TG2: Selection of Variabes and Functional Form for Multivariable Models

Georg Heinze

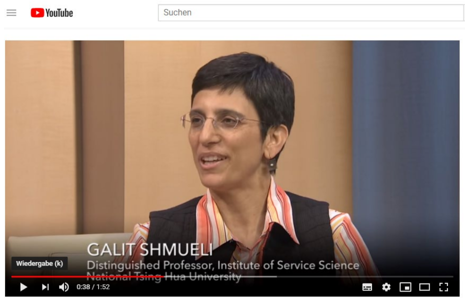
Regression modeling – what‘s important?

- **Which variables to include?**
  Driven by the interpretation of the model as:
  - Prediction
  - Causal explanation
  - Description
- **How to specify the functional form of continuous variables?**

- We assume ‚homework‘ has been done:
  - Initial data analysis (TG3)
    - Missing values, univariate X, multivariate X, keep outcome Y separate!
  - Reasonable model class was chosen
    - Linear, binary, multinomial, censoring, …

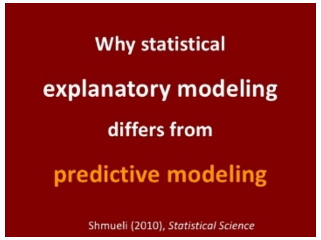
Why statistical modeling?

- To explain or to predict?


YouTube
Suchen
GALIT SHMUELI
Distinguished Professor, Institute of Service Science
National Tsing Hua University
Wiedergabe (k)
0:38 / 1:52

Galit Shmueli discusses the distinction between explaining and predicting (Preview)

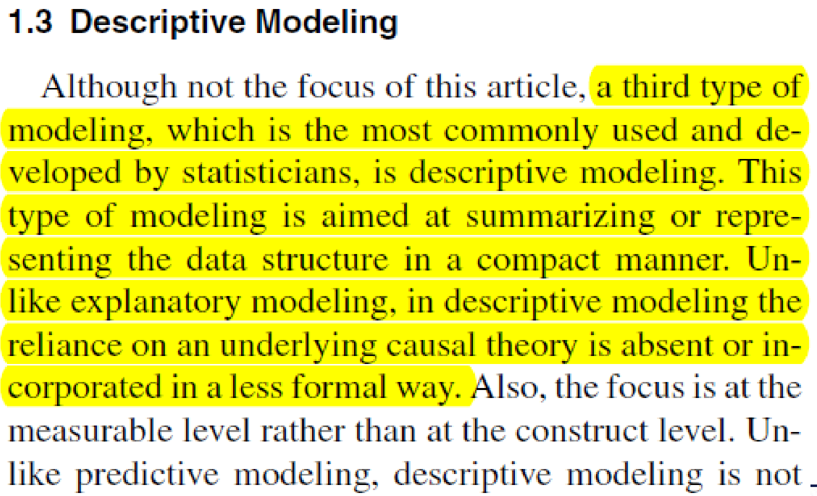
Galit Shmueli, 2010, *Statistical Science*



Why statistical explanatory modeling differs from predictive modeling

Shmueli (2010), *Statistical Science*

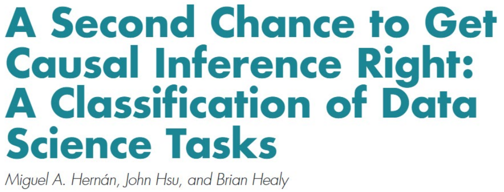
- … or to describe?



1.3 Descriptive Modeling

Although not the focus of this article, a third type of modeling, which is the most commonly used and developed by statisticians, is descriptive modeling. This type of modeling is aimed at summarizing or representing the data structure in a compact manner. Unlike explanatory modeling, in descriptive modeling the reliance on an underlying causal theory is absent or incorporated in a less formal way. Also, the focus is at the measurable level rather than at the construct level. Unlike predictive modeling, descriptive modeling is not

Galit Shmueli, 2010, *Statistical Science*



A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks

Miguel A. Hernán, John Hsu, and Brian Healy

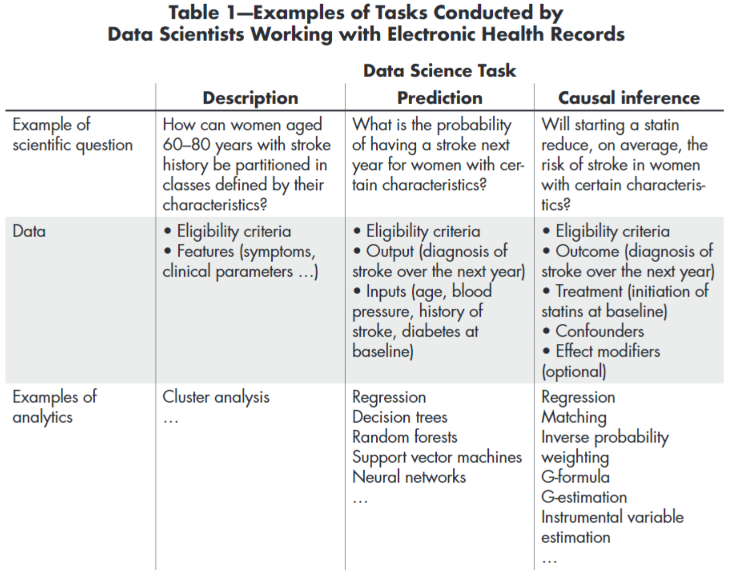


Table 1—Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	• Eligibility criteria • Features (symptoms, clinical parameters …)	• Eligibility criteria • Output (diagnosis of stroke over the next year) • Inputs (age, blood pressure, history of stroke, diabetes at baseline)	• Eligibility criteria • Outcome (diagnosis of stroke over the next year) • Treatment (initiation of statins at baseline) • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

To Explain or to Predict?

- **Descriptive models**
  - Capture the data structure parsimoniously: which variables are associated wtih the outcome and how?
  - Often useful transparent prediction models, in special cases even causal conclusions possible
- **Prediction models**
  - Interest in accurate predictions for future application.
  - No concern about causality and confounding (association), but explainability.
  - Prognostic and diagnostic prediction models.
- **Explanatory models**
  - Interest in causal contrasts (e.g., coefficients)
  - Often achieved by counterfactual prediction
  - Confounder selection

(Shmueli, 2010)

More about models‘ aims

- **Description**:
  - Just X and Y: understand how Y is associated with X‘s
  - Simple: make general, widely valid statements about these associations
  - Often misspecified ‚by intention‘
- **Prediction**:
  - Transparent: formula-based predictions can be explained as/decomposed in contributions of X‘s
  - Simple: model is more easily applicable with few variables
  - Misspecification may lead to locally biased predictions
- **Explanation** (causal inference):
  - Interest in effect of an intervention on an outcome
  - Main concern: correct adjustment for confounders
  - Misspecification leads to biased effect estimate
  - Simplicity not ultimately needed; may reduce variance

Role of (algorithmic) variable selection vs. prespecification

- **Descriptive models**
  - Prespecify – if we want to describe the data in that way
  - Variable selection – to identify the main associations (‚remove noise‘)?
- **Prediction models**
  - Prespecify – predictors chosen based on availability, costs, accuracy, reliability, …
  - Variable selection – to decrease prediction error by removing noisy inputs?
- **Explanatory models**
  - Prespecify – select confounders based on strong assumptions (positivity, DAGs, …)
  - Variable selection – to decrease MSE of estimator?

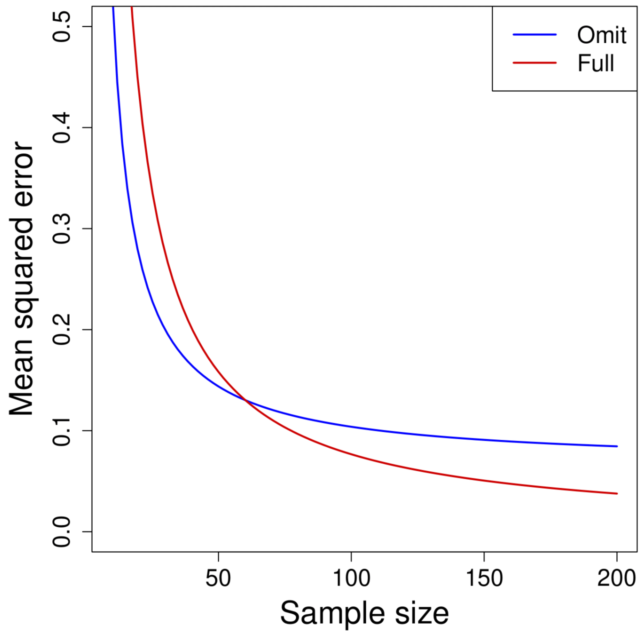
Motivation for omission: to reduce MSE?

- M$_1$: $\beta_0 + \beta_1 X_1 + \beta_2 X_2$
- M$_2$: $\theta_0 + \theta_1 X_1$
- Omission of X_2 successful (in terms of MSE of β_1) if:

$$\text{Bias}^2_{\text{omit}} < \text{Variance}_{\text{full}} - \text{Variance}_{\text{omit}}$$

Independent of *N*

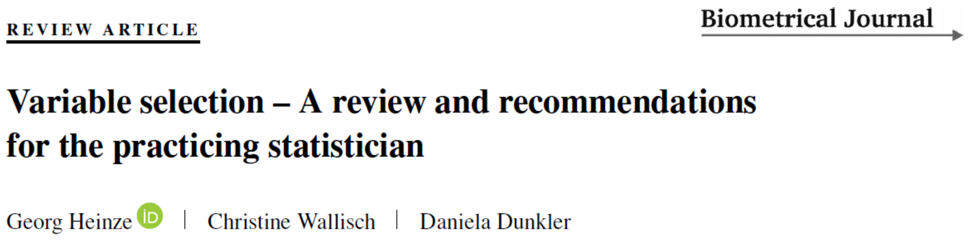
Inversely proportional to *N*



Success of ‚always omit‘ depends on sample size

Luijken K, Groenwold RHH, van Smeden M, Strohmaier S, Heinze G*:*
A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data
Biometrical Journal 2022, accepted

