of-the-art** documents and the necessary evidence
- Help to identify questions requiring much more primary research

The overarching long-term aim is to improve key parts of design and statistical analyses of observational studies in practice

STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative

Willi Sauerbrei,^{a*†} Michal Abrahamowicz,^b
Douglas G. Altman,^c Saskia le Cessie,^d and[‡] James Carpenter^e
on behalf of the STRATOS initiative

Statistics in Medicine 2014

2011	ISCB Ottawa, Epidemiology Sub-Comm.	Preliminary ideas
2012	ISCB Bergen	Discussions, SG
2013	ISCB Munich	Initiative launched
2014-16	ISCB	Invited Sessions

<http://www.stratos-initiative.org/>

Basic information

Topic Group		Chairs and further members	
1	Missing data	Chairs:	James Carpenter, Kate Lee
		Members:	Melanie Bell, Els Goetghebeur, Joe Hogan, Rod Little, Andrea Rotnitzky, Kate Tilling, Ian White
2	Selection of variables and functional forms in multivariable analysis	Chairs:	Michal Abrahamowicz, Aris Perperoglou, Willi Sauerbrei
		Members:	Heiko Becher, Harald Binder, Frank Harrell, Georg Heinze, Patrick Royston, Matthias Schmid
3	Initial data analysis	Chairs:	Marianne Huebner, Saskia le Cessie, Werner Vach
		Members:	Maria Blettner, Dianne Cook, Heike Hofmann, Hermann-Josef Huss, Lara Lusa
4	Measurement error and misclassification	Chairs:	Laurence Freedman, Victor Kipnis
		Members:	Raymond Carroll, Veronika Deffner, Kevin Dodd, Paul Gustafson, Ruth Keogh, Helmut Küchenhoff, Pamela Shaw, Janet Tooze
5	Study design	Chairs:	Mitchell Gail
		Members:	Doug Altman, Gary Collins, Luc Duchateau, Neil Pearce, Peggy Sekula, Elizabeth Williamson, Mark Woodward
6	Evaluating diagnostic tests and prediction models	Chairs:	Gary Collins, Carl Moons, Ewout Steyerberg
		Members:	Patrick Bossuyt, Petra Macaskill, Ben van Calster, Andrew Vickers
7	Causal inference	Chairs:	Els Goetghebeur
		Members:	Bianca De Stavola, Saskia le Cessie, Niels Keiding, Erica Moodie, Ingeborg Waernbaum, Michael Wallace
8	Survival analysis	Chairs:	Michal Abrahamowicz, Per Kragh Andersen, Terry Therneau
		Members:	Richard Cook, Pierre Joly, Torben Martinussen, Maja Pohar-Perme, Jeremy Taylor
9	High-dimensional data	Chairs:	Lisa McShane, Joerg Rahnenfuehrer
		Members:	Axel Benner, Harald Binder, Anne-Laure Boulesteix, Tomasz Burzykowski, W. Evan Johnson, Lara Lusa, Stefan Michiels, Sherri Rose

Cross-cutting panels

Panels		Chairs
1	Glossary (GP)	Simon Day, Marianne Huebner, Jim Slattery
2	Data Sets (DP)	Saskia Le Cessie, Aris Perperoglou, Hermann Huss
3	Publications (PP)	Stephen Walter
		Co- Chairs: Bianca De Stavola, Mitchell Gail, Petra Macaskill
4	New Membership (MP)	James Carpenter, Willi Sauerbrei
5	Website (WP)	Joerg Rahnenfuehrer, Willi Sauerbrei
6	Literature Review (RP)	Gary Collins, Carl Moons
7	Simulation Studies (SP)	Michal Abrahamowicz, Harald Binder
8	Contact with Other Societies and Organizations (OP)	Willi Sauerbrei
9	Knowledge Transfer (TP)	Suzanne Cadarette

On requirements for an evidence supported guidance document

—

Issues in variable and function selection

(consider low dimensional data and not 'too small' sample sizes)

TG2: Selection of variables and functional forms in multivariable analysis

In multivariable analysis, it is common to have a **mix of binary, categorical (ordinal or unordered) and continuous variables** that may influence an outcome. While **TG6** considers the situation where the **main task is predicting the outcome** as accurately as possible, the main focus of **TG2** is to **identify influential variables** and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) **main challenges** are **selection of variables** for inclusion in a multivariable explanatory model and **choice of the functional forms for continuous variables**.

