UNIVERSITÄTS KLINIKUM FREIBURG

Lecture 2
Multivariable model-building – general considerations and issues in variable selection

Willi Sauerbrei, Edwin Kipruto

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

WS 2025/26

Learning objectives 1

To understand:

- ‘Full’ model is unsuitable if ‘more than a predictor’ is required
- Many variable selection strategies are available. For none of them properties are known. All have problems.
- Aim of a study strongly influences analysis strategy (describe, predict, explain)
- Model complexity is a key criterion
- Backward elimination is often a suitable procedure
- Parameter estimates after variable selection are biased
- Combine selection and shrinkage
- Often simple predictors have advantages

- Low-dimensional data in mind

Plan 1

1. General issues of building regression models
2. Methods of variable selection
3. Examples – Different procedures select different models
4. Bias introduced by variable selection
5. Selection bias
   - Can we correct by shrinkage?
   - Combine variable selection and shrinkage
6. Complexity of predictors
7. Handling categorical predictors
8. Summary

Implicit assumptions (for this lecture)

- Subject matter knowledge (if available) determines (parts) of the model
- About 5 to 30 candidate variables
- No ‚small sample size' situation
- Assumption of linear effect for continuous variables is acceptable (relaxed later).
- No missing data problem

1. General issues of building regression models

- Regression models in observational studies
- Under- and overfitting
- Before model building starts – importance of initial data analysis

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

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

Central issue

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

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

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler

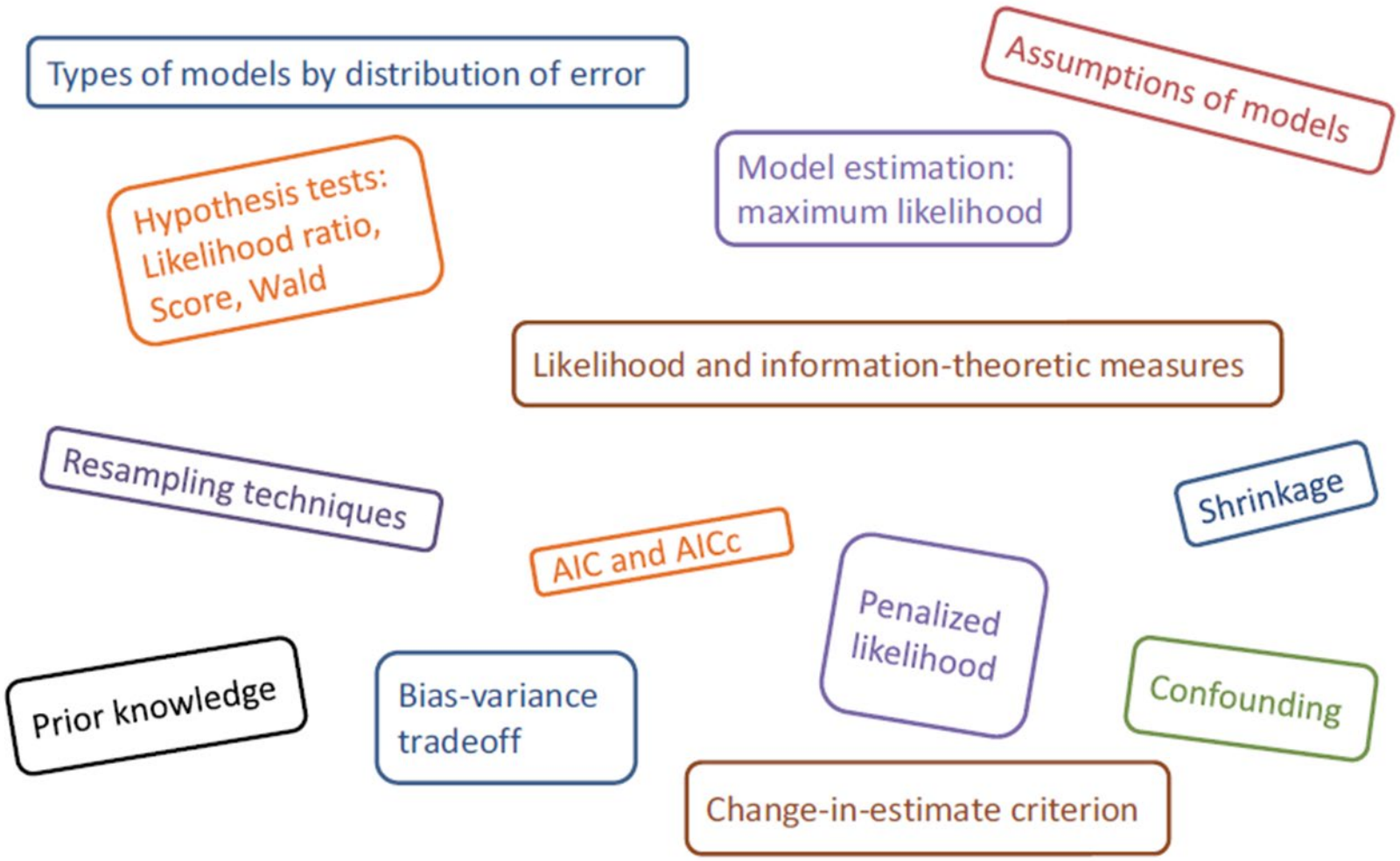
Biometrical Journal. 2018;60:431–449.

Focus of this presentation

- Methods and consequences of variable selection

Heinze & Dunkler, 03-2016

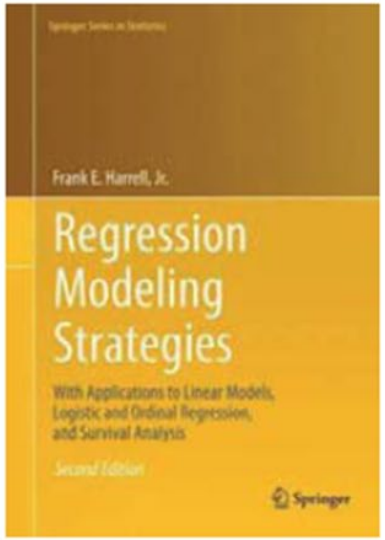
Statistical prerequisites

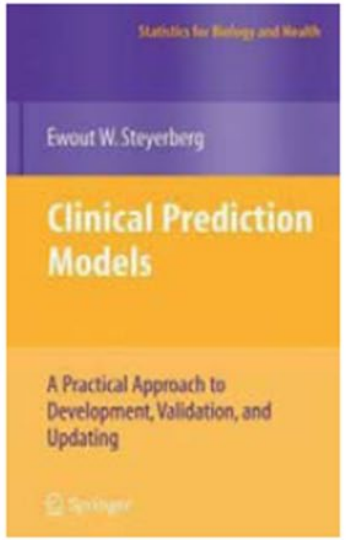


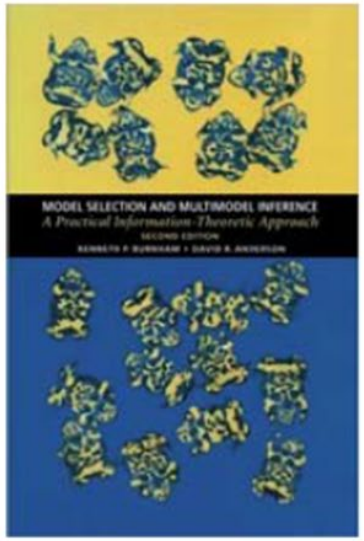
Opinions on variable selection

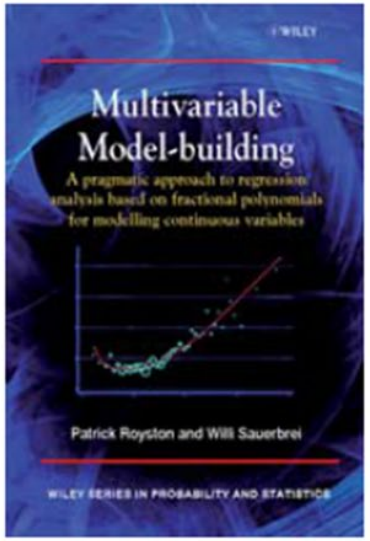
for models with focus on prediction and explanation.

Variable selection





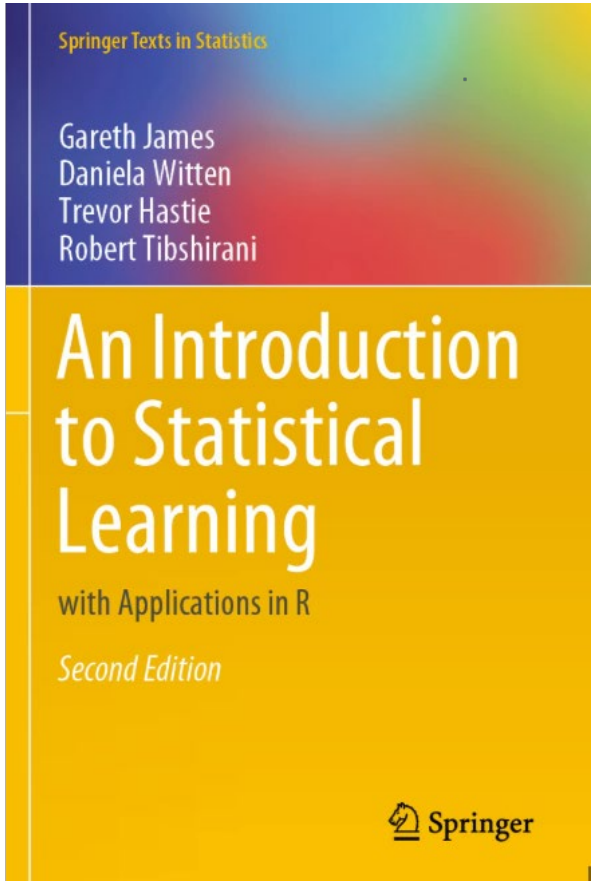


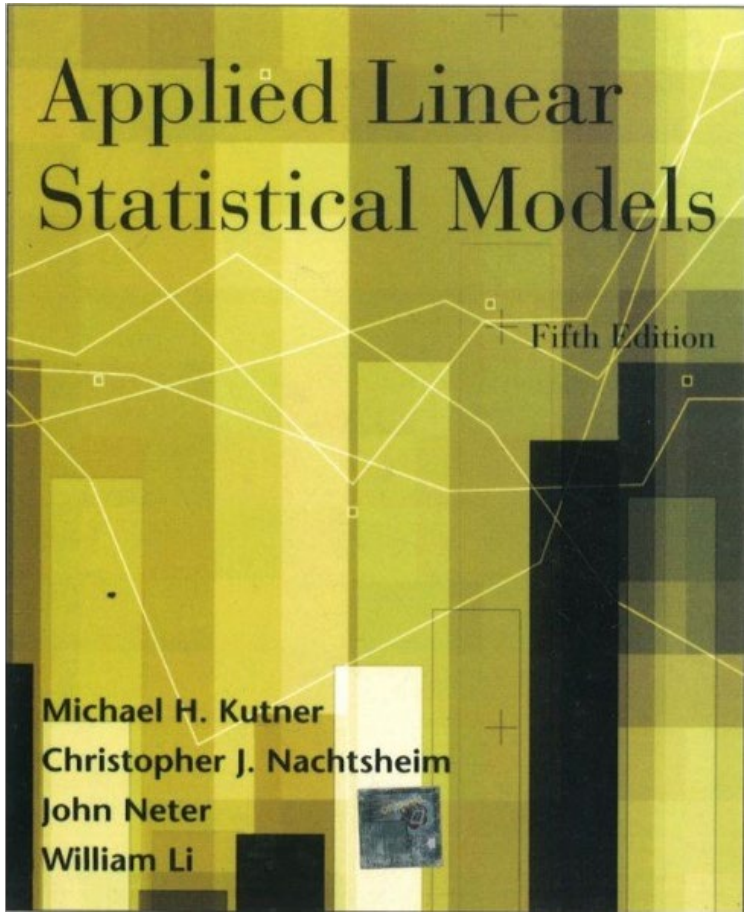


(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

→Emphasis on different aims
→ Different philosophies
→Joint work (except B&A) **in STRATOS!**

Further books about model building





Which variables should be included?