Consequences of variable selection



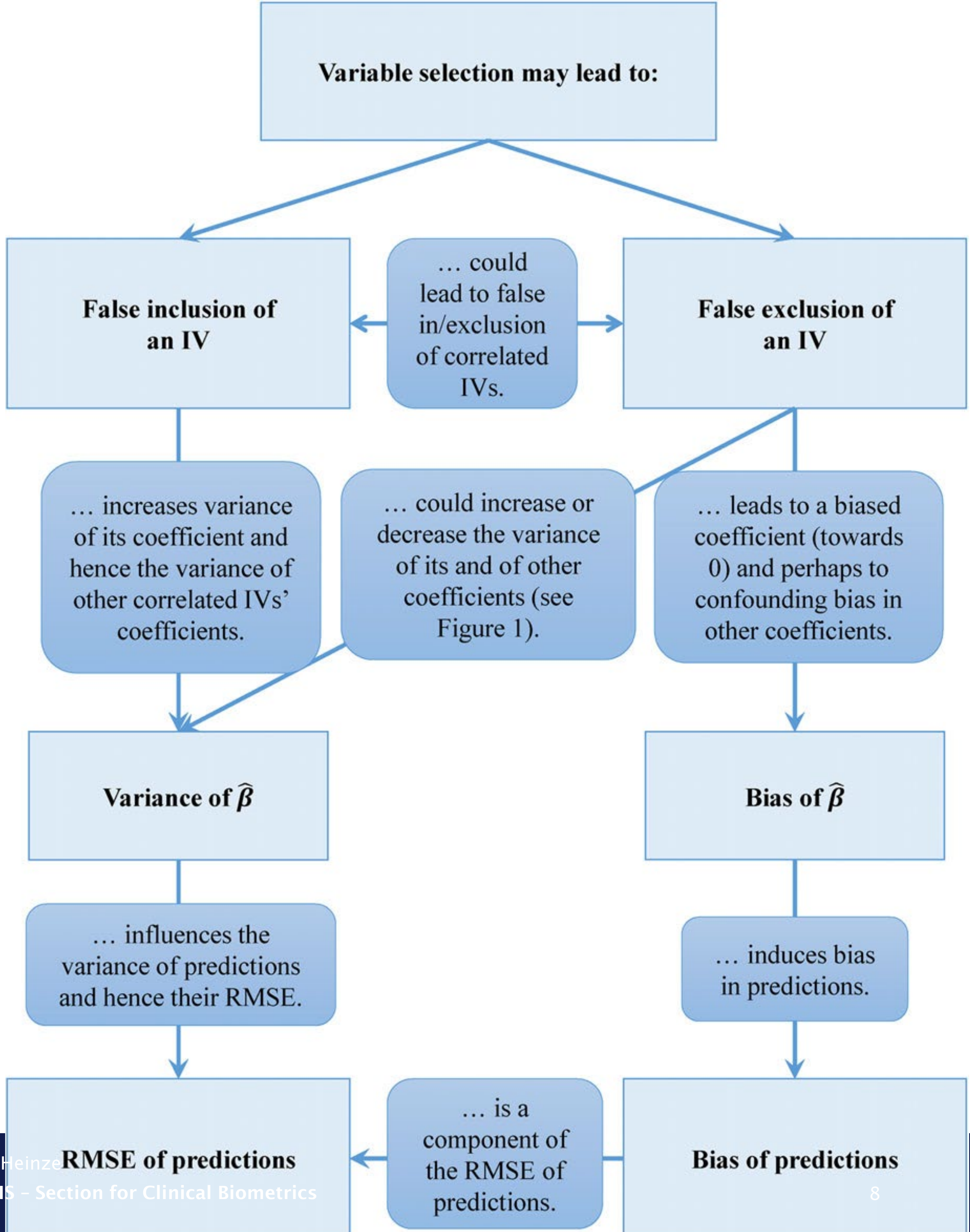
REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

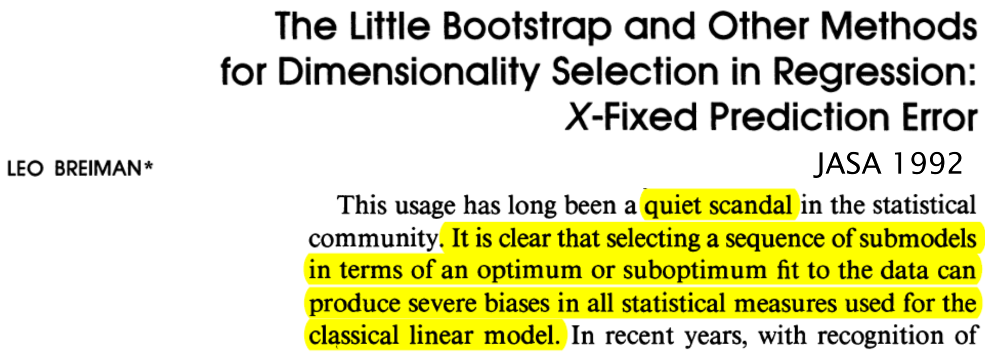
Georg Heinze | Christine Wallisch | Daniela Dunkler

- The probability of false selections is quite high (multiplicity, sequential testing, sampling variability, …)
- Simulations and resampling suggest that the ‚true' Data Generating Model can hardly be identified.



Variable selection may lead to:
False inclusion of an IV
… could lead to false in/exclusion of correlated IVs.
False exclusion of an IV
… increases variance of its coefficient and hence the variance of other correlated IVs' coefficients.
… could increase or decrease the variance of its and of other coefficients (see Figure 1).
… leads to a biased coefficient (towards 0) and perhaps to confounding bias in other coefficients.
Variance of $\widehat{\beta}$
Bias of $\widehat{\beta}$
… influences the variance of predictions and hence their RMSE.
… induces bias in predictions.
RMSE of predictions
… is a component of the RMSE of predictions.
Bias of predictions

Variable selection- a quiet scandal?



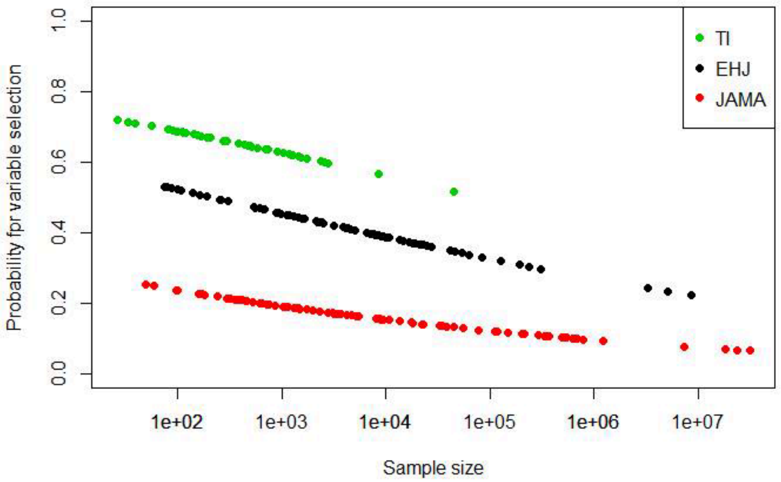
The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: *X*-Fixed Prediction Error

LEO BREIMAN*

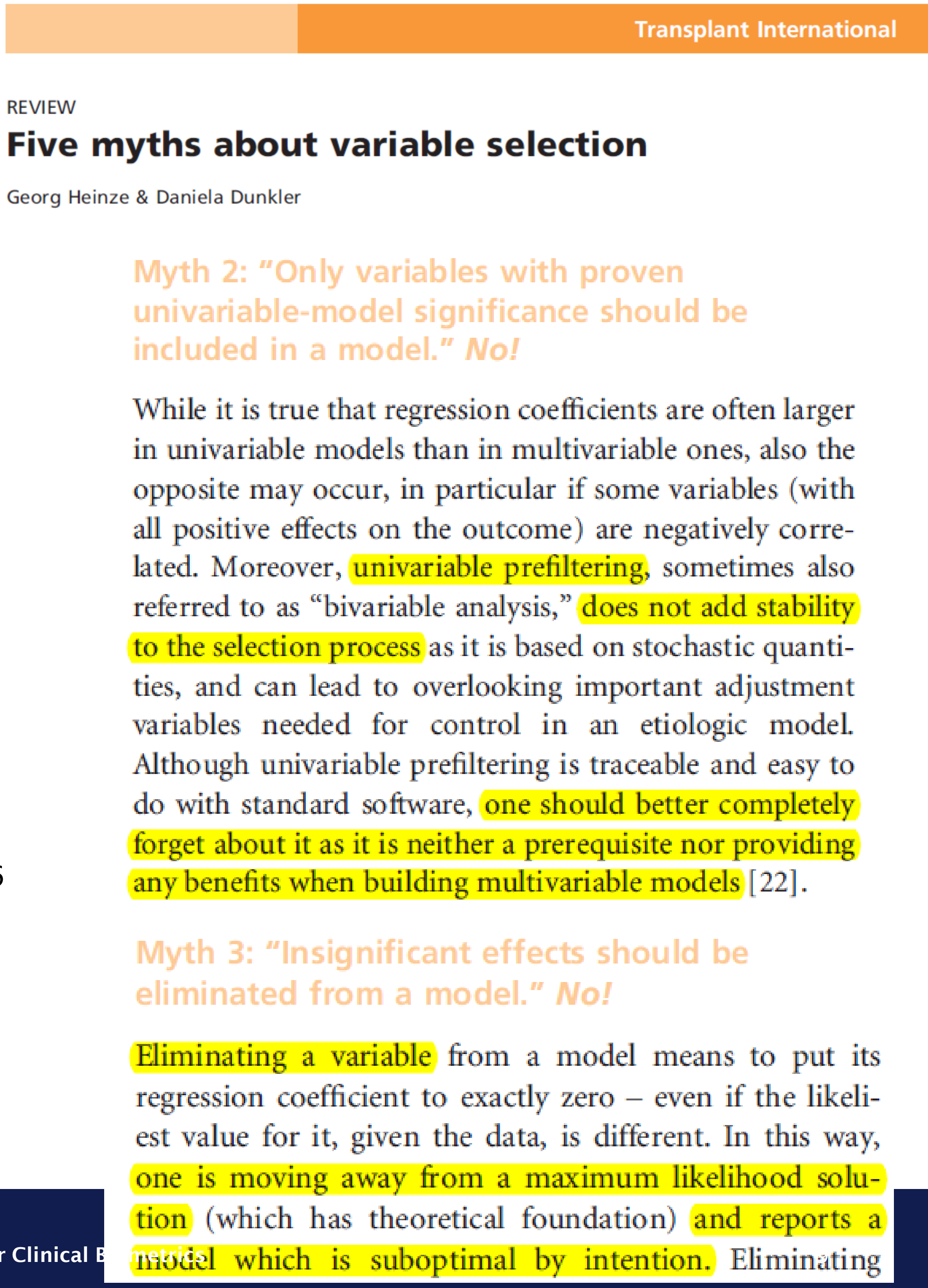
JASA 1992

This usage has long been a quiet scandal in the statistical community. It is clear that selecting a sequence of submodels in terms of an optimum or suboptimum fit to the data can produce severe biases in all statistical measures used for the classical linear model. In recent years, with recognition of