[...] The effects of **continuous predictors are typically modeled by either categorizing** them (which raises such issues as the number of categories, cutpoint values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cutpoints) **or assuming linear relationships** with the outcome, possibly after a simple transformation (e.g. logarithmic or quadratic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed and the **validity of the underlying assumptions is not assessed**.

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including **fractional polynomials** and **several 'flavors' of splines**.

[...] collaborations with other TGs to account for such **complexities** as **missing data, measurement errors, time-varying confounding** or issues specific to modeling continuous predictors in survival analyses.

TG2: Part 1 - Selection of variables

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

The following slides are taken from the 'Statistics in Practice' presentation at the meeting of the German Region of the Biometric Society, March 2016

<http://www.biometrische-gesellschaft.de/arbeitsgruppen/weiterbildung/education-for-statistics-in-practice.html>

Education for Statistics in Practice, DAGStat 2016

Variable selection – a review and recommendations for the practicing statistician


Updated version!

Georg Heinze & Daniela Dunkler
Medical University of Vienna
CeMSIIS – Section for Clinical Biometrics

georg.heinze@meduniwien.ac.at, daniela.dunkler@meduniwien.ac.at


Focus of this presentation:

- Methods and consequences of variable selection



Complexity is your enemy. Any fool can make something complicated. It is hard to keep things simple.