Effect of underfitting and overfitting

3 predictors

Correct model M_1	$y =$	$1\,x_1 + 2x_2 + 0\,x_3 + \varepsilon$
M2 overfitting	$y =$	$ß_1x_1 + ß_2x_2 + ß_3x_3 + \varepsilon$
M3 underfitting	$y =$	$ß_2x_2 + \varepsilon$

Underfitting and overfitting

3 predictors:

$ß_1$= 1, $ß_2$= 2, $ß_3$= 0; $r_{1,2}$ = 0.5, $r_{1,3}$ = 0, $r_{2,3}$ = 0.7;

N = 400, σ^2 = 1

Means of 3 runs

	M_1 (true)	M_2 (overfitting)	M_3 (underfitting)
$\hat{\beta}_1$	1.050 (0.059)	1.04 (0.073)	-
$\hat{\beta}_2$	1.950 (0.060)	1.98 (0.105)	2.53 (0.068)
$\hat{\beta}_3$	-	-0.03 (0.091)	-
$\hat{\sigma}^2$	1.060	1.060	1.90
R^2	0.875	0.875	0.77

M_2 (overfitting) Standard errors larger (variance inflation)

M_3 (underfitting) ‚biased', different interpretation, R^2 smaller, stand. error (VIF↓, σ^2 ↑)?

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

Classes of multivariable models

1. The model is predefined. All that remains is to estimate the parameters and check the main assumptions.
2. The aim is to develop a good predictor. The number of variables should be small.
3. The aim is to develop a good predictor. Limiting the model complexity is not important.
4. The aim is to assess the effect of one or several (new) factors of interest, adjusting for some established factors in a multivariable model.
5. The aim is to assess the effect of one or several (new) factors of interest, adjusting for confounding factors determined in a data-dependent way by multivariable modelling.
6. Hypothesis generation of possible effects of factors in studies with many covariates.

Building multivariable regression models – Preliminaries 1

- ‚Reasonable' model class was chosen
- Comparison of strategies
  - Theory
    - only for limited questions, unrealistic assumptions
  - Examples or simulation
    - Examples from literature
      - oversimplifies the problem
      - data clean
      - ‚relevant' predictors given
      - number predictors managable

Building multivariable regression models – Preliminaries 2

- Data from defined population, relevant data available (‚zeroth problem', Mallows 1998)

- Examples based on published data
  - rigorous pre-selection → what is a full model?

Building multivariable regression models – Preliminaries 3

- Several ‚problems' need a decision before the analysis can start
  - Eg. Blettner & Sauerbrei (1993), searching for hypotheses in a case-control study (more than 200 variables available)

Problem 1. *Excluding variables prior to model building.*
Problem 2. *Variable definition and coding.*
Problem 3. *Dealing with missing data.*
Problem 4. *Combined or separate (eg. by sex) models*.
Problem 5. *Choice of nominal significance level and selection procedure.*

Building multivariable regression models – Preliminaries 4

- More problems are available,
  - see discussion on initial data analysis in Chatfield (2002) section ‚*Tackling real life statistical problems*‘ and

    Mallows (1998)

‚*Statisticians must think about the real problem, and must make judgements as to the relevance of the data in hand, and other data that might be collected, to the problem of interest ... one reason that statistical analyses are often not accepted or understood is that they are based on unsupported models. It is part of the statistician's responsibility to explain the basis for his (or her) assumption.*‘

STRATOS initiative - TG 3 Initial Data Analysis

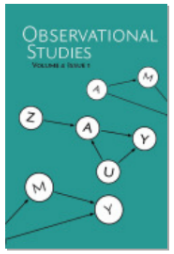
STRengthening Analytical Thinking for Observational Studies

A Contemporary Conceptual Framework for Initial Data Analysis

Marianne Huebner, Saskia le Cessie, Carsten O. Schmidt, Werner Vach

Observational Studies, Volume 4, Issue 1, 2018, pp. 171-192 (Article)

Published by University of Pennsylvania Press
DOI: https://doi.org/10.1353/obs.2018.0014



Huebner *et al.*, 2018

Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses

Marianne Huebner[1,2*], Werner Vach[3], Saskia le Cessie[4], Carsten Oliver Schmidt[5], Lara Lusa[6,7] and on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, http://www.stratos-initiative.org)

Huebner *et al.*, 2020

Mark Baillie[1], Saskia le Cessie[2], Carsten Oliver Schmidt[3], Lara Lusa[4], Marianne Huebner[5]*, for the Topic Group "Initial Data Analysis" of the STRATOS Initiative[¶]

1 Novartis, Basel, Switzerland, **2** Department of Clinical Epidemiology and Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands, **3** Institute for Community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany, **4** Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technology, University of Primorska, Koper, Slovenia, **5** Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, United States of America

¶ Membership of the STRATOS Initiative is provided in the Acknowledgments.
* huebner@msu.edu

Baillie *et al.*, 2022

Regression without regrets –initial data analysis is a prerequisite for multivariable regression

Georg Heinze[1*], Mark Baillie[2], Lara Lusa[3,4], Willi Sauerbrei[5], Carsten Oliver Schmidt[6], Frank E. Harrell[7], Marianne Huebner[8] on behalf of TG2 and TG3 of the STRATOS initiative

Heinze *et al.*, 2024

2. Methods of variable selection

- Overview
- ‘Recommendations’
- ‘Full’ model?
- Backward elimination is a sensible approach

Multivariable models – 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?

WHAT ABOUT THE FUNCTIONAL FORM?

Stepwise procedures

Central Issue: significance level

Criticism

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

All subset selection

- Minimize an information criterion over the 2^k available models
  - For a given model M
  - IC= -2l +a dim (M)

All subset selection
(normal errors regression model)

criteria for best model

- fixed number of covariables: $R^2 = 1 - (SSE / SYY)$

- models with different number of covariables (p)

i)	Mallows' C_P	$= (SSE / - n \hat{\sigma}^2)$	$+ p\ 2$
ii)	Akaike's AIC	$= n \ln (SSE / n)$	$+ p\ 2$
iii)	BIC	$= n \ln (SSE / n)$	$+ p \ln (n)$
		fit	*penalty*

other criteria with minor variations

Several approaches transferred for generalized linear models and models for survival data

Other procedures

- Variable clustering
- Incomplete principal components
- Change-in-estimate
- Bootstrap selection
- Penalized likelihood (Lasso, Garotte, ...)
- Directed acyclic graph (DAG-) based selections

For an overview, see **Sauerbrei *et al.* (2020)**

Theoretical results for model building strategies

'Exact distributional results are virtually impossible to obtain, even for simplest of common subset selection algorithms'

Picard & Cook, JASA,1984

Traditional procedures

"Recommendations" from the literature

(up to 1990, after more than 20 years of use and research)

- **Mantel (1970)**

'... advantageous properties of the stepdown regression procedure (BE) ..., in comparison to StS

- **Draper & Smith (1981)**

'... own preference is the stepwise procedure. To perform all regressions is not sensible, except when there are few predictors'

- **Weisberg (1985)**

'Stepwise methods must be used with caution. The model selected in a stepwise fashion need not optimize any reasonable criteria for choosing a model. Stepwise may seriously overstate significance results'

- **Wetherill (1986)**

`Preference should be given to the backward strategy for problems with a moderate number of variables, in comparison to StS'

- **Sen & Srivastava (1990)**

'We prefer all subset procedures. It is generally accepted that the stepwise procedure (StS) is vastly superior to the other stepwise procedures'.

Harrell 2001, Regression Modeling Strategies

Stepwise variable selection ...if ... just been proposed ... likely be rejected because it violates every principle of statistical estimation and hypothesis testing.

... no currently available stopping rule was developed for data-driven variable selection. Stopping rules as AIC or Mallows´ C_p are intended for comparing only two *prespecified* models.

Full model fits have the advantage of providing meaningful confidence intervals using standard formulas

... Bayes several advantages ... → will not be considered

LASSO-Variable selection and shrinkage → later

… AND WHAT TO DO?

Full or selected model?

What is the 'full' model?

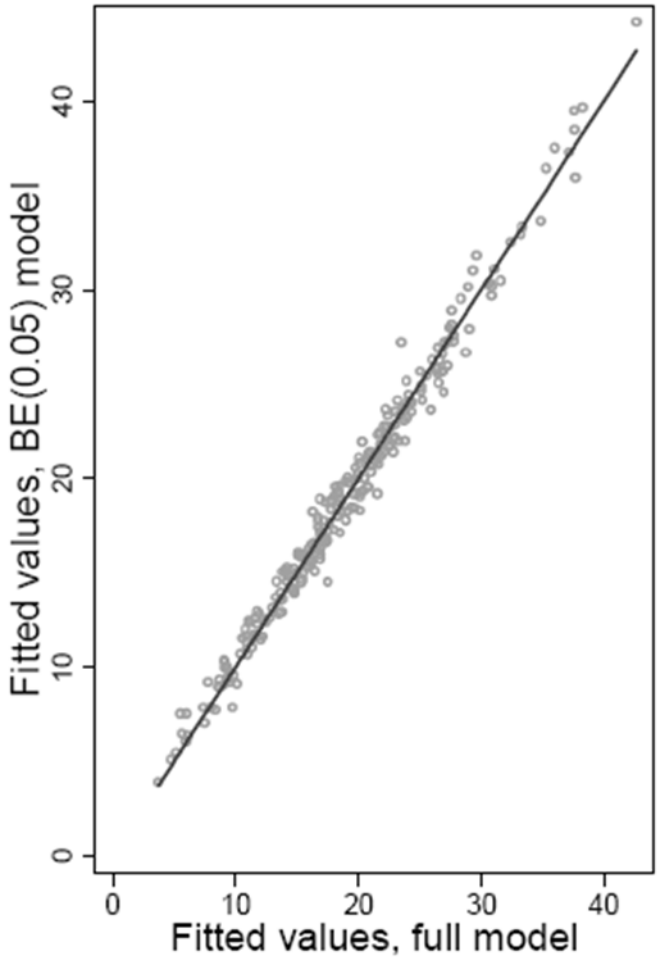
Var	Full model			BE (0.05)			Full model excluding x_6		
	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE	$\widehat{\beta}$	SE	$\widehat{\beta}$/SE
x_1	0.074	0.032	2.31	0.056	0.024	2.35	0.211	0.034	6.20
x_2	−0.019	0.067	−0.28				0.227	0.074	3.08
x_3	−0.249	0.191	−1.30	−0.322	0.121	−2.65	−0.915	0.212	−4.32
x_4	−0.394	0.234	−1.68				−0.378	0.278	−1.36
x_5	−0.119	0.108	−1.10				0.150	0.124	1.21
x_6	0.901	0.091	9.90	0.774	0.033	23.26	–	–	–
x_7	−0.146	0.144	−1.02				0.163	0.166	0.98
x_8	0.178	0.146	1.22				0.231	0.173	1.33
x_9	−0.041	0.245	−0.17				−0.095	0.291	−0.33
x_{10}	0.185	0.220	0.85				−0.053	0.259	−0.21
x_{11}	0.178	0.170	1.04				−0.066	0.200	−0.33
x_{12}	0.277	0.207	1.34				0.058	0.244	0.24
x_{13}	−1.830	0.529	−3.46	−1.943	0.406	−4.78	−2.692	0.620	−4.34