- While instabilities are to be expected, we found that use of variable selection methods is inversely correlated with sample size



Gläser & Hillebrand, B.Sc. Thesis, 2016



Transplant International

REVIEW

Five myths about variable selection

Georg Heinze & Daniela Dunkler

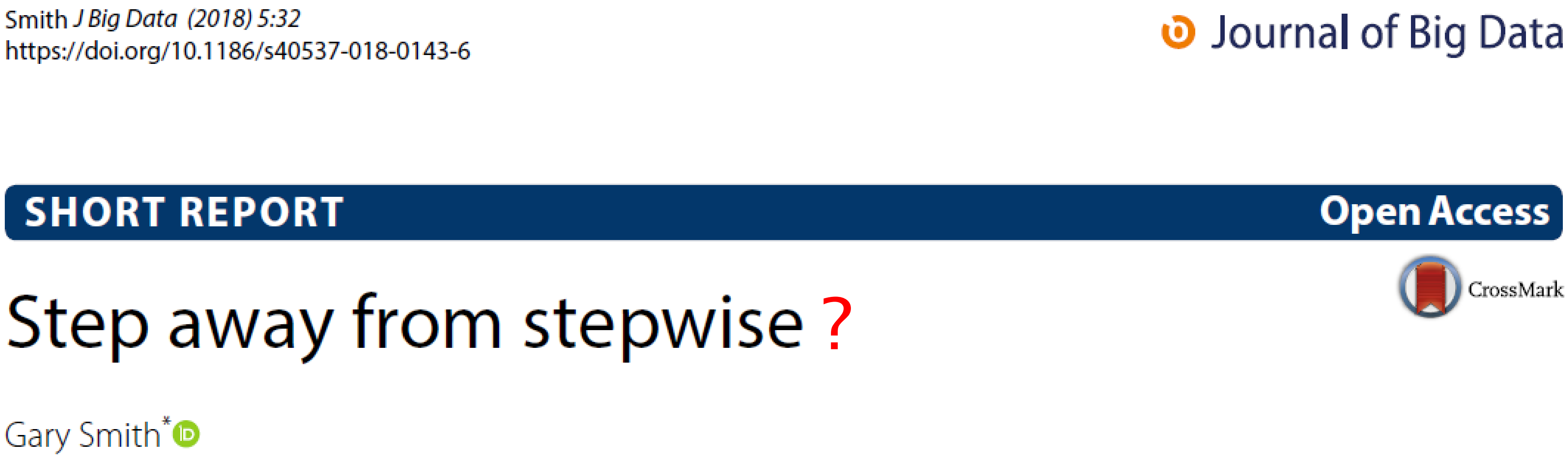
Myth 2: "Only variables with proven univariable-model significance should be included in a model." *No!*

While it is true that regression coefficients are often larger in univariable models than in multivariable ones, also the opposite may occur, in particular if some variables (with all positive effects on the outcome) are negatively correlated. Moreover, univariable prefiltering, sometimes also referred to as "bivariable analysis," does not add stability to the selection process as it is based on stochastic quantities, and can lead to overlooking important adjustment variables needed for control in an etiologic model. Although univariable prefiltering is traceable and easy to do with standard software, one should better completely forget about it as it is neither a prerequisite nor providing any benefits when building multivariable models [22].

Myth 3: "Insignificant effects should be eliminated from a model." *No!*

Eliminating a variable from a model means to put its regression coefficient to exactly zero – even if the likeliest value for it, given the data, is different. In this way, one is moving away from a maximum likelihood solution (which has theoretical foundation) and reports a model which is suboptimal by intention. Eliminating

So…



Smith *J Big Data (2018) 5:32*
https://doi.org/10.1186/s40537-018-0143-6

Journal of Big Data

SHORT REPORT **Open Access**

CrossMark

Step away from stepwise ?

Gary Smith*

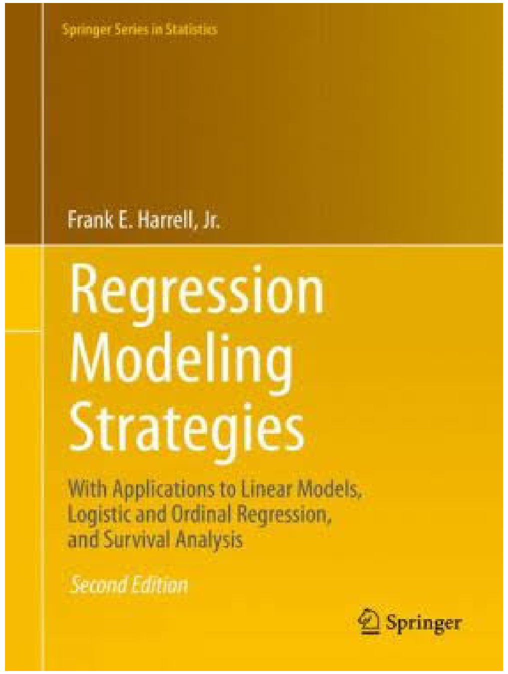
- *"Exact distributional results are virtually impossible to obtain, even for simplest of common subset selection algorithms"*

  Picard & Cook, JASA, 1984

Opinions on variable selection

- for models with focus on prediction and description.

☹ ← **Variable selection** → ☺


Springer Series in Statistics

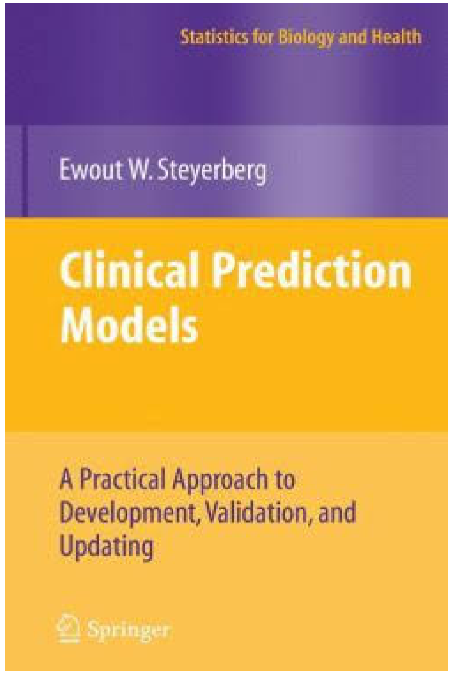
Frank E. Harrell, Jr.

Regression Modeling Strategies

With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis

Second Edition

Springer

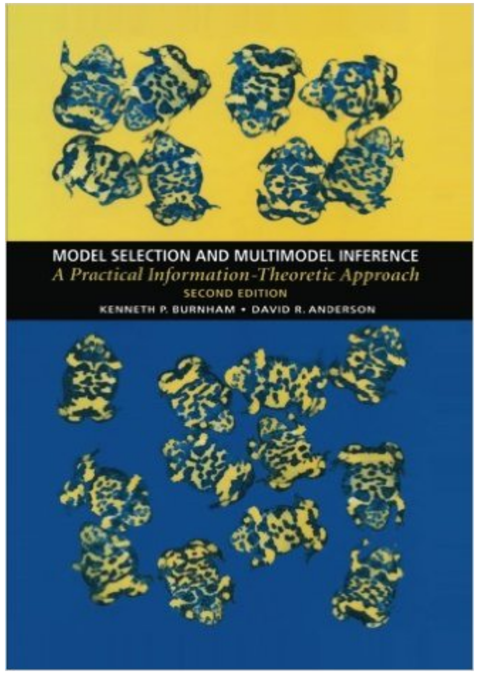

Statistics for Biology and Health

Ewout W. Steyerberg

Clinical Prediction Models

A Practical Approach to Development, Validation, and Updating

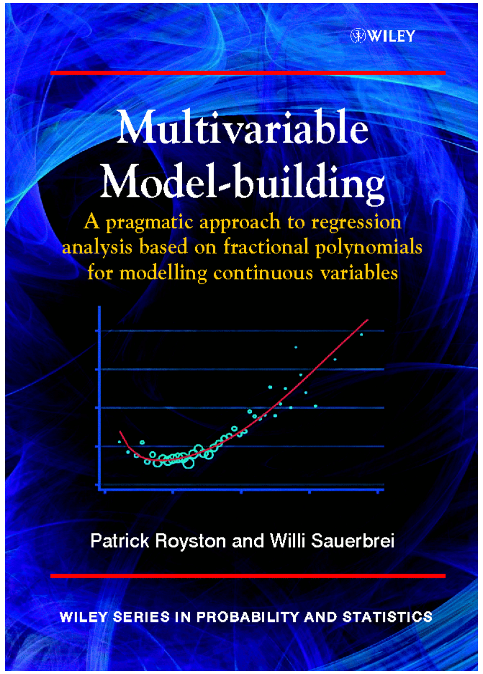
Springer


MODEL SELECTION AND MULTIMODEL INFERENCE

A Practical Information-Theoretic Approach

SECOND EDITION

KENNETH P. BURNHAM • DAVID R. ANDERSON


WILEY

Multivariable Model-building

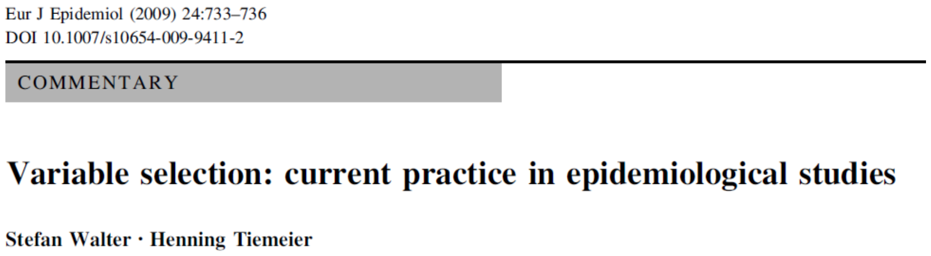
A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables

Patrick Royston and Willi Sauerbrei

WILEY SERIES IN PROBABILITY AND STATISTICS

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

Some reviews on use of variable selection methods

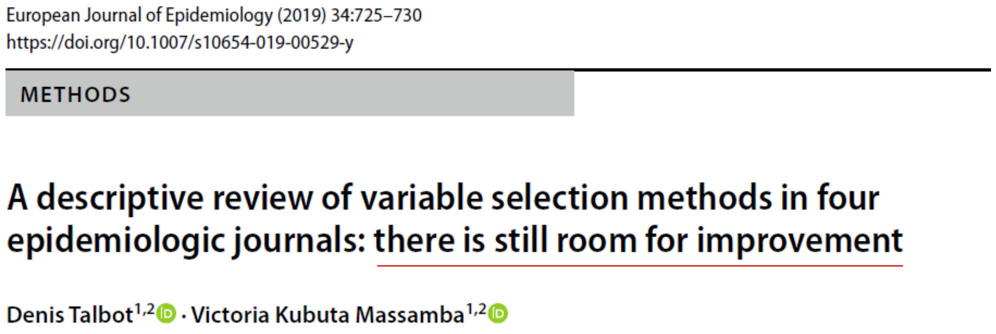


Eur J Epidemiol (2009) 24:733–736
DOI 10.1007/s10654-009-9411-2

COMMENTARY

Variable selection: current practice in epidemiological studies

Stefan Walter · Henning Tiemeier

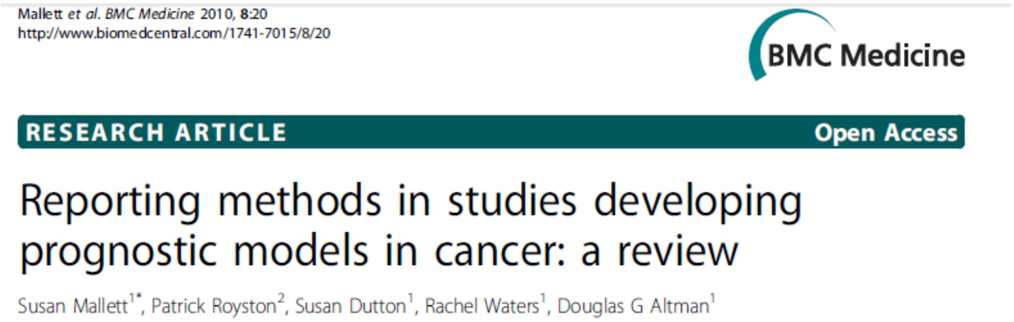


European Journal of Epidemiology (2019) 34:725–730
https://doi.org/10.1007/s10654-019-00529-y

METHODS

A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement

Denis Talbot[1,2] · Victoria Kubuta Massamba[1,2]



Mallett *et al. BMC Medicine* 2010, 8:20
http://www.biomedcentral.com/1741-7015/8/20

BMC Medicine

RESEARCH ARTICLE Open Access

Reporting methods in studies developing prognostic models in cancer: a review

Susan Mallett[1*], Patrick Royston[2], Susan Dutton[1], Rachel Waters[1], Douglas G Altman[1]

Intensive Care Med (2021) 47:1487–1489
https://doi.org/10.1007/s00134-021-06535-7

LETTER

Quiet scandal: variable selection in three major intensive care medicine journals

Check for updates

Charles-Hervé Vacheron[1,2,3*], Arnaud Friggeri[1,4], Bernard Allaouchiche[2,5,6], Delphine Maucort-Boulch[3,7,8] and Esla Coz[3,7]

Use of VS ranged from:
Stepwise: 2-33%
Univariate sel.: 2-44%
much depending on quality/culture of journal.

ELSEVIER

Journal of Clinical Epidemiology 139 (2021) 12–19

Epidemiology

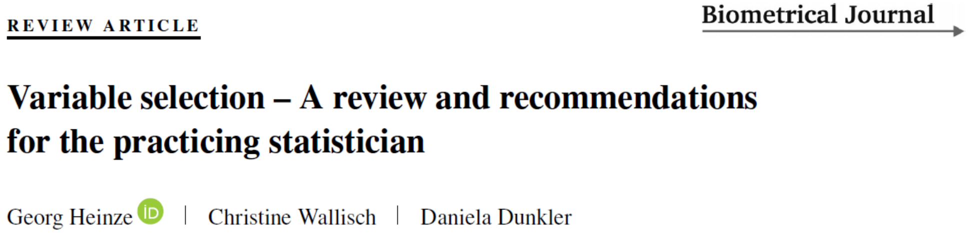
ORIGINAL ARTICLE

Variable selection methods were poorly reported but rarely misused in major medical journals: Literature review

T. Pressat-Laffouilhère[a,b,c,*], R. Jouffroy[d,e,f], A. Leguillou[g], G. Kerdelhue[b], J. Benichou[a,e], A. Gillibert[a]

Variable selection reviews

- Traditional, Modern, Bayesian

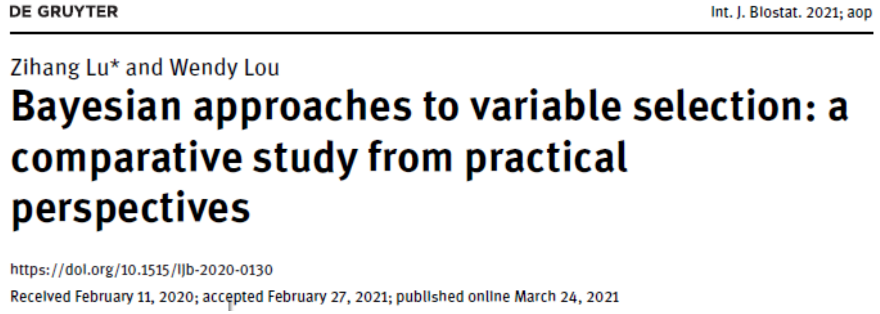


REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler



DE GRUYTER

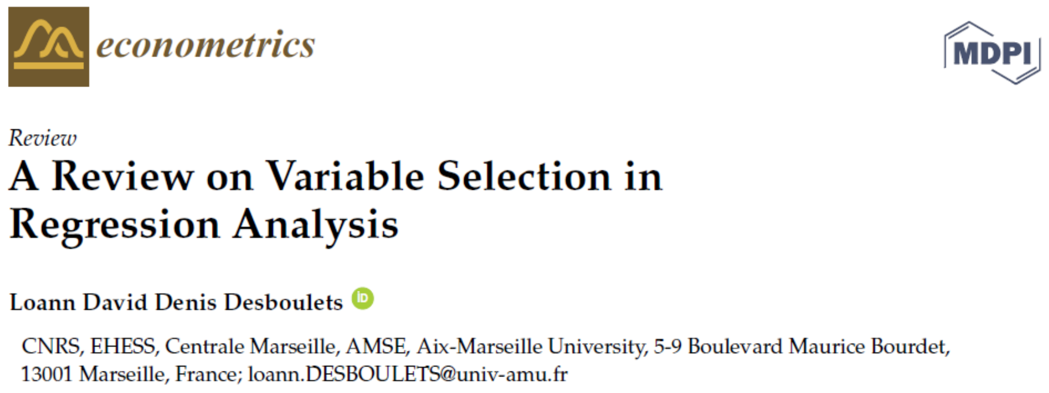
Int. J. Biostat. 2021; aop

Zihang Lu* and Wendy Lou

Bayesian approaches to variable selection: a comparative study from practical perspectives

https://doi.org/10.1515/ijb-2020-0130

Received February 11, 2020; accepted February 27, 2021; published online March 24, 2021



econometrics

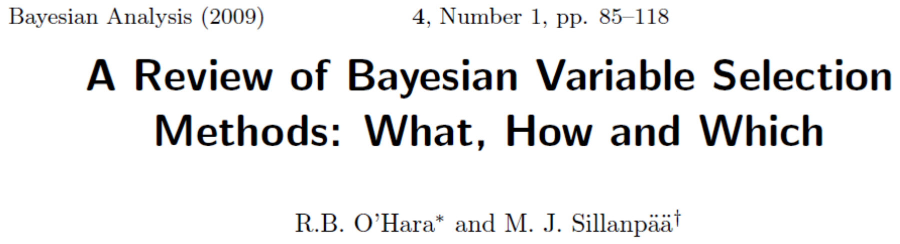
MDPI

Review

A Review on Variable Selection in Regression Analysis

Loann David Denis Desboulets

CNRS, EHESS, Centrale Marseille, AMSE, Aix-Marseille University, 5-9 Boulevard Maurice Bourdet, 13001 Marseille, France; loann.DESBOULETS@univ-amu.fr

