Variable selection for statistical models: a review and recommendations for the practicing statistician

Georg Heinze, Christine Wallisch and Daniela Dunkler

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

for Topic Group 2 of the STRATOS initiative
(www.stratos-initiative.org)

E-mail: georg.heinze@meduniwien.ac.at

Statistical models are important tools in empirical medical research. They facilitate individualized outcome prognostication conditional on covariates as well as adjustments of estimated effects of covariates on the outcome. Theory of statistical models is well-established if the set of covariates to consider is fixed and small, such that we can assume that effect estimates are unbiased and the usual methods for confidence interval estimation are valid. In routine work, however, it is not known a priori which covariates should be included in a model, and often we are confronted with the number of candidate variables in the range 10-25. This number is often too large to be considered in a statistical model.

In recent decades many statisticians have extensively studied variable selection procedures for various purposes, e.g., for adjusting the effect of a risk factor of interest for confounders or other covariates, for hypothesis testing, or for deriving multivariable prediction models. It has turned out that no selection procedure is generally superior to other procedures and there is no generally accepted state of the art for variable selection [1]. Unfortunately, in medical papers it is still not uncommon to use univariable selection as a screening approach to eliminate non-significant variables and use the remaining variables to build the multivariable model. This approach has severe weaknesses.   We will provide an overview of variable selection methods which are based on

a) significance or information criteria, [2; Ch. 2]
b) penalized likelihood, [3]
c) the change-in-estimate criterion, [4]
d) background knowledge, [5] or
e) combinations thereof. [6]
These methods were usually developed in the context of a linear regression model and then transferred to more general models like generalized linear models or models for censored survival data.

In this half-day workshop, we will exemplify applications of variable selection using scientific questions and data from real medical studies with different research questions focusing on descriptive models and transparent prediction models. Data of these studies are publicly available, and their analysis will be discussed by means of worked exercises with accompanying R notebooks. We will also interactively present a simulation study to investigate implications of variable selection, e.g., on uncertainty and stability of the final model [7,8], on bias and variability of regression coefficients [9], and on the validity of confidence intervals [10].

We will give pragmatic recommendations for the practitioner by suggesting typical steps to be done when variable selection is considered. We give guidance on how to pre-select candidate covariates, how to choose an appropriate variable selection method, and how to report the final model and its stability in scientific reports [11,12]. These recommendations are based on data settings with a mix of 5-25 continuous and categorical covariates that are moderately correlated (r<0.8). We also discuss some open issues that still need further investigation [1].

We will mix visual presentations with check-up questions to the audience and will demonstrate worked exercises interactively in R-Studio. Participants can follow these analyses with their own notebook, but it is not required to bring a notebook to attend and follow this course.

References:

[1] Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell Jr FE, Royston P, Heinze G, for TG2 of the STRATOS initiative. State of the art in selection of variables and functional forms in multivariable analysis – outstanding issues. Diagnostic and Prognostic Research 4:3, 2020.

[2] Royston P, Sauerbrei W. Multivariable Model-Building. A pragmatic approach to regression analysis based on fractional polynomials for modeling continuous variables. Wiley, Chichester, 2008

[3] Tibshirani R. Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58: 267–288, 1996

[4] Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. American Journal of Epidemiology 129: 125–137, 1993

[5] VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics 67: 1406–1413, 2011

[6] Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: A pragmatic and puposeful way to develop statistical models. PloS One 9(11): e113677, 2014

[7] Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. Biometrics 53: 603-618, 1997

[8] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. Statistics in Medicine 11: 2093–2109, 1992

[9] Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. Journal of Clinical Epidemiology 64(12), 1464-5, 2011

[10] Austin PC. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. Statistics in Medicine 27, 3286-3300, 2008

[11] Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. Biometrical Journal 60, 431-449, 2018

[12] Wallisch C, Dunkler D, Rauch G, de Bin R, Heinze G. Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. Statistics in Medicine 40:369-381, 2021