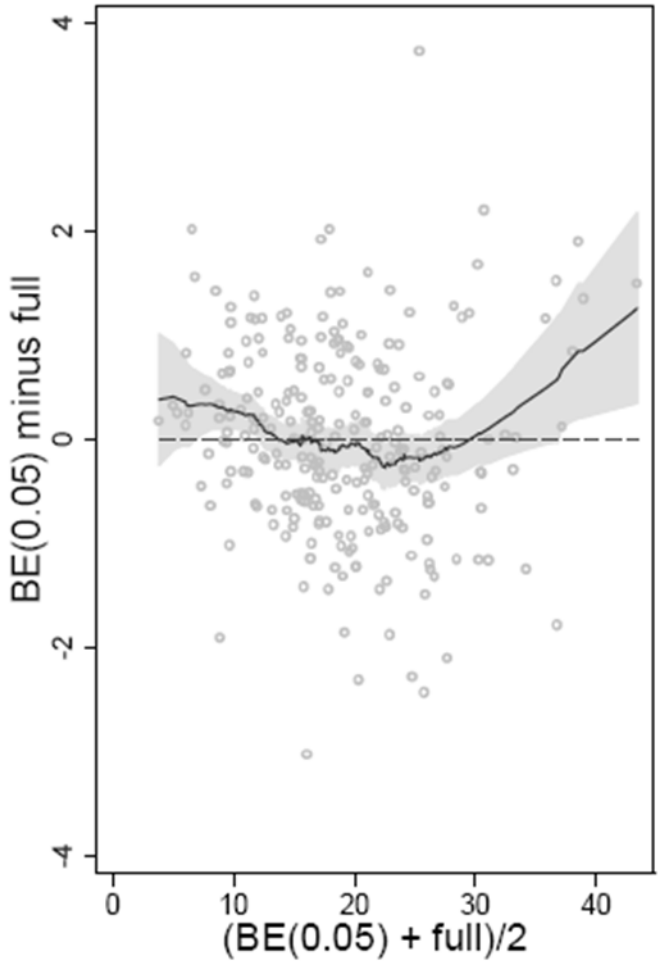
Educational body fat data (Royston & Sauerbrei 2008, Tab 2.3)

Var-	Full	Nominal significance level, α								All-subsets	
iable	model	0.01		0.05		0.10		0.157			
		BE	FS	BE	FS	BE	FS	BE	FS	AIC	BIC
x_1	*			✓		✓		✓	✓	✓	
x_2			✓		✓		✓		✓		✓
x_3		✓		✓		✓		✓		✓	
x_4								✓		✓	
x_5								✓		✓	
x_6	*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
x_7											
x_8									✓		
x_9											
x_{10}											
x_{11}							✓		✓	✓	
x_{12}								✓			
x_{13}	*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Educational body fat data (R&S 2008, Tab 2.4)

- Predictors from simple and complex models are often very similar





Educational body fat data (R&S 2008, Fig 2.1)

Bland-Altman plot

Variable selection

All procedures have severe problems!

$\Rightarrow$ Full model? No!

Illustration of problems

Too often with small studies

(sample size versus no. variables)

Arguments for the full model

Often by using published data

Heavy pre-selection!

What is the full model?

Type I error of selection procedures

Actual significance level (linear regression model)

For all-subset methods in good agreement with asymptotic results for one additional variable (Teräsvirta & Mellin, 1986)

- for moderate sample size only slightly higher than

BE	~	α_{in}
All-AIC	~	15.7 %
All-BIC	~	P (χ_1^2 > ln (n))
		0.032 N = 100
		0.014 N = 400

Increases with correlation to variable with effect (‚wrong' variable selected)

Backward elimination is a sensible approach

- Significance level can be chosen depending on the modelling aim
  - (eg. P=0.01 for diagnostic criteria, P=0.20 for confounders)
- Reduces overfitting
- We assume that re-inclusion is allowed

Of course required:

- Checks
- Sensitivity analysis
- Stability analysis

Results of BE and Stepwise (Forward with re-exclusion) are very often identical in low correlation and large sample sizes

3. Examples - different procedures select different models

- ‘Bad’ data
  - Highly correlated data
  - Sample size too small

  → Results differ often

- More appropriate situation

- Problems caused by variable selection

Different selection procedures, different results

SHOCK 1

Risk factors for CHD
N=7088, 456 events (McGee et al. 1984)

Selection method	Factor			
	Calories	Protein	Fat	Carbohydrates
Full model	XX	X	X	
BE	XXX		XX	
SS				XXX
β/SE (Wald)	3.05	2.14	2.43	1.13

X – 5%; XX – 1%; XXX – 0.1%

Extreme situation, strong correlation! Selection sensible?

Just estimate the parameters in the model with 4 variables

Risk factors for CHD - Correlation

The 4 factors are highly correlated

- Calories with (Protein, Fat, Carbohydrates): 0.75-0.77,
- Protein – Fat: 0.72,
- Protein – Carbohydrates: 0.41,
- Fat – Carbohydrates: 0.29

Another SHOCK 2

Prognostic factors for multiple myeloma, N = 65, 48events, Kuk (1984)

	Full model (5%)	AII - AIC	BE (0.05)	StS (0.05)
1	x	x	X	X
2				X
3	x	x	X	
4		x	X	
5		x		
6	x	x	X	
7		x	X	
8				
9				
10	x	x		
11	x	x		
12			X	
13			X	
14				
15				
16				

Effective sample size is very small (48)

More appropriate situation

Prognostic factors for brain tumor (glioma, N=411, 274 deaths)
15 variables, multicollinearity

Compare models selected with BE and StS
Consider different significance levels (0.01, 0.05, 0.10, 0.157)
Compare AIC with BE (0.157)

All models include X_3, X_5, X_6, X_8 (call it M_B)
(in the full model these 4 variables have $p < 0.05$,
no other variable with $p < 0.05$).

Glioma study – models selected

MB- ´basis model including x3,x5,x6,x8

Procedure	Sign. level	Model selected
BE	0.01	M_B
StS	0.01	M_B
BE	0.05	$M_B + X_{12}$
StS	0.05	$M_B + X_{12}$
BE	0.10	$M_B + X_{12} + X_4 + X_{11} + X_{14}$
StS	0.10	$M_B + X_{12} + X_1$
BE	0.157	$M_B + X_{12} + X_4 + X_{11} + X_{14} + X_9$
StS	0.157	$M_B + X_{12} + X_4 + X_{11} + X_{14} + X_9$
AIC	–	$M_B + X_{12} + X_4 + X_{11} + X_9 + X_{13}$

Glioma study – Estimation after selection

411 patients (274 events) with complete data

Var	$\hat{\beta}$	
	full	BE(0.05)
X_1	-0.09	
X_2	-0.06	
X_3	0.31	0.38
X_4	0.12	
X_5	0.45	0.43
X_6	-0.14	-0.16
X_7	-0.02	
X_8	-0.31	-0.33
X_9	-0.10	
X_{10}	0.04	
X_{11}	0.12	
X_{12}	-0.13	-0.14
X_{13}	0.03	
X_{14}	0.11	
X_{15}	-0.07	

Problems caused by variable selection

- Biased estimation of individual regression parameter
- Overoptimism of a score (here not considered)
- Under- and Overfitting
- Replication stability (see session 6.2)

Severity of problems influenced by complexity of models selected

Specific aim influences complexity

4. Bias introduced by variable selection

- Selection bias
  - Effect of beta, sample size
- Omission bias
  - Correlated variables

Selection Bias

Selection and estimation from one data set

- Copas & Long (1991)
  - Choice of variables depends on estimated coefficients rather than their true values. X is more likely to be included if the regression coefficient is overestimated.

- Miller (1990)
  - Competition Bias: Best subset for fixed number of parameters

  - Stopping Rule Bias: Criterion for number of parameters
    ...the more extensive the search for the choosen model the greater the selection bias

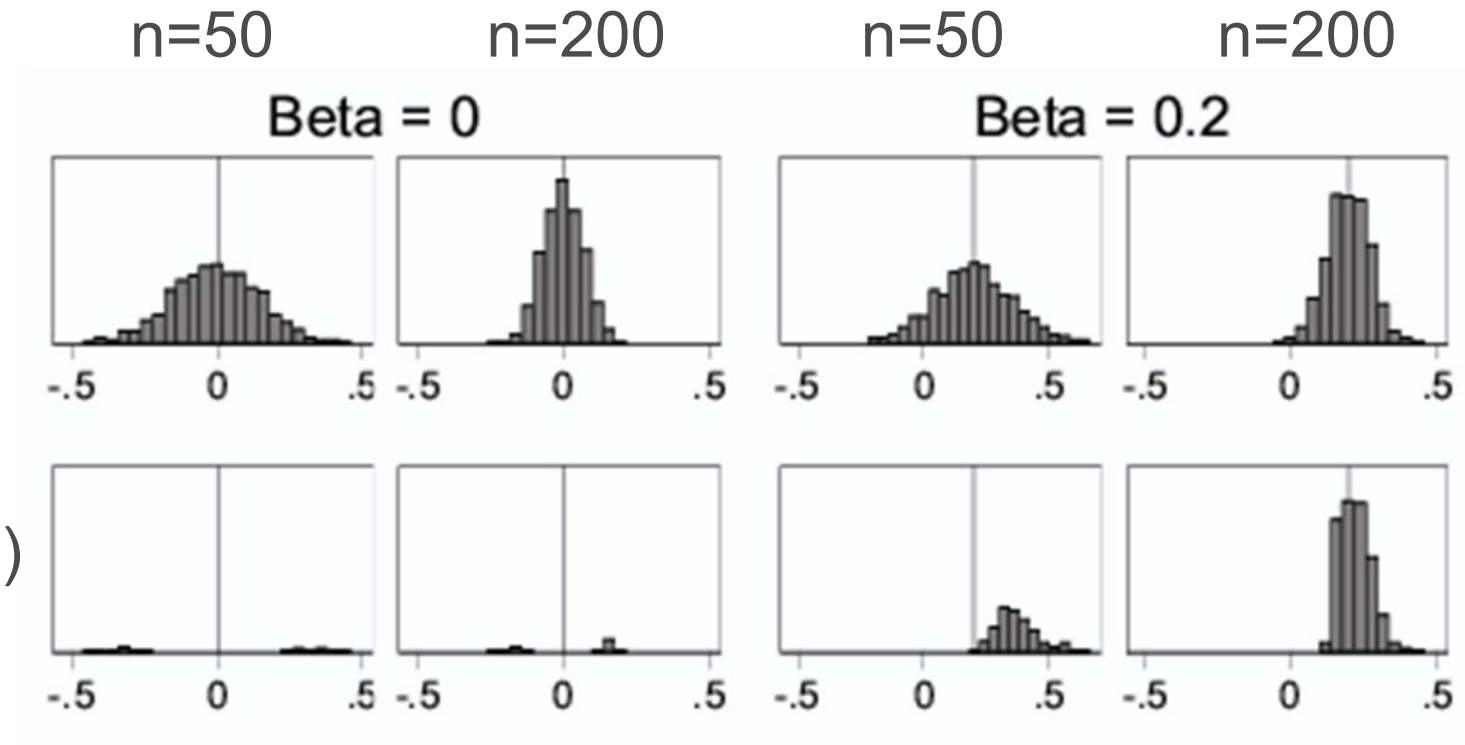
Selection Bias

- Key point: standardized β (β/SE) small or large
  (absolute ß, sign ignored in the following)