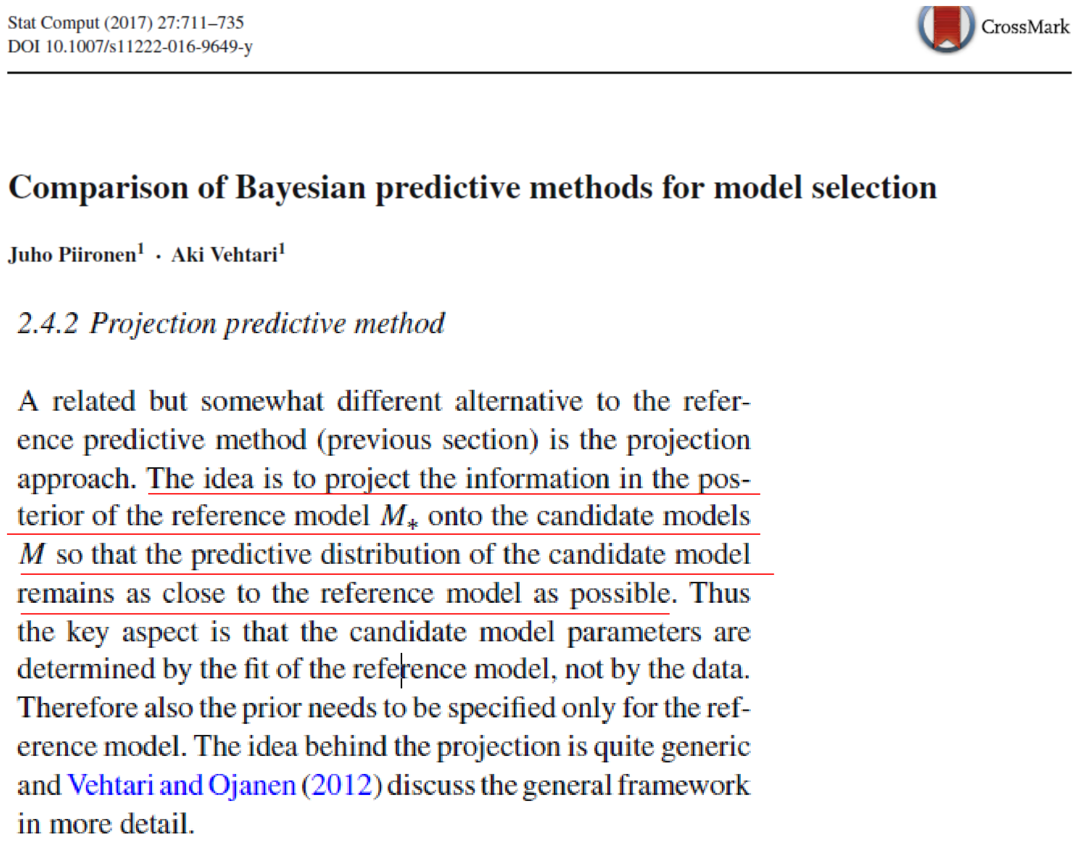


Bayesian Analysis (2009) 4, Number 1, pp. 85–118

A Review of Bayesian Variable Selection Methods: What, How and Which

R.B. O'Hara* and M. J. Sillanpää†

Projection prediction



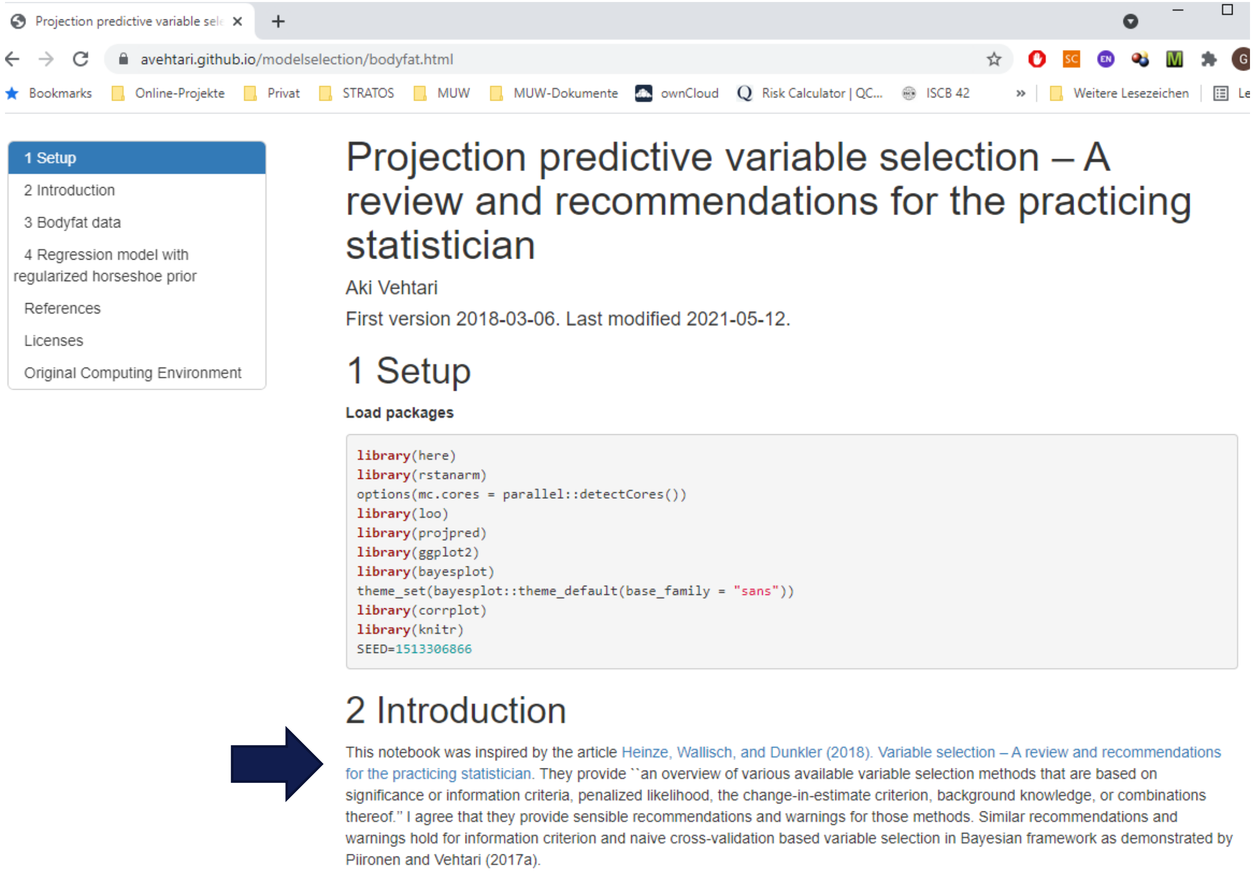
Stat Comput (2017) 27:711–735
DOI 10.1007/s11222-016-9649-y

Comparison of Bayesian predictive methods for model selection

Juho Piironen[1] · Aki Vehtari[1]

2.4.2 Projection predictive method

A related but somewhat different alternative to the reference predictive method (previous section) is the projection approach. The idea is to project the information in the posterior of the reference model M_* onto the candidate models M so that the predictive distribution of the candidate model remains as close to the reference model as possible. Thus the key aspect is that the candidate model parameters are determined by the fit of the reference model, not by the data. Therefore also the prior needs to be specified only for the reference model. The idea behind the projection is quite generic and Vehtari and Ojanen (2012) discuss the general framework in more detail.



Projection predictive variable selection – A review and recommendations for the practicing statistician

Aki Vehtari

First version 2018-03-06. Last modified 2021-05-12.

1 Setup

Load packages

```
library(here)
library(rstanarm)
options(mc.cores = parallel::detectCores())
library(loo)
library(projpred)
library(ggplot2)
library(bayesplot)
theme_set(bayesplot::theme_default(base_family = "sans"))
library(corrplot)
library(knitr)
SEED=1513306866
```

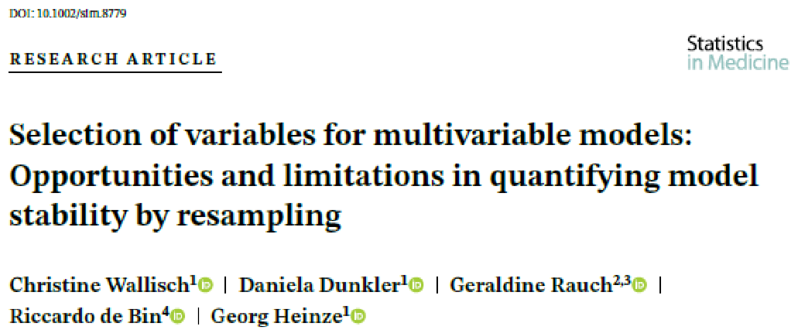
2 Introduction

This notebook was inspired by the article Heinze, Wallisch, and Dunkler (2018). Variable selection – A review and recommendations for the practicing statistician. They provide ``an overview of various available variable selection methods that are based on significance or information criteria, penalized likelihood, the change-in-estimate criterion, background knowledge, or combinations thereof.'' I agree that they provide sensible recommendations and warnings for those methods. Similar recommendations and warnings hold for information criterion and naive cross-validation based variable selection in Bayesian framework as demonstrated by Piironen and Vehtari (2017a).

Inspired by our review, reanalyzed the bodyfat data set

Traditional VS techniques – towards recommendations

- Purpose of modeling?? Do you need to let your data decide on included variables?
- Accept that model selection is part of estimation process and comes with uncertainties.
- Any VS method makes a compromise between false negative and false positive selections.
- Therefore, **describe instabilities**!
  → Bootstrap, subsampling



