Education for Statistics in Practice – DAGStat 2016, Göttingen

Do. 17. März 2016
13:00-14:20 und 14:50-16:10 Uhr
Raum (vorläufig): ZHG009

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

Georg Heinze and Daniela Dunkler

Section for Clinical Biometrics
Center for Medical Statistics, Informatics and Intelligent Systems
Medical University of Vienna
Vienna, Austria
georg.heinze@meduniwien.ac.at

Statistical models are handy tools for empirical medical research. They facilitate individualized outcome prognostication conditional on covariates as well as adjustments of estimated effects of risk factors on the outcome by covariates. 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 10-30. 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, but almost all selection procedures are superior to selecting those variables for a multivariable model which show significant effects in univariable models. Nevertheless, this univariable screening method is still the most popular approach in the medical literature. We will provide an overview of variable selection methods which are based on

a) significance or information criteria, [1; Ch. 2]

b) penalized likelihood, [2]

c) the change-in-estimate criterion, [3]

d) background knowledge, [4] or

e) combinations thereof. [5]

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 tutorial, we will exemplify application of variable selection using scientific questions and data from real medical studies with binary and censored survival endpoints. We will also discuss implications of variable selection, e.g., on uncertainty and stability of the final model [6,7], on bias of regression coefficients [8], and on the validity of confidence intervals [9]. We will give pragmatic recommendations for the practitioner, suggesting typical steps to be done when variable selection is conducted, from selection of candidate covariates, over choosing an appropriate variable selection method to reporting the final model in scientific reports. These recommendations will consider the case of moderately correlated covariates (r<0.8) of mixed type. We will further provide a brief outlook on methodologic extensions to deal with missing values, nonlinear effects and effect modification. Finally, we will provide an overview of software implementations in SAS, R and SPSS.

References:

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

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

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

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

[5] 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

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

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

[8] 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

[9] 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