- Depends on true effect of β and sample size

Selection Bias:

a problem of small standardized ß

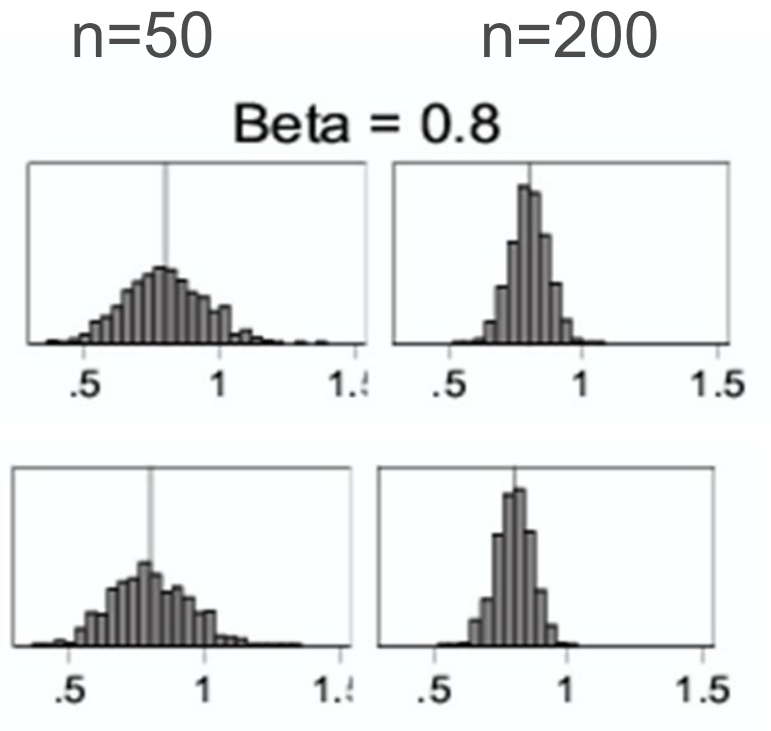


	n=50	n=200	n=50	n=200
% incl.	4.2	4.0	28.5	80.4

R&S 2008

Selection Bias:

a problem of small standardized ß (cont.)



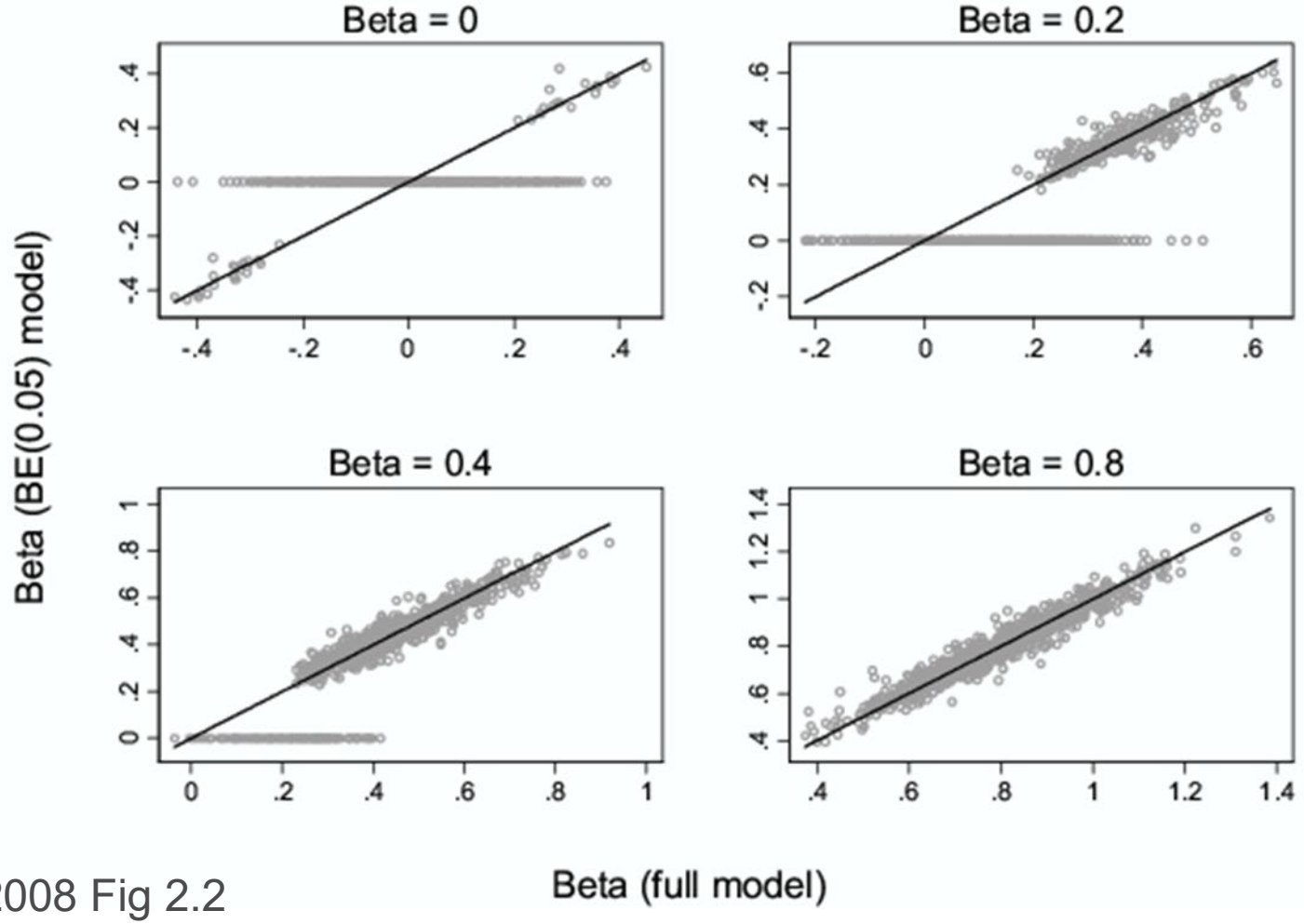
	n=50	n=200
% incl.	100	100

R&S 2008

Selection Bias:

Estimate after variable selection with BE (0.05)

Simulation 5 predictors, 4 are ‚noise'



R&S 2008 Fig 2.2

Omission Bias:

Reasons for the bias

- True model Y = X_1 β_1 + X_2 β_2 + ε
  Decision for model with subset X1

  X_1 and X_2 may include several variables
  β_1 and β_2 may be vectors

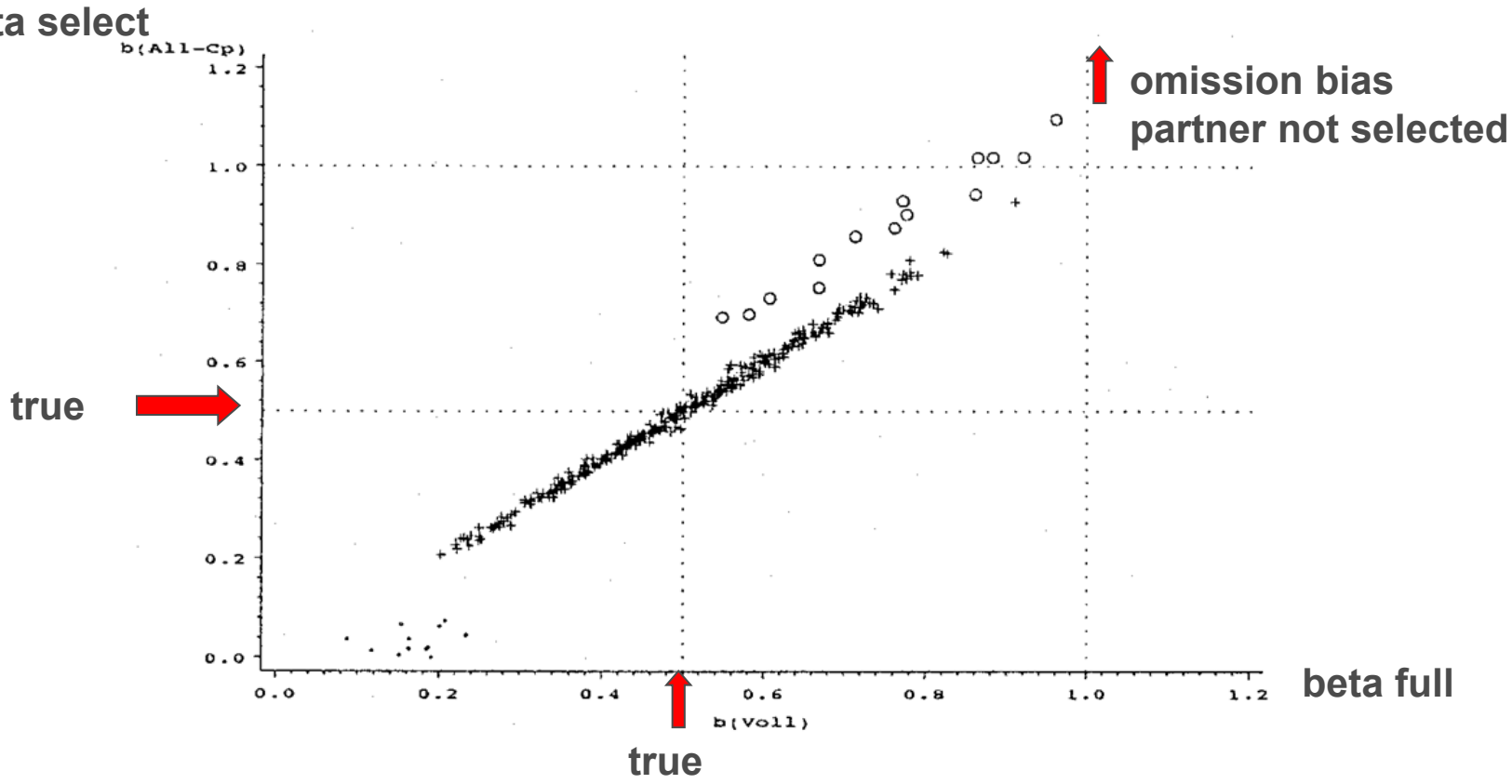
- Estimation with new data

$$E(\beta_1) = \beta_1 + (X_1' X_1)^{-1} X_1 X_2 \beta_2$$

$$|\text{.....Omission bias}|$$

Selection and omission bias

Estimates of β_2 from full and selected model, corr. variables x1 and x2 , power is high