DOI: 10.1002/sim.8779

RESEARCH ARTICLE

Statistics in Medicine

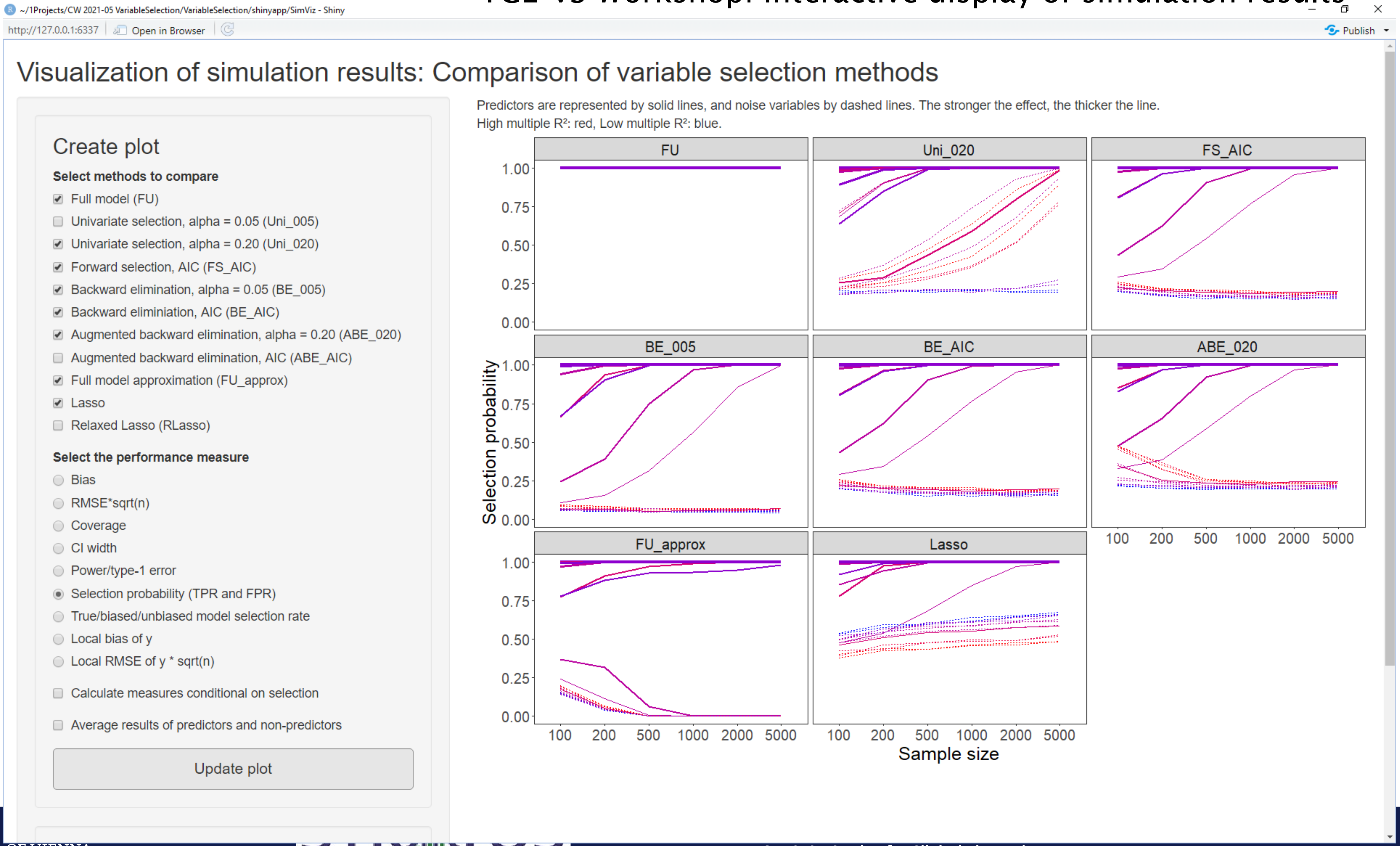
Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling

Christine Wallisch[1] | Daniela Dunkler[1] | Geraldine Rauch[2,3] | Riccardo de Bin[4] | Georg Heinze[1]

StatMed 2021

- Success of VS is largely driven by sample size
- Recommendations must take sample size into accout.

TG2 VS Workshop: interactive display of simulation results



~/1Projects/CW 2021-05 VariableSelection/VariableSelection/shinyapp/SimViz - Shiny

http://127.0.0.1:6337 | Open in Browser | Publish

Visualization of simulation results: Comparison of variable selection methods

Create plot

Select methods to compare

- ☑ Full model (FU)
- □ Univariate selection, alpha = 0.05 (Uni_005)
- ☑ Univariate selection, alpha = 0.20 (Uni_020)
- ☑ Forward selection, AIC (FS_AIC)
- ☑ Backward elimination, alpha = 0.05 (BE_005)
- ☑ Backward eliminiation, AIC (BE_AIC)
- ☑ Augmented backward elimination, alpha = 0.20 (ABE_020)
- □ Augmented backward elimination, AIC (ABE_AIC)
- ☑ Full model approximation (FU_approx)
- ☑ Lasso
- □ Relaxed Lasso (RLasso)

Select the performance measure

- ○ Bias
- ○ RMSE*sqrt(n)
- ○ Coverage
- ○ CI width
- ○ Power/type-1 error
- ● Selection probability (TPR and FPR)
- ○ True/biased/unbiased model selection rate
- ○ Local bias of y
- ○ Local RMSE of y * sqrt(n)

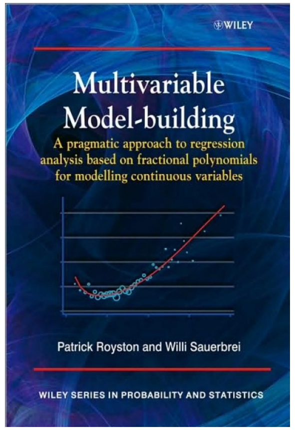
□ Calculate measures conditional on selection

□ Average results of predictors and non-predictors

Update plot

Predictors are represented by solid lines, and noise variables by dashed lines. The stronger the effect, the thicker the line.
High multiple R²: red, Low multiple R²: blue.

Combining variable and function selection



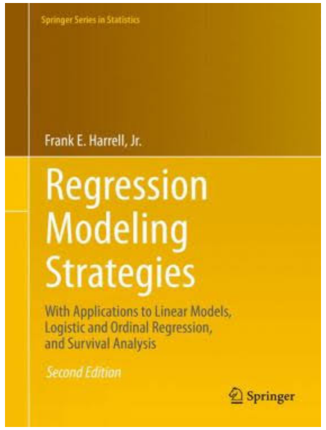
WILEY

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



Springer Series in Statistics

Frank E. Harrell, Jr.

Regression Modeling Strategies

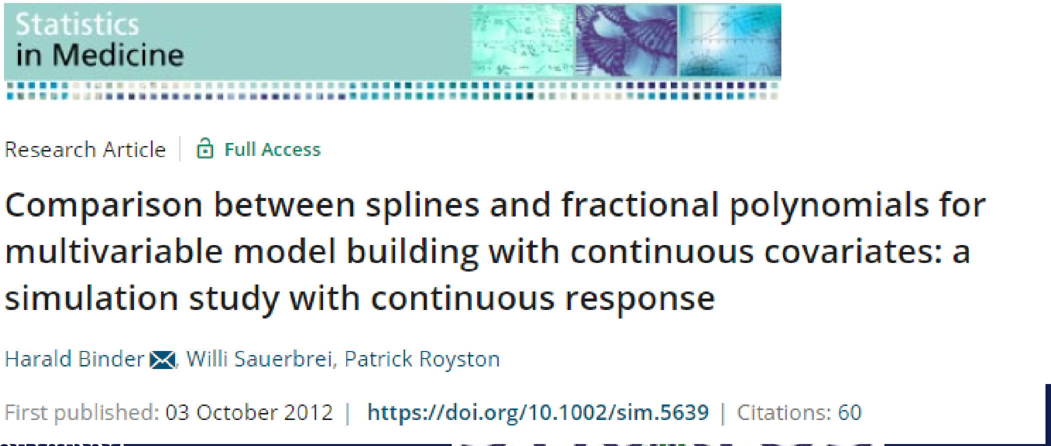
With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis

Second Edition

Springer

- Two inter-related questions in model building

- Multivariable fractional polynomials (MFP)
- Various spline based approaches
- Comparisons:
- MFP vs. Splines:



Statistics in Medicine

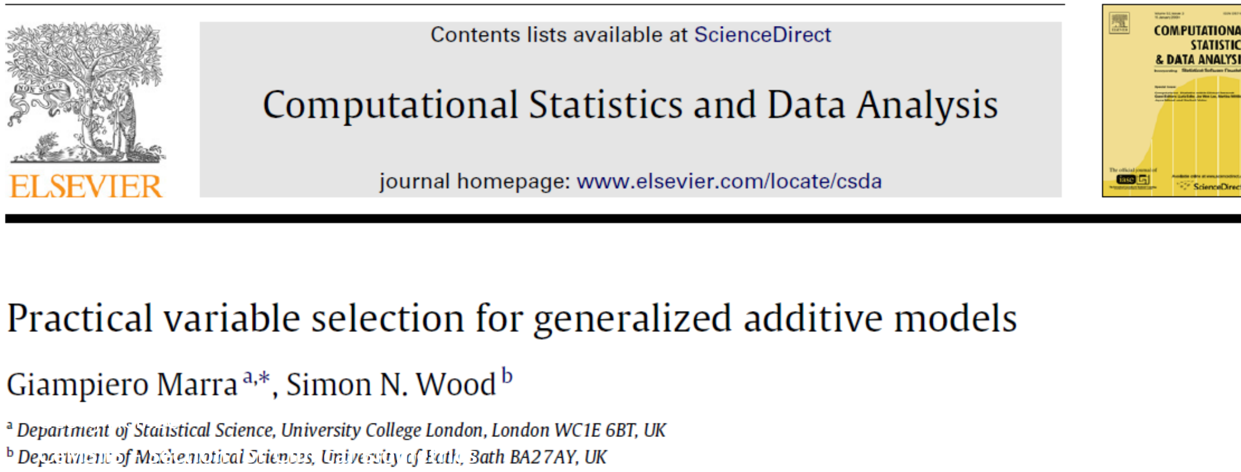
Research Article | Full Access

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

Harald Binder, Willi Sauerbrei, Patrick Royston

First published: 03 October 2012 | https://doi.org/10.1002/sim.5639 | Citations: 60

Different versions of (penalized) splines:



Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

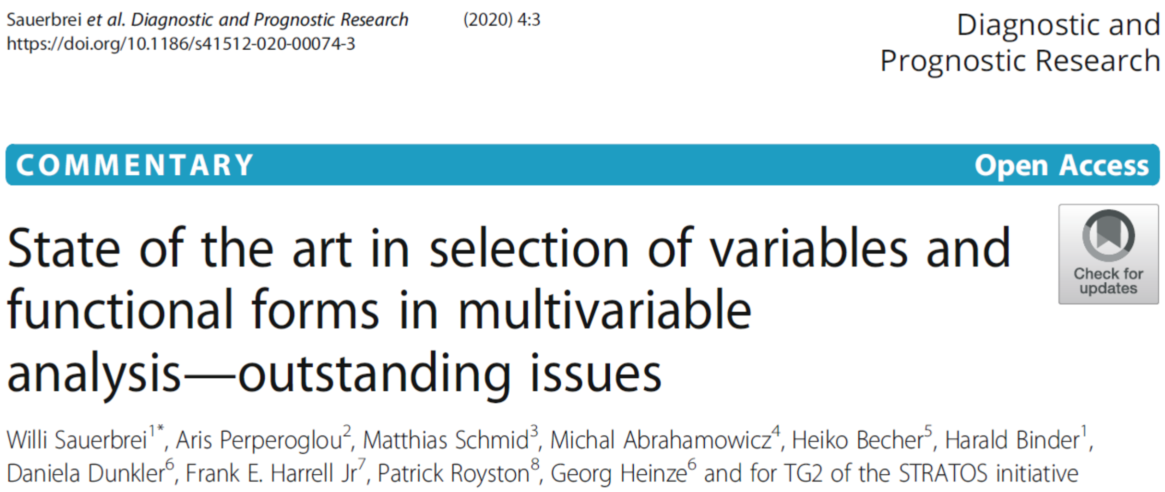
ELSEVIER

Practical variable selection for generalized additive models

Giampiero Marra [a,*], Simon N. Wood [b]

[a] Department of Statistical Science, University College London, London WC1E 6BT, UK

[b] Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK



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!