Sir Richard Branson
founder of Virgin Group

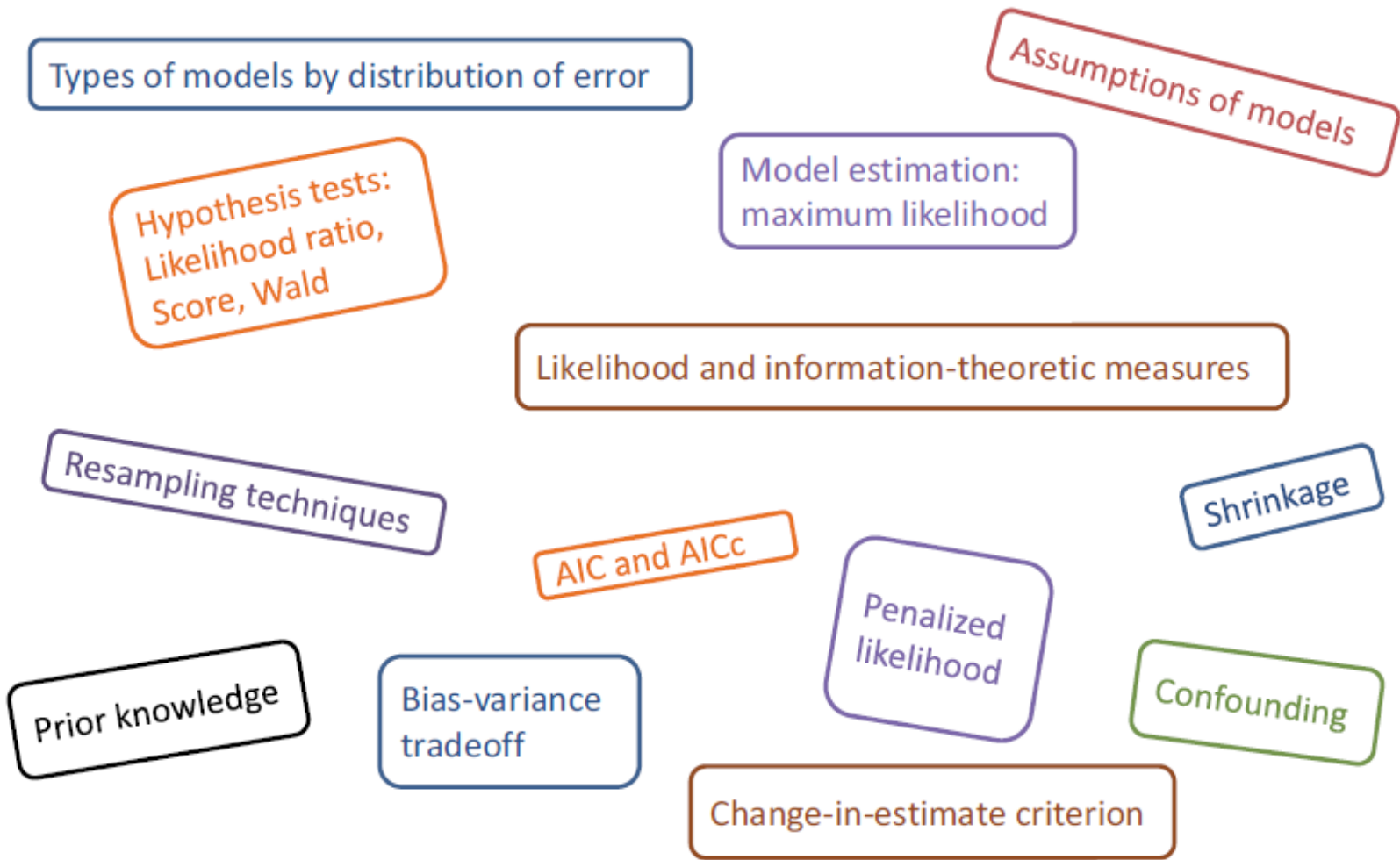


When I do my own makeup, I keep things pretty simple.

(Jordana Brewster)

izquotes.com

Statistical prerequisites



Basic algorithms

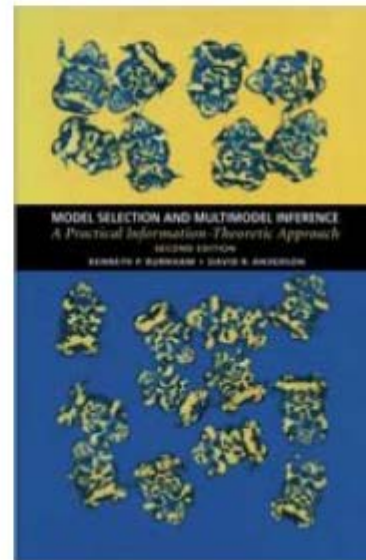
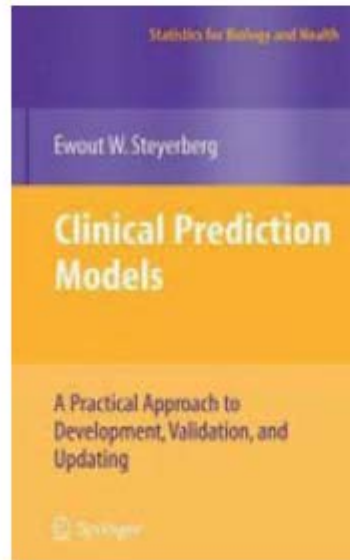
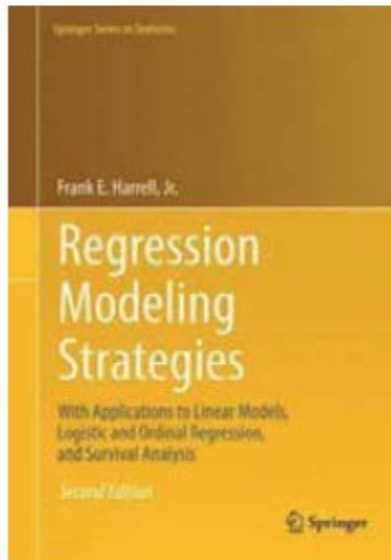
- 'Full' model
- Univariable filtering
- Best subset selection
- Forward selection
- Backward elimination
- Change-in-estimate: Purposeful variable selection and augmented backward selection
- Information-theoretic approach
- Directed acyclic graph (DAG)-based selection