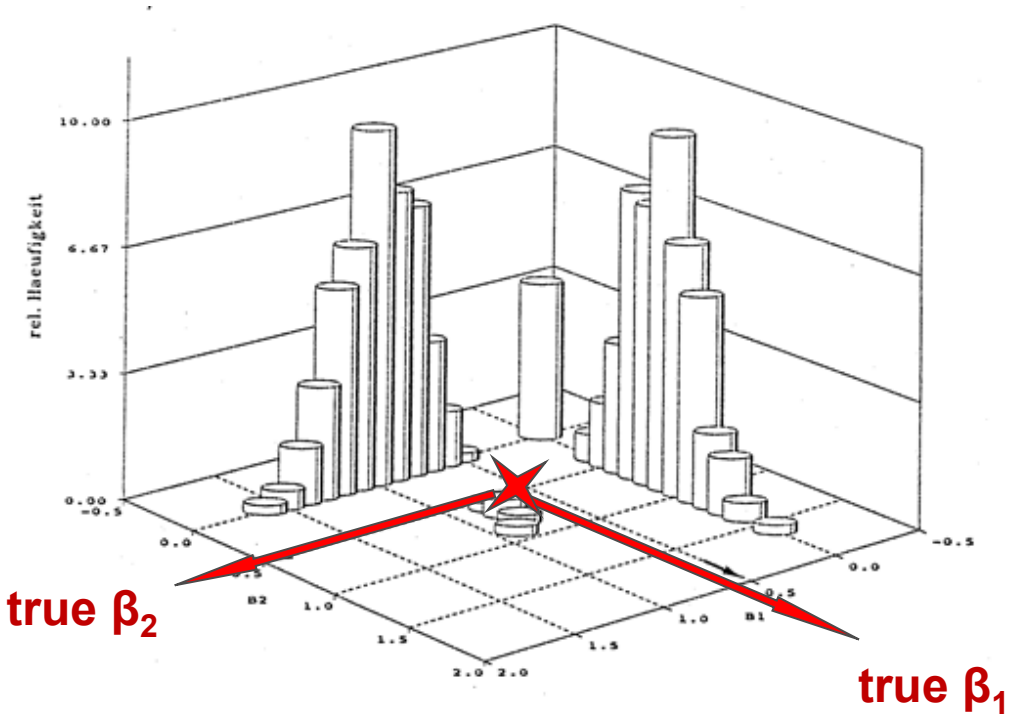
+ X_1 and X_2 in C_p model selected ●X_2 in C_p model not selected
○X_1 in C_p model not selected

Sauerbrei (1992)

Selection and omission bias

Two correlated variables with weak effects, small sample size

Often one ‚representative' is selected
Estimates of β_1 and β_2 from selected model



Sauerbrei (1992)

Summary (1)

Model building in observational studies, low dimensional data

(many issues are easier in randomized trials)

- All models are wrong, some, though are better than others and we can search for the better ones. Another principle is not to fall in love with one model, to the exclusion of alternatives (Mc Cullagh & Nelder 1983)

Summary (2)

- More than 10 strategies for variable selection
- Nominal significance level is the key factor
- Usual estimates after selection may be (heavily) biased, especially for small studies
- Specific aim of a study has
  - influence on selection strategy
  - influence on importance of the problems
  - replication stability
  - under- and overfitting
  - biased estimation of regression parameters
  - overoptimism
- Personal preference against over complex models
- Importance of other aspects as categorization, functional relationship often underrated (comes later)

Discussion and Outlook

- Properties of selection procedures need further study
- More prominent role for complexity and stability in analyses required
  - resampling methods well suited
- Combination of selection and shrinkage
- Model uncertainty concept
- Triggered by high-dimensional data

  Many new proposals, but what about properties and comparisons?

5. Selection bias caused by variable selection - can we correct by shrinkage?

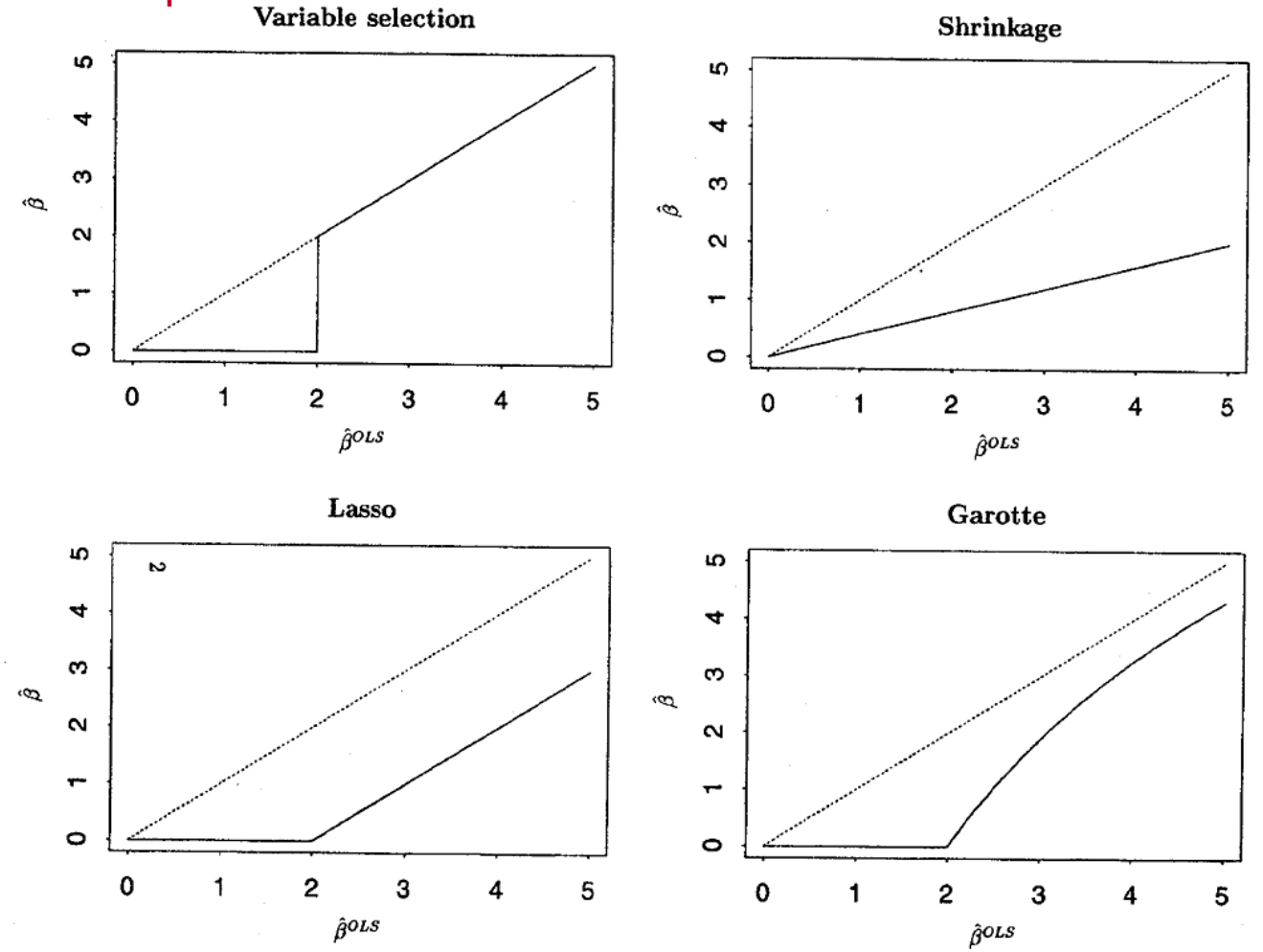
- Variable selection and shrinkage
- Global and parameterwise shrinkage factors

- Shrinkage to correct for selection bias?

Variable Selection and Shrinkage

Regression coefficients as functions of OLS estimates

Principle for one variable



Regr. coefficients zero ⇒ variable selection R&S 2008 Fig. 2.4

Variable Selection and Shrinkage

OLS	$\operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$
Var Sel	$\operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$ under the constraint $\beta_j = 0$ for $j \notin I$
Shrinkage by CV calibration	
-global	$\operatorname{argmin}_{c}\left\{\sum_{i=1}^{n}\left(y_i - c\cdot\sum_{j=1}^{p}\hat{\beta}^{OLS}_{j(-i)}x_{ij}\right)^2\right\}$ calculated for a model selected
- PWSF	$\operatorname{argmin}_{c_j}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j\in I}^{p}c_j\cdot\hat{\beta}^{OLS,I}_{j(-i)}x_{ij}\right)^2\right\}$ calculated for a model selected
Garotte	$\operatorname{argmin}_{c}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}c_j\hat{\beta}^{OLS}_{j}x_{ij}\right)^2\right\}$ under the constraint $\sum_{j=1}^{p}c_j \le t$ with $c_j \ge 0$ optimal t determined by cross - validation
Lasso	$\operatorname{argmin}_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}$ under the constraint $\sum_{j=1}^{p}\lvert\beta_j\rvert \le t$ optimal t determined by cross - validation

Combine variable selection and shrinkage

Ridge regression – Shrinkage, but no selection

Within estimation shrinkage

- Garotte, Lasso and newer variants

Combine variable selection and shrinkage, optimization under different constraints

Post estimation shrinkage using CV (shrinkage of a selected model)

- Global
- Parameterwise (PWSF, heuristic extension)

Penalized Regression Methods for Linear Models

- This approach introduces a penalty term to the loss function, $\ell(\theta|y,x)$
- The parameters are obtained by minimizing $K(\theta)$

$$K(\theta) = \ell(\theta|y,x) + \lambda P(\theta)$$

- In normal error regression models we minimize the residual sum of squares (RSS) plus a penalty term as shown below

$$K(\theta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\hat{\beta}_j X_j) + \lambda P(\theta) = RSS + \lambda P(\theta)$$

  - Where
    - $\boldsymbol{P}$ is a penalty function such as $\sum_{j=1}^{p}{\beta_j}^2$ for ridge regression
    - $\boldsymbol{\lambda}$ is the tuning parameter often obtained using k-fold cross-validation
- Most penalized approaches such as Nonnegative Garrote and Lasso conduct variable selection and shrinkage simultaneously