We need you!

- Review,
- Applications,
- Neutral simulation studies,
- Recommendations.

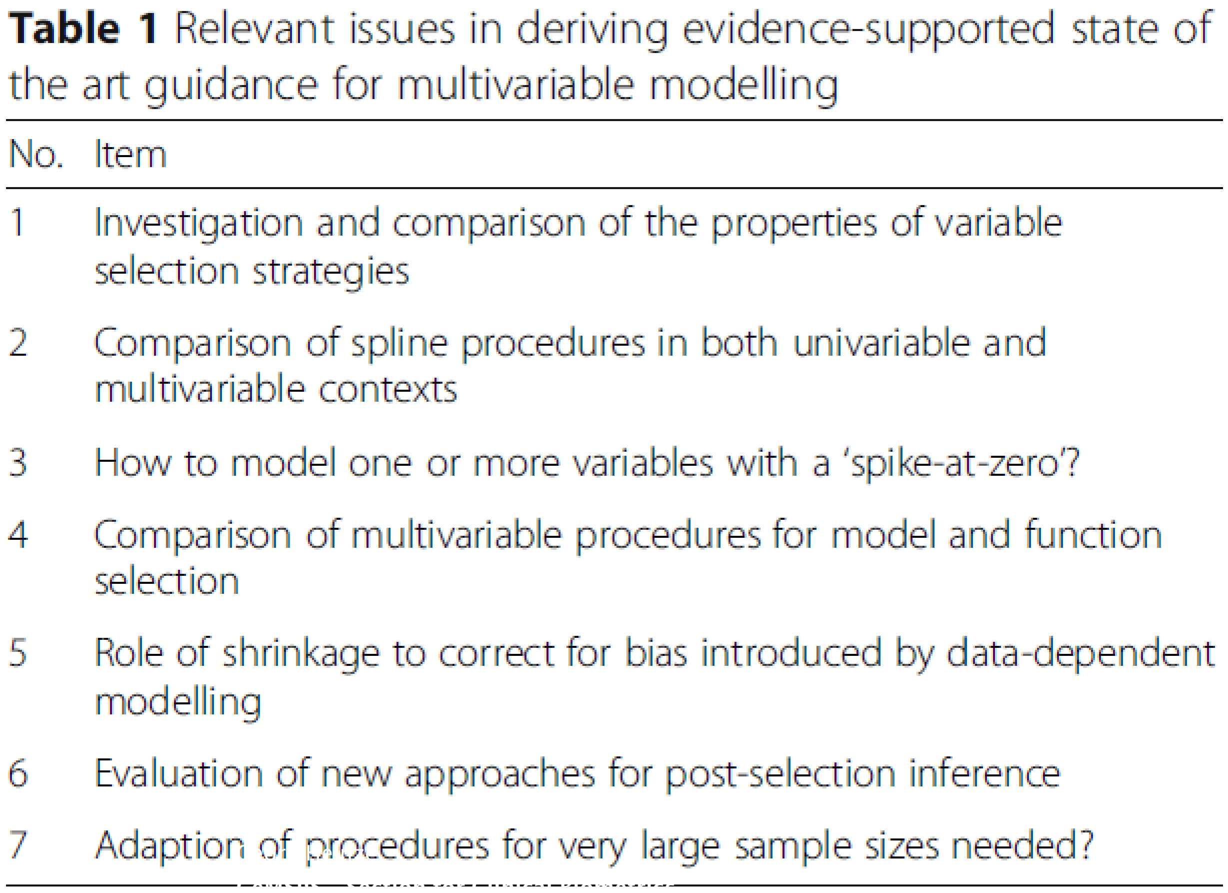
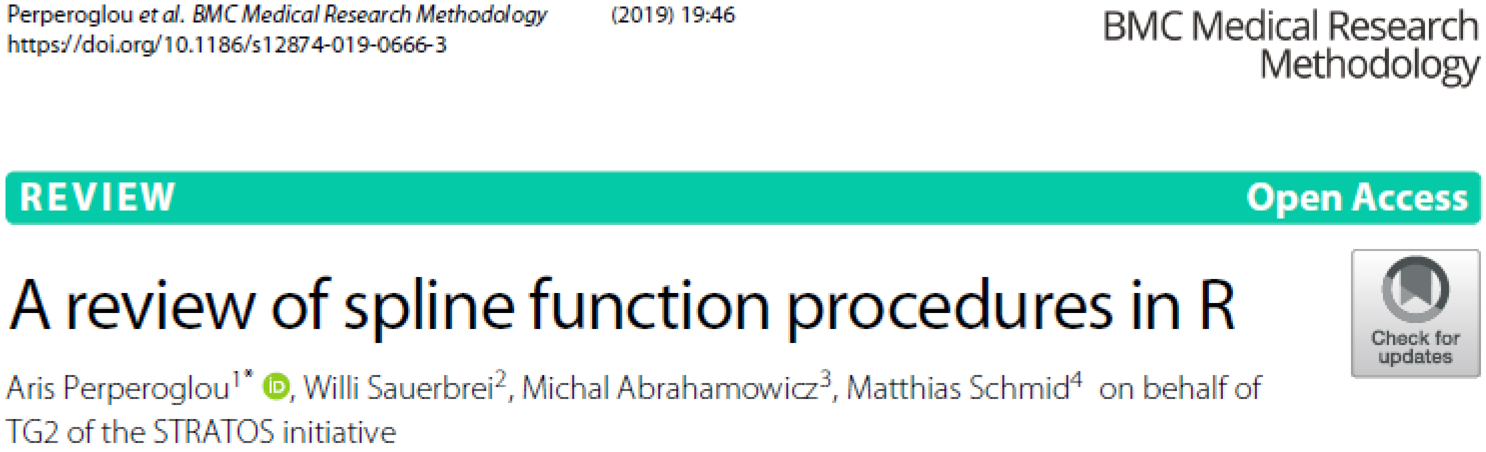


Table 1 Relevant issues in deriving evidence-supported state of the art guidance for multivariable modelling

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?

Splines - a brief overview of regression packages in R



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

Check for updates

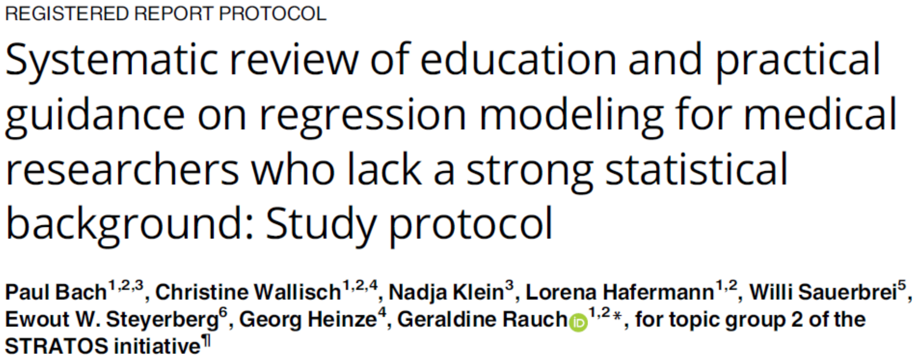
Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

Package	Downloads	Vignette	Book	Website	Datasets
quantreg	5099669	X	X		**8**
survival	3511997	X	X		**38**
mgcv	3217720	X	X		**2**
gbm	668984			X	**0**
VGAM	662399	X	X	X	**50**
gam	459497		X	X	**4**
gamlss	**210761**	**X**	**X**	**X**	**43**

What guidance is out there for data analysis with limited statistical training?

PLOS ONE

2020

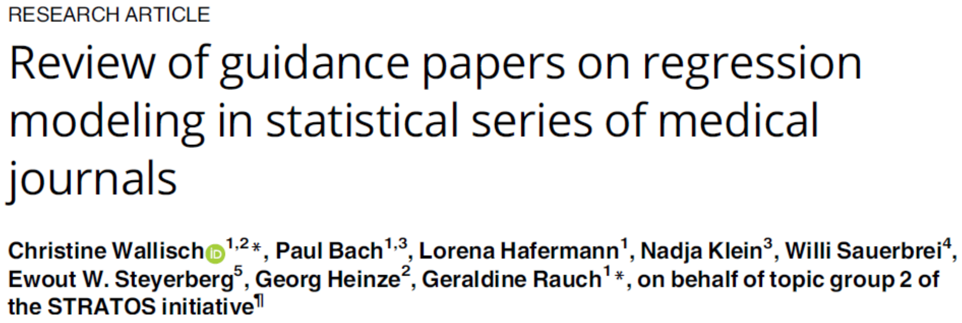


REGISTERED REPORT PROTOCOL

Systematic review of education and practical guidance on regression modeling for medical researchers who lack a strong statistical background: Study protocol

Paul Bach[1,2,3], Christine Wallisch[1,2,4], Nadja Klein[3], Lorena Hafermann[1,2], Willi Sauerbrei[5], Ewout W. Steyerberg[6], Georg Heinze[4], Geraldine Rauch[1,2]*, for topic group 2 of the STRATOS initiative¶

2021



RESEARCH ARTICLE

Review of guidance papers on regression modeling in statistical series of medical journals