Opinions on variable selection

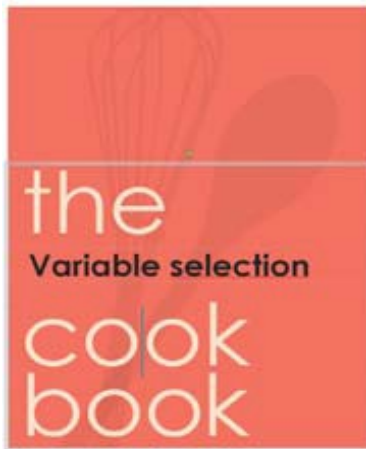
for models with focus on prediction and explanation.



Variable selection



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



Recipe for disaster

- Prepare a long list of poorly conceived predictors.
- Add only small n .
- Mix together in an extensive iterative data dredging.
- Select the model with the smallest p-values.
- Present this final model without further considerations.

Bon appétit!



TG2: Part 2 - Selection of functions for continuous variables

- Categorization of continuous variables is still very popular (despite well known weaknesses)
- Linearity is often assumed without checking
- Often better alternatives
 - Fractional polynomials
 - Splines

Fractional polynomials

Fractional polynomials and the multivariable fractional polynomial (MFP) approach

Royston and Altman (1994)

Sauerbrei and Royston (1999)

Royston and Sauerbrei (2008)

The MFP approach combines

- *Selection of variables by using backward elimination (BE) with*
- *Selection of fractional polynomial (FP) functions of continuous variables*

Although relatively simple and easily understood by researchers familiar with the basics of regression models, the selected models often extract most of the important information from the data. Models derived are **relatively easy to interpret and to report, a pre-requisite for transportability and general use in practice.**

Easy to use software is available.

<http://mfp.imbi.uni-freiburg.de/>

Splines

Several approaches. Guidance is urgently needed.

Multivariable Regression Modelling

A review of available spline packages in R.

Aris Perperoglou for TG2

ISCB 2015

Splines

- Splines are piecewise polynomial functions
 - Given the range of a continuous variables, define points on this interval (knots)
- Fit a simple polynomial between these knots.
- Depending on the selection of knots and the type of polynomial the type of spline is named as:
 - polynomial, natural, restricted regression splines (**de Boor 1978, Harrel 2013**)
 - b-splines (**de Boor 1978**), p-splines (**Marx and Eilers 1996**), penalized regression splines (**Wood 2006**)
 - smoothing splines, Generalized Additive Models (**Hastie and Tibshirani 1990**)...