Least absolute shrinkage and selection operator (Lasso) and adaptive Lasso

- Standardization of predictors X to a mean of 0 and variance 1 is required
- Lasso does not need the initial estimates from the full model
- The adaptive Lasso (ALASSO) needs initial estimates from the full model in order to calculate its adaptive weights denoted by $\widehat{w}_j = \frac{1}{\left|\widehat{\beta}_j^{\,init}\right|^{\gamma}}$
- Lasso and ALASSO minimize objective function $K(\beta_L)$ and $K(\beta_A)$, respectively

$$K(\beta_L) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_j\right)^2 + \lambda\sum_{i=1}^{p}|\beta_j| \qquad \text{LASSO}$$

$$K(\beta_A) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_j\right)^2 + \lambda\sum_{i=1}^{p} w_j|\beta_j|\,, \qquad \widehat{w}_j = \frac{1}{\left|\widehat{\beta}_j^{\,init}\right|^{\gamma}} \qquad \text{ALASSO}$$

- $\boldsymbol{\lambda}$ is the key tuning parameter for Lasso
- $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are the key tuning parameters for adaptive Lasso
- $\widehat{\beta}^{init}$ is usually **OLS**, but **ridge** or any other initial estimates can be used

Nonnegative Garrote (NNG)

- The NNG starts with the initial estimates $\hat{\beta}_j$ from the full model. In the linear regression model, OLS estimates are often used
- Unlike Lasso, standardization of variables is not necessary
- The shrinkage factor, c_j, is obtained by minimizing the objective function $K(c_j)$. The **intercept**, $\hat{\beta}_0$, is not penalized

$$K(c_j) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} c_j \hat{\beta}_j X_j \right)^2 + \lambda \sum_{i=1}^{p} c_j \,, \qquad c_j \geq 0 \,, \lambda \geq 0$$

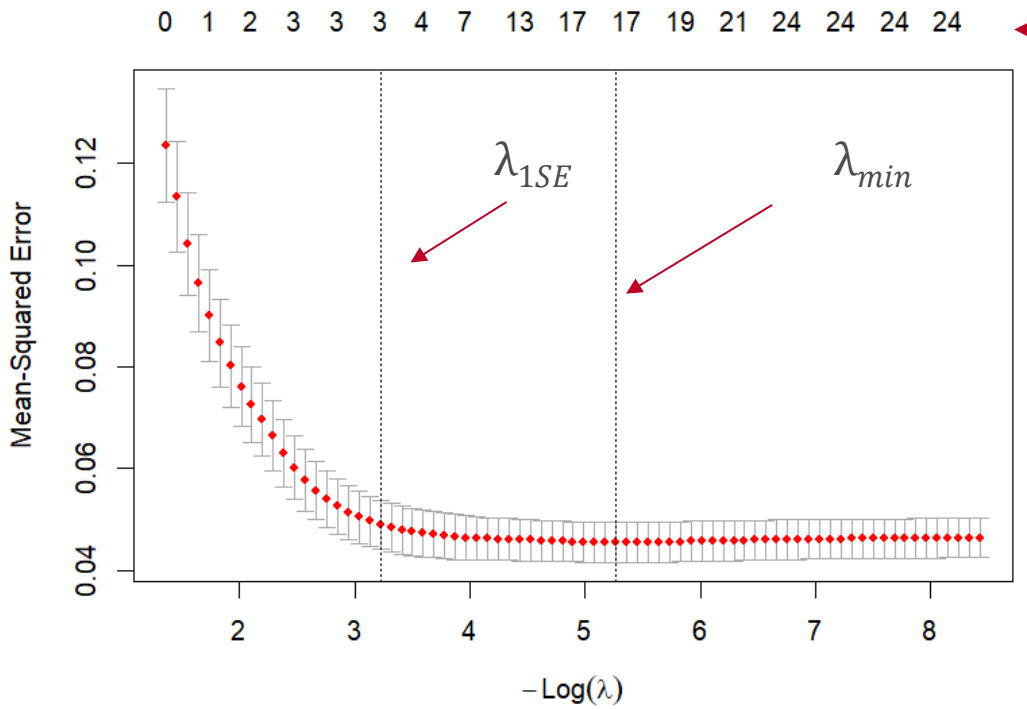
  - Where $\boldsymbol{\lambda}$ is the key tuning parameter that controls model complexity.
- The nonnegative garrote estimate is given by ${\beta_j}^G = c_j \hat{\beta}_j$
- $c_j = 1$ implies that no shrinkage is needed for variable x_j while $c_j = 0$ implies exclusion of variable x_j from the model
- Note that the c_j are not restricted to the range of 0 and 1. It can exceed 1 for correlated variables (consequence of omission bias)

Choice of Tuning Parameters

- The K-fold cross-validation (K = 10) method is often recommended for choosing the tuning parameter(s) for Lasso, Garrote and Adaptive Lasso
- In Lasso, the value of λ that minimizes the mean cross-validation error is chosen as the tuning parameter
- Adaptive Lasso has two parameters (λ, γ). Two-dimensional 10-fold CV can be used to obtain the optimal pairs.
- The pair that minimizes the mean cross-validation error is chosen as the tuning parameters.

Determining Tuning Parameters

- Choose the grid values of λ, often the default values of the software is used
- In Lasso, if **prediction is the aim:** choose λ_{min} , the value that minimizes the cross-validated mean squared error (MSE).
- Or if **variable selection is the aim:** choose λ_{1SE} the largest value of λ whose cross-validated MSE is within one standard error of the minimum.




```
Call:  cv.glmnet(x = x, y = y)

Measure: Mean-Squared Error

     Lambda Index Measure       SE Nonzero
min 0.00515    43 0.04549 0.003951      17
1se 0.03986    21 0.04898 0.004897       3
```

6. Complexity of predictors

- Models with a small or a large number of predictors
- Models with simple or ‘complex’ functions (next sessions)

Complexity of models

- Main (clinical) aim of the model has strong influence on choice of complexity
- Variable selection strategies:
  AIC, BIC or stepwise strategies select on different nominal significance levels
- Complexity has influence on problem of overfitting/ underfitting

Main aim prediction

- Predictors are ‚dominated' by some strong factors

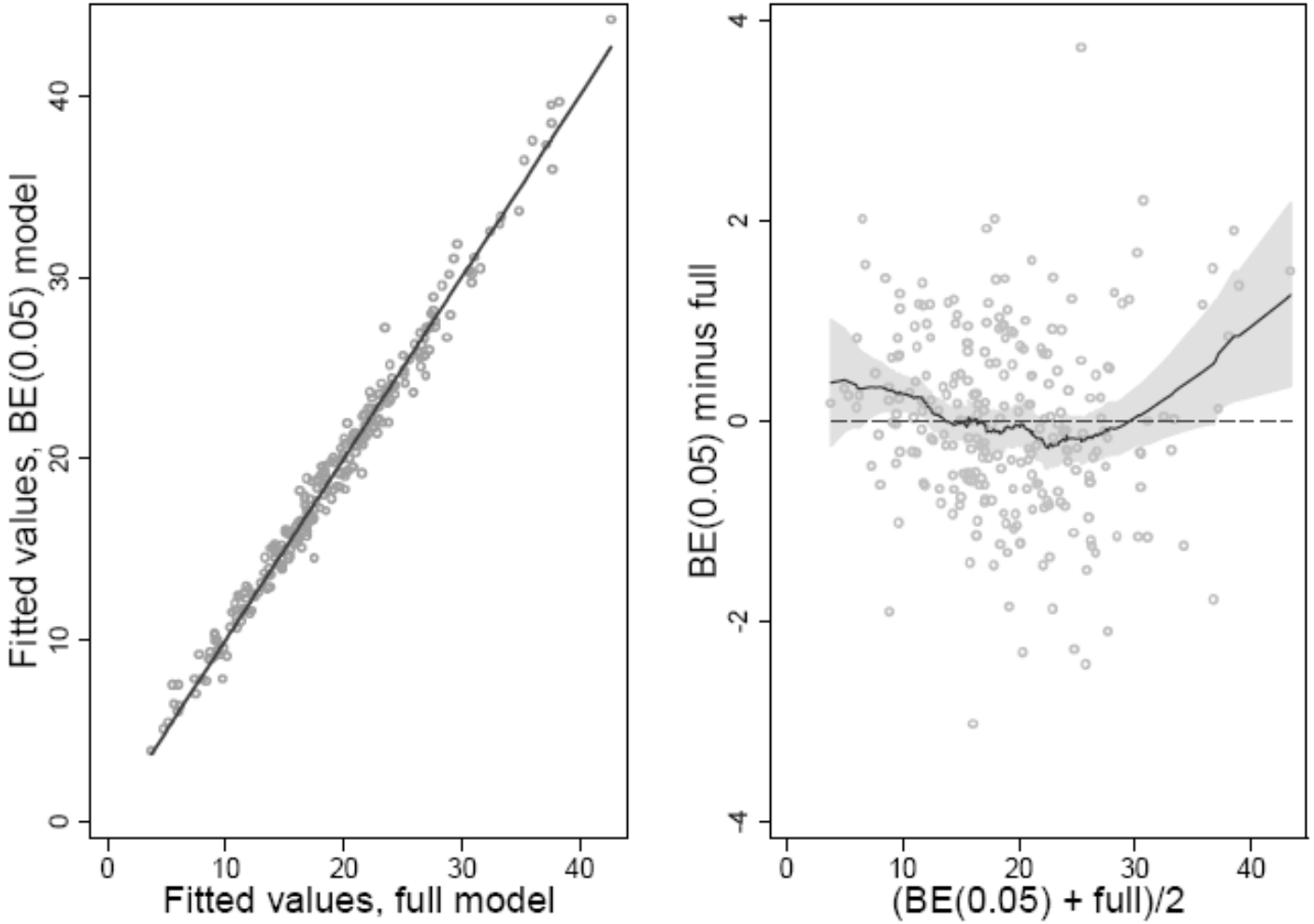
Simple or complex predictor?