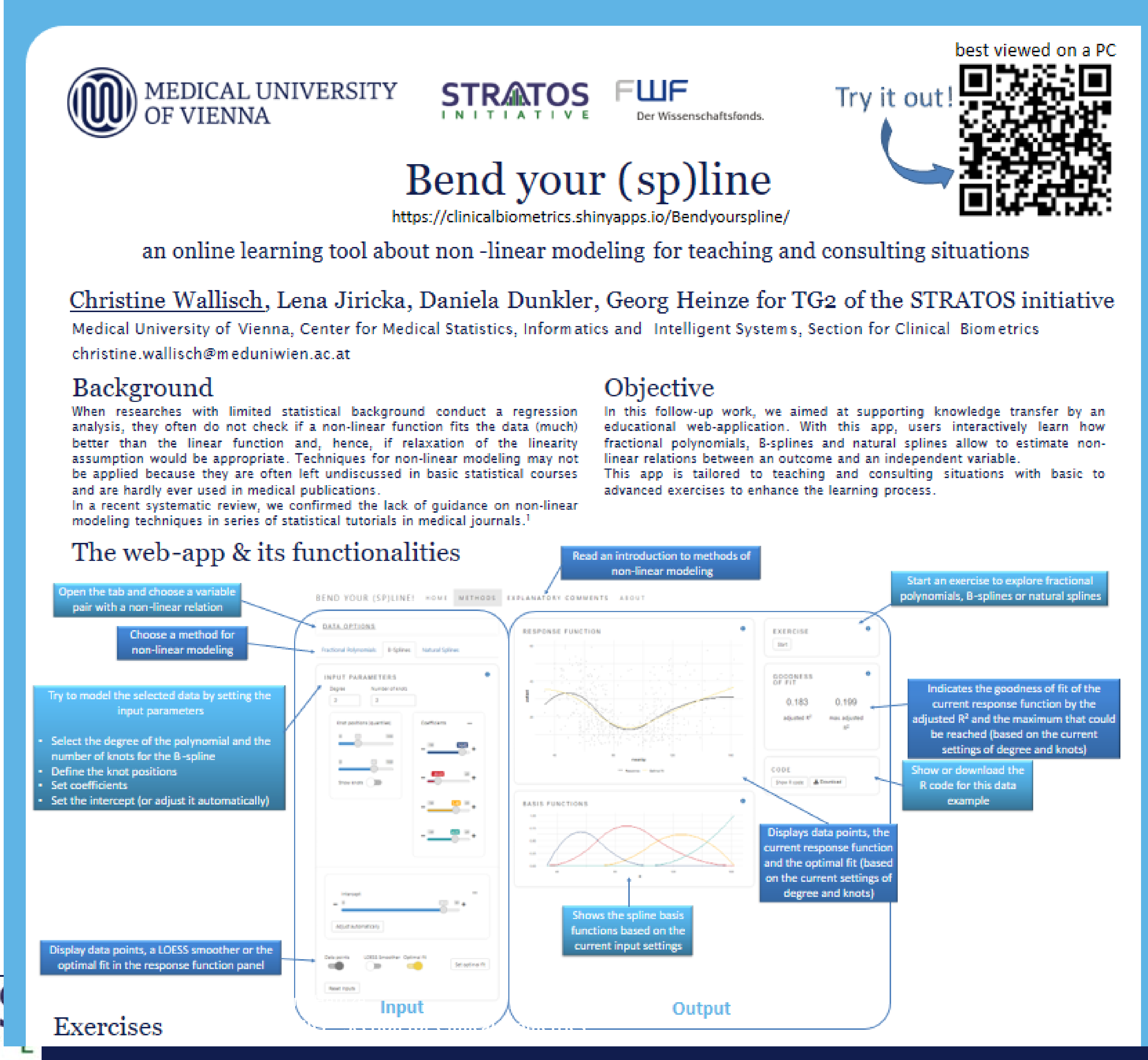
Christine Wallisch[1,2]*, Paul Bach[1,3], Lorena Hafermann[1], Nadja Klein[3], Willi Sauerbrei[4], Ewout W. Steyerberg[5], Georg Heinze[2], Geraldine Rauch[1]*, on behalf of topic group 2 of the STRATOS initiative¶

- We identified 23 series including 57 topic-relevant articles. Within each article, two independent raters analyzed the content by investigating 44 predefined aspects on regression modeling.
- Some papers could be recommended
  e.g. Nature Methods series

Further activities

- The shiny app ‚Bend your (sp)line‘:

- (See the poster at DAGStat!)



MEDICAL UNIVERSITY OF VIENNA

STRATOS INITIATIVE

FWF Der Wissenschaftsfonds.

best viewed on a PC

Try it out!

Bend your (sp)line

https://clinicalbiometrics.shinyapps.io/Bendyourspline/

an online learning tool about non -linear modeling for teaching and consulting situations

Christine Wallisch, Lena Jiricka, Daniela Dunkler, Georg Heinze for TG2 of the STRATOS initiative

Medical University of Vienna, Center for Medical Statistics, Informatics and Intelligent Systems, Section for Clinical Biometrics

christine.wallisch@meduniwien.ac.at

Background

When researches with limited statistical background conduct a regression analysis, they often do not check if a non-linear function fits the data (much) better than the linear function and, hence, if relaxation of the linearity assumption would be appropriate. Techniques for non-linear modeling may not be applied because they are often left undiscussed in basic statistical courses and are hardly ever used in medical publications.

In a recent systematic review, we confirmed the lack of guidance on non-linear modeling techniques in series of statistical tutorials in medical journals.[1]

Objective

In this follow-up work, we aimed at supporting knowledge transfer by an educational web-application. With this app, users interactively learn how fractional polynomials, B-splines and natural splines allow to estimate non-linear relations between an outcome and an independent variable.

This app is tailored to teaching and consulting situations with basic to advanced exercises to enhance the learning process.

The web-app & its functionalities

Read an introduction to methods of non-linear modeling

Open the tab and choose a variable pair with a non-linear relation

Choose a method for non-linear modeling

Try to model the selected data by setting the input parameters

- Select the degree of the polynomial and the number of knots for the B-spline
- Define the knot positions
- Set coefficients
- Set the intercept (or adjust it automatically)

Display data points, a LOESS smoother or the optimal fit in the response function panel

Start an exercise to explore fractional polynomials, B-splines or natural splines

Indicates the goodness of fit of the current response function by the adjusted R² and the maximum that could be reached (based on the current settings of degree and knots)

Show or download the R code for this data example

Displays data points, the current response function and the optimal fit (based on the current settings of degree and knots)

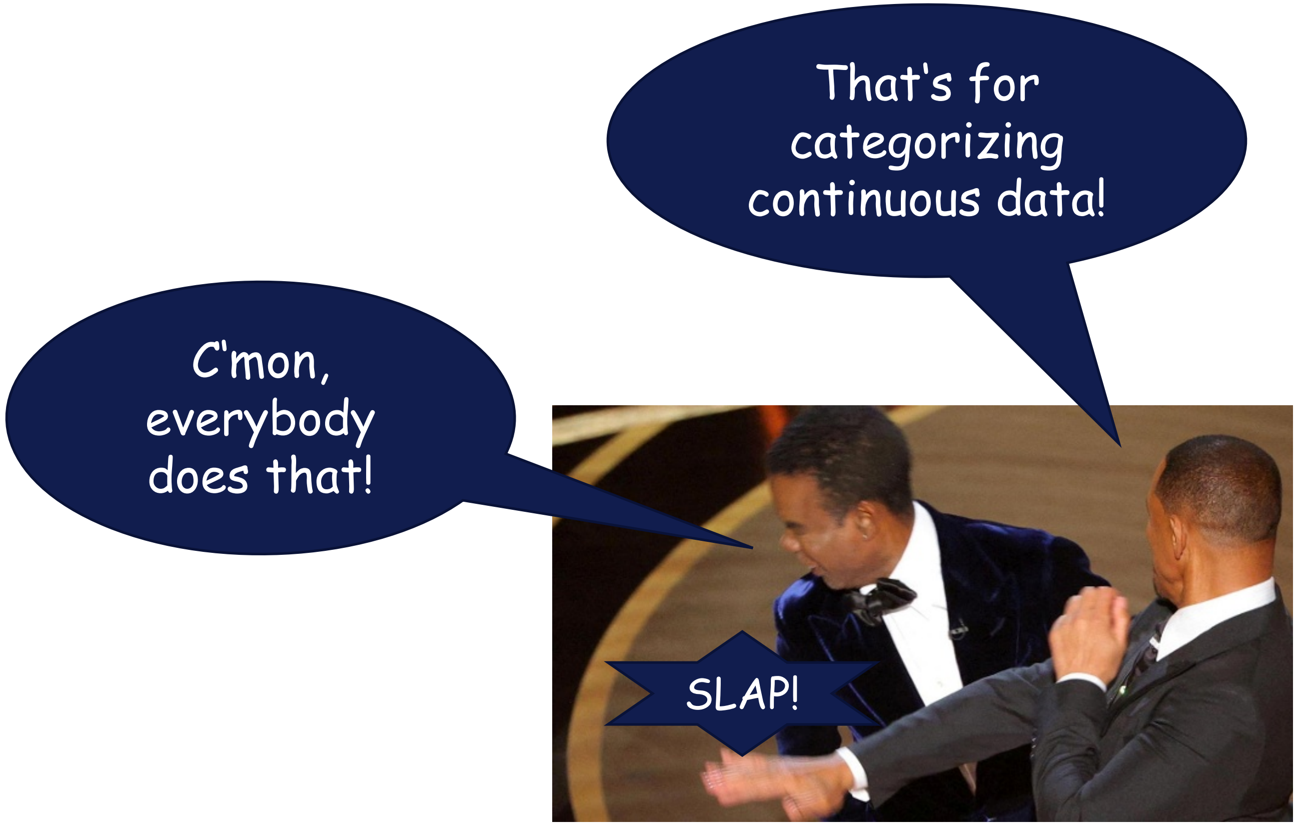
Shows the spline basis functions based on the current input settings

Exercises

Further activities

- A series of short videos:

- We plan to create a series of short videos explaining concepts in compact way:
- Topics e.g.:
  - *What is a statistical model?*
  - *Categorization*
  - *Linear or nonlinear functional relationship?*
  - *Variable selection and instability of selected models*
  - *Full model specification or variable selection?*


That's for categorizing continuous data!
C'mon, everybody does that!
SLAP!

Members of TG2

- Georg Heinze, Aris Perperoglou, Willi Sauerbrei (co-chairs)
- Michal Abrahamowicz, Heiko Becher, Harald Binder, Thomas Cowling, Daniela Dunkler, Rolf Groenwold, Frank Harrell, Nadja Klein, Geraldine Rauch, Patrick Royston, Matthias Schmid, Christine Wallisch (members)
- Edwin Kipruto, Kim Luijken, Michael Kammer
  (early career adjunct members)

- @Georg__Heinze
  georg.heinze@meduniwien.ac.at