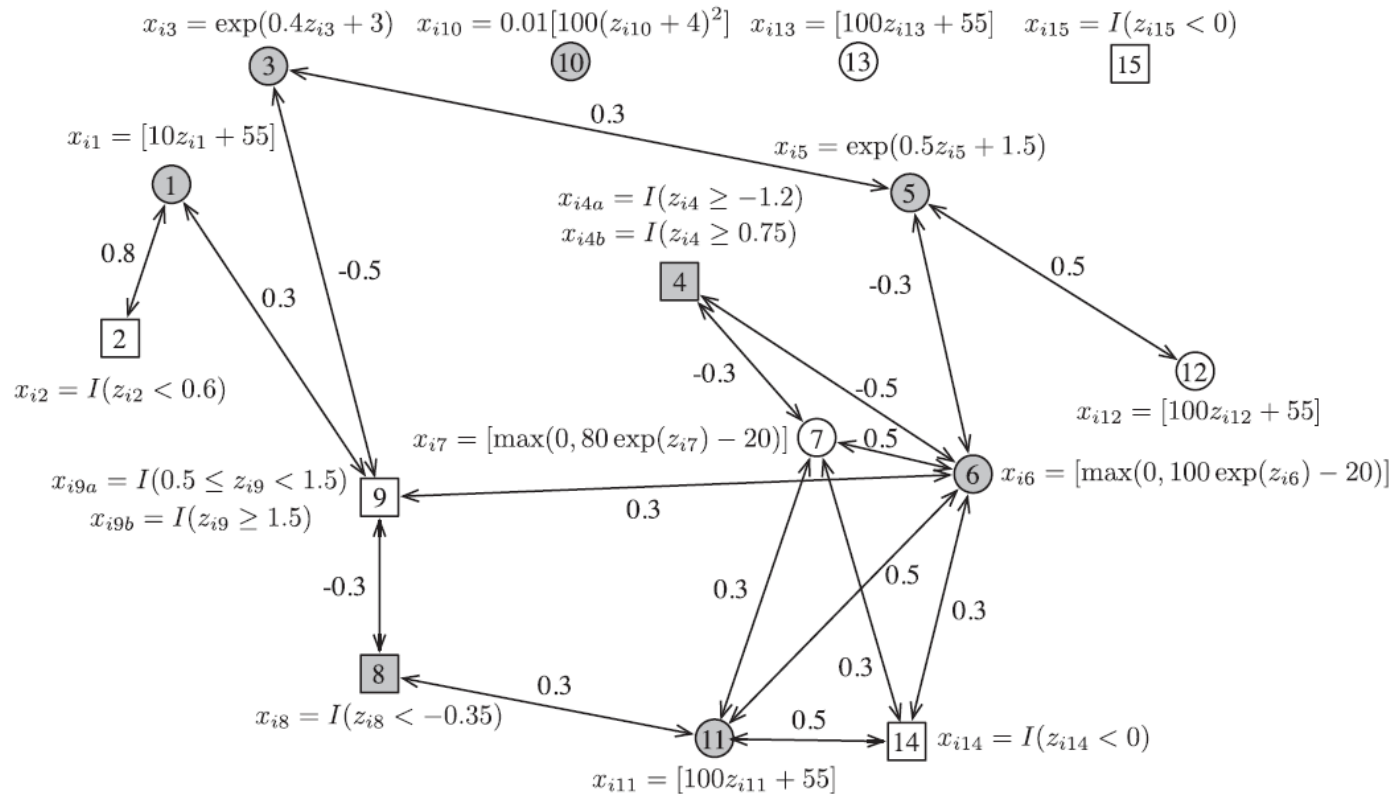
Decisions required for spline modelling

- Although splines are a powerful tool, in practice the number of decisions that the user has to make complicate the modelling problem.
 - Type of spline
 - Number of knots
 - Position of knots
 - Order of the spline (complexity)

Comparisons of approaches required

Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response

Harald Binder, Willi Sauerbrei and Patrick Royton; Statistics in Medicine, 2013



Simulation studies: Simple designs do often not help

Steps towards guidance documents

Selection of multivariable models for explanation (TG2)

- **Strategies for variable selection**
 - Better understanding of advantages and disadvantages
 - Role of model complexity, stability and shrinkage
- **Review of the literature about methods**
 - Strategies used in practice
 - Comparison of strategies for model building
- **Comparison of spline procedures**
- **Specific role of 'spike at zero' variables?**
- **Comparison of approaches for variable selection and choices of functional form**
- **Guidance documents for variable and function selection**

Summary

Many issues...

- A large number of variable selection strategies has been proposed
- There are several spline based procedures

...how to derive evidence to support guidance documents

- Theoretical investigations?
- Large and meaningful simulation studies!!!
- Good examples