Often very close agreement
Glioma study (Full – 15 variables, BE – 4 variables)

Bland- Altman plot

Body- fat data

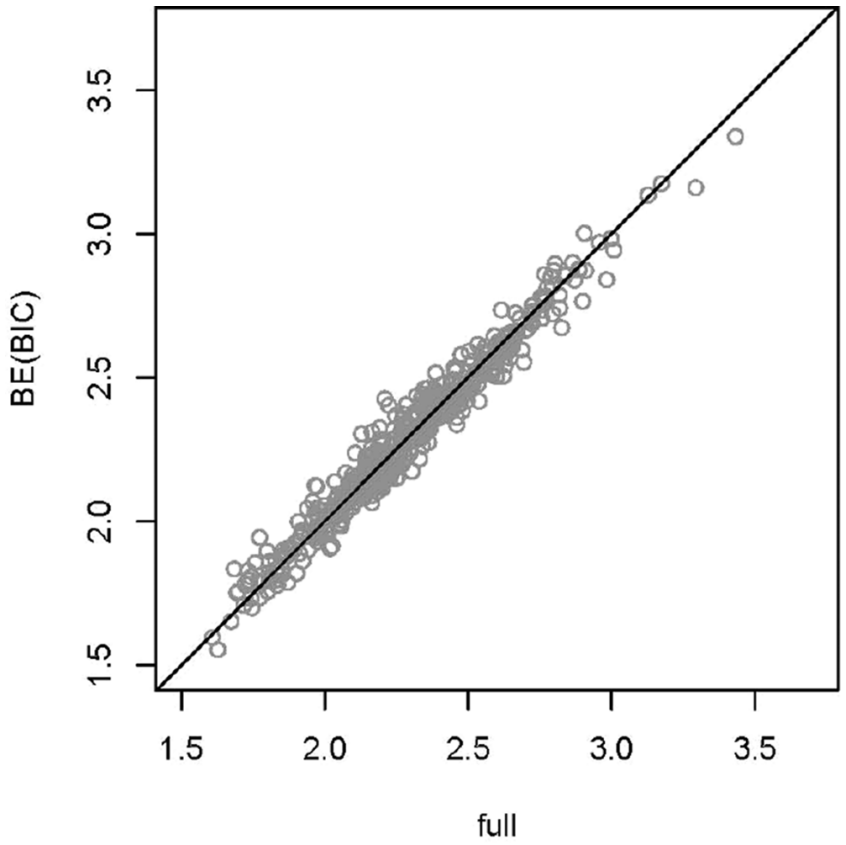
Full-13 variables, BE- 4 variables



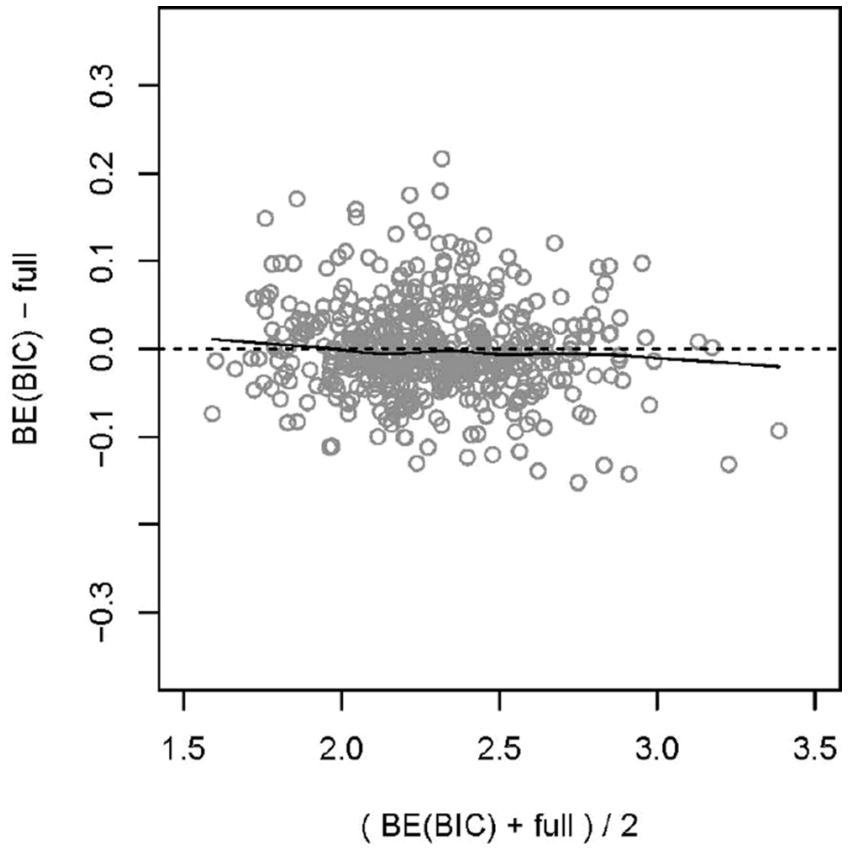
R&S 2008, Fig 2.1

Similarity of predictors

Ozon data - outcome



Full model – 24 variables
BIC model – 4 variables

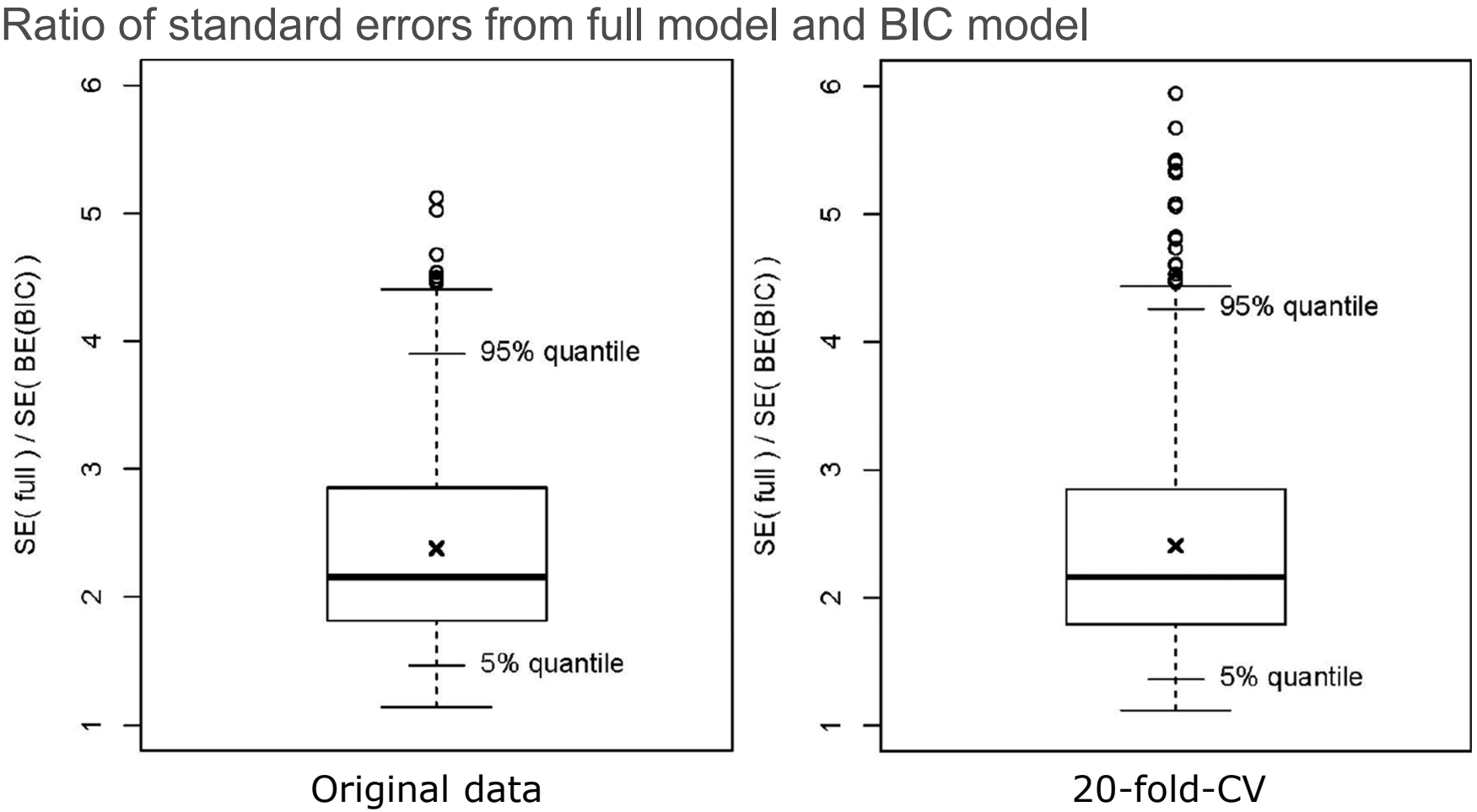


Bland – Altman plot

Sauerbrei et al. 2015, Fig 1

Similarity of predictors

Ozon data - Variances are very different



Sauerbrei et al. 2015, Fig 2

Simple or complex predictor?

- Often a simple and a complex predictor (eg derived by using significance level 0.01 or 0.157) are highly correlated.
- Main reason: usually both include the ‚strong' factors which dominate the predictor.
- In linear error models: many models have similar R^2.
- Glioma example: 4 models (full, 0.157, 0.05,0.01)
- Pearson correlation coefficient between full model and simplest model with 4 variables is 0.94, values of other pairs of models are higher.

If also modelling non-linear functions, the issue is more complicated, but the principle is similar.

7. Handling categorical predictors

5 types of covariates

Type	Meaning	Example
Binary	Two classes	Gender (M/F)
Nominal	Categorical with no natural ordering	Continent (Europe, Asia, etc.)
Ordinal	Categorical with ordering	Tumour grade in cancer
Count	Number of instances of something	Number of children
Continuous	Can in principle take any real value	Serum cholesterol concentration

R&S 2008, Tab 3.1

Coding schemes for an ordinal variable

dose	Metric coding		Dummy-variable coding				Dead/Total	(%)
			Categorical		Ordinal			
	M-1	M-2	DC1	DC2	DO1	DO2		
1 - low	1	1	0	0	0	0	40/100	(40)
2 - medium	2	2	1	0	1	0	70/130	(54)
3 - high	3	2.5	0	1	1	1	90/170	(53)
Total							200/400	(50)

R&S 2008, Tab 3.2

Odds ratio

Dependence on coding scheme and selection strategy

dose	Crude OR	Model-based OR					
		Metric coding		Dummy-variable coding			
				Categorical		Ordinal	
		M-1*	M-2	FS	BE	FS	BE
2 vs. 1	1.75	1.0 (1.26)	1.43	1.0	1.75	1.71	1.71
3 vs. 2	0.96	1.0 (1.26)	1.19	1.0	0.96	1.0	1.0

R&S 2008, Tab 3.3

Cervical cancer data

Variable	Explanation	Coding					
		0	1	2	3	4	5
caco	Case-Control status	control	case				
age	Age group		15-30	31-35	36-40	41-45	46-60
ocuse	Previous oral contraceptive use	no	yes				
duroc	Months of oral contraceptive use	0	1-12	13-24	25-60	≥ 61	
sexrel	Number of sexual relations	0	1	≥ 2			
agesex	Age at first sexual relationship	none	8-16	17-18	19-20	21-23	≥ 24
npaps	No. of Papanicolaou smears	none	1-2	3-4	5-6	≥ 7	
vagdis	Doctor visited for abnormal vaginal discharge	no	yes				
tpreg	Total number of pregnancies	none	1-2	3-4	5-6	≥ 7	

R&S 2008, Tab 3.4

Coding schemes for ordinal variables

duroc	Metric	Categorical coding				Ordinal coding			
	M-1	DC1	DC2	DC3	DC4	DO1	DO2	DO3	DO4
never	1	0	0	0	0	0	0	0	0
1-12	2	1	0	0	0	1	0	0	0
13-24	3	0	1	0	0	1	1	0	0
25-60	4	0	0	1	0	1	1	1	0
>60	5	0	0	0	1	1	1	1	1

R&S 2008, Tab 3.5

Influence of duroc
- Comparison to reference category

duroc	Cases/n	(%)	OR	P
never	88/629	(14)	–	
1-12	15/89	(17)	1.25	0.47
13-24	3/45	(7)	0.44	0.17
25-60	14/75	(19)	1.41	0.28
> 60	21/61	(34)	3.23	< 0.001

R&S 2008, Tab 3.6

Influence of duroc

logOR result depending on coding scheme

duroc	Categorical	Ordinal	Metric
never	0	0	0
1-12	0.22	0.22	0.21
13-24	−0.82	−1.04	0.42
25-60	0.34	1.17	0.63
> 60	1.17	0.83	0.83

R&S 2008, Tab 3.7

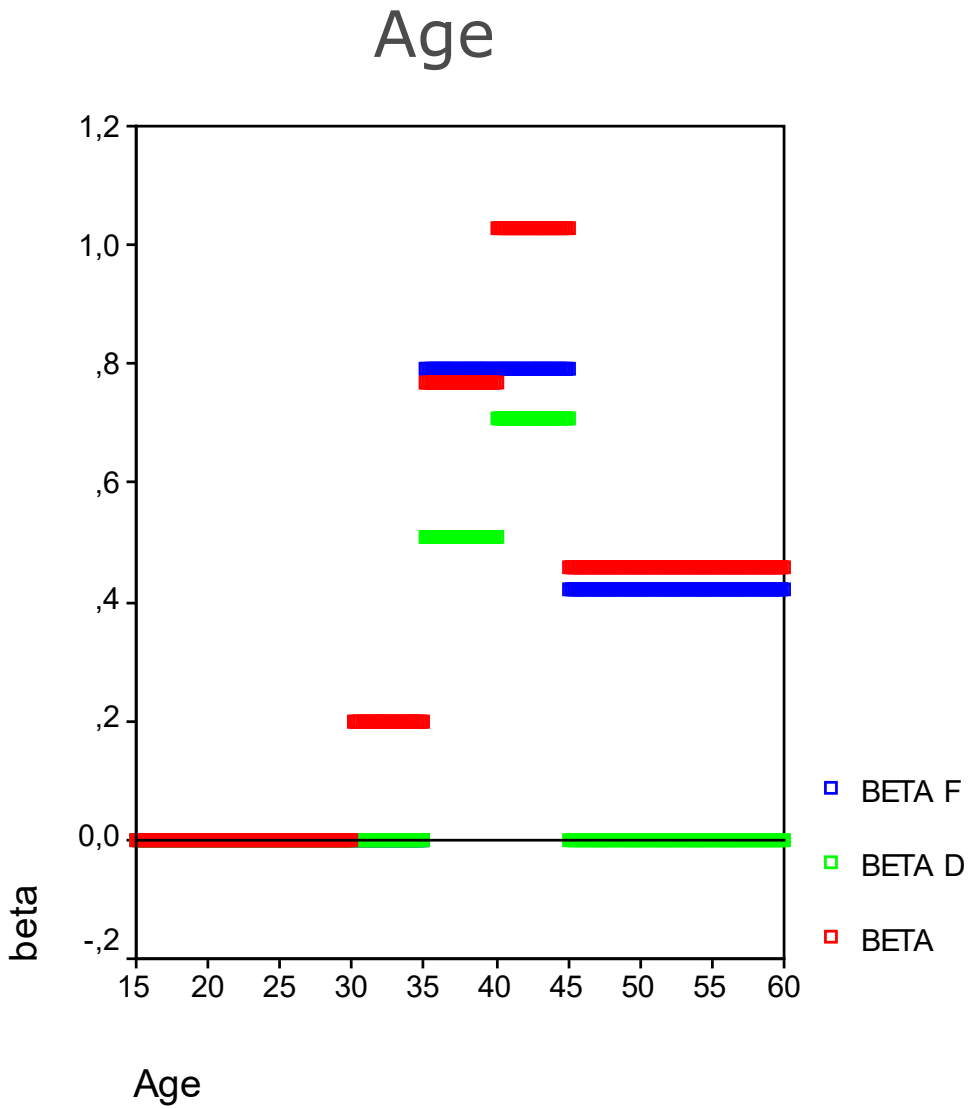
Variable	Full model		BE(0.05)	Duration*	
	$\widehat{\beta}$	P	$\widehat{\beta}$	$\widehat{\beta}$	SE
`age`	0.12	0.2	–	–	–
`duroc`	0.15	0.04	–	0.13	0.07
`sexrel`	0.38	0.2	–	–	–
`agesex`	0.13	0.04	0.15	0.14	0.06
`npaps`	−0.48	< 0.01	−0.43	−0.47	0.14
`vagdis`	0.91	< 0.01	0.90	0.89	0.21
`tpreg`	0.42	< 0.01	0.52	0.51	0.08

R&S 2008, Tab 3.9

Variable	Category	β	p	D(0.05)	O(0.05)	F
age	15-30	0	—	0	0	0
	31-35	0.20	0.68	0	0	0
	36-40	0.77	0.06	0.51	0.68	0.79
	41-45	1.03	0.02	0.71	0.68	0.79
	46-60	0.46	0.31	0	0.68	0.42
OC use	never	0	—	0	0	0
durability	1-12	-0.03 (0.34)	0.93	0	0	0
	13-24	-0.93 (0.63)	0.14	0	0	0
	25-60	0.11 (0.35)	0.75	0	0	0
	61 or more	0.83 (0.33)	0.01	0.82	0.86	0.82
no. sexrelations	none, one	0	—	0	0	0
	more than one	0.25	0.42	0	0	0
age first sex	never	0	—	0	0	0
	8-16	2.41	0.04	2.74	2.96	2.95
	17-18	2.75	0.02	3.10	2.96	2.95
	19-20	3.06	0.01	3.41	2.96	2.95
	21-23	2.75	0.02	3.04	2.96	2.95
	24 or older	2.39	0.04	2.73	2.96	2.95
no. pap smears	none	0	—	0	0	0
	1-2	-0.38	0.17	0	0	0
	3-4	-1.16	‹.00	-1.07	-1.00	-1.04
vag.discharge	no	0	—	0	0	0
	yes	0.76	0.00	0.71	0.73	0.70
no. pregnancies	none	0	—	0	0	0
	1-2	0.39	0.50	0	0	0
	3-4	0.95	0.10	0.69	0.67	0.72
	5-6	0.84	0.17	0.70	0.67	0.72
	7 or more	1.05	0.10	0.91	0.67	0.72

R&S 2008, Tab 3.8

Results from 3 models



Cervical cancer – coding ‚linear‘

Var	Full model M L1 $\hat{\beta}$	p	BE(0.05) M L2 $\hat{\beta}$	Duration, adj. BE(0.157) M L3 $\hat{\beta}$	SE
age	0.12	0.20	—	—	—
duration	0.15	0.04	—	0.13	0.07
no. sex rel.	0.38	0.21	—	—	—
age first sex	0.13	0.04	0.15	0.14	0.06
no. pap smears	-0.48	<0.01	-0.43	-0.47	0.14
vag. discharge	0.91	<0.01	0.90	0.89	0.21
no. pregnancies	0.42	<0.01	0.52	0.51	0.08

Interpretation from model M L3

Odds ratio duration (13-24) vs never: exp (2· 0.13) = 1.30
duration (> 60) vs never: exp (4· 0.13) = 1.68

R&S 2008, Tab 3.9

OC use and cervical cancer

Comparison of models

Modell	no. variates	-2 logL	AUC
Full	21	645.3	0.783
BE – Ord			
0.05	6	662.1	0.756
0.157	11	652.0	0.772
BE – Dum			
0.05	13	653.2	0.774
Linear Coding	7	687.2	0.750
Lin.cod. and BE	4	693.7	0.739
Final	7	659.7	0.762

Example: OC use as risk factor for cervical carcinoma

Summary of multivariable analysis presented
(no interactions considered)

- Data categorized – problematic
- Coding is very important
  - Coding linear – very bad
  - Coding Dummy – often preferred for presentation
    problematic with stepwise selection
  - Coding Ordinal – preferable for stepwise selection
- Full modell or selected model
  - Interpretation of full model can be difficult
  - BE can give sensible model if Ordinal-Coding is used
  - selection level important

8. Summary

Taken from **Royston and Sauerbrei 2008** book

- Selection of variables
- Handling categorical and continuous predictors
- Further issues

Selection of variables – Summary (Chap 2)

1. Because subject-matter knowledge in observational studies is usually limited, data-driven model selection has an important role.
2. Multivariable modelling has different possible goals; the main distinction is between predictive and explanatory models. The former aims for a good predictor, whereas the latter aims to identify important variables.
3. Despite claimed theoretical advantages, the full model is not a practical proposition in most studies.
4. Interpretability and practical usefulness are essential attributes of explanatory models. Simple models are more likely to have such properties.

5. Several model selection algorithms have been suggested. The most used are stepwise methods, the best of which is backward elimination. Methods based on information criteria (AIC and BIC) are the main competitors.
6. With stepwise methods, the key tuning parameter is the nominal P -value for selecting or eliminating a variable. Larger P -values produce larger models.
7. Replication instability of selected models should be assessed by bootstrap.
8. Parameter estimates of selected models, irrespective of the strategy used, are prone to different types of bias. With small samples the bias can be large.
9. Techniques combining selection of variables with shrinkage of their parameter estimates reduce bias, but their properties require further exploration.

Handling categorical and continuous predictors – Summary (Chap 3)

1. Observational studies comprise a mixture of binary, ordinal, nominal and continuous covariates. Preliminary to the analysis one must decide whether data manipulations are required.

2. For categorical variables, sparse categories may be combined, guided by the distribution of the variable and sometimes by subject-matter knowledge. Sparse categories may also indicate a small group of subjects who should have been excluded.

3. Reference categories should not be too small and should make scientific sense as a comparator for the other categories.

4. Having chosen the categorization, a satisfactory coding scheme is needed when variable selection is to be done.

5. Analysis of continuous variables using cutpoints should be avoided.

6. Investigation of non-linearity is required.

7. Interpretability and transportability of functions for continuous predictors are important. Generally, simple functions are preferable to complex ones.

8. Global-influence functions, which include polynomials and FPs, are generally simpler than functions with local features.

Further issues

- Interpretability and Stability should be important features of a model.
- Validation (internal and external) needs more consideration
- Resampling methods

  give important insight, but theoretically not well developed should become integrated part of analysis lead to more careful interpretation of results

- Transportability and practical usefulness are important criteria (Prognostic models: clinically useful or quickly forgotten? Wyatt & Altman 1995)

Be carefull with too complex models

STRATOS initiative – TG2

A more recent overview

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and Prognostic Research

COMMENTARY Open Access

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei[1*], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], Georg Heinze[6] and for TG2 of the STRATOS initiative

Towards state of the art – research required!

Table 1 Relevant issues in deriving evidence-supported state of the art guidance for multivariable modelling

No.	Item
1	Investigation and comparison of the properties of variable selection strategies
2	Comparison of spline procedures in both univariable and multivariable contexts
3	How to model one or more variables with a 'spike-at-zero'?
4	Comparison of multivariable procedures for model and function selection
5	Role of shrinkage to correct for bias introduced by data-dependent modelling
6	Evaluation of new approaches for post-selection inference
7	Adaption of procedures for very large sample sizes needed?