Variable selection – a review and recommendations for the practicing statistician

Georg Heinze [*,1], **Christine Wallisch**[1], and Daniela Dunkler[1]

[1] Medical University of Vienna, Center for Medical Statistics, Informatics and Intelligent Systems, Section for Clinical Biometrics, Spitalgasse 23, Vienna 1090, Austria

This workshop is mainly based on our recently published review paper on variable selection (Heinze, Wallisch, Dunkler, 2018).

Extended abstract

Statistical models are used in medical research for prediction or for explanation (Shmueli, 2010). In particular, these models allow to prognosticate individualized outcomes conditional on independent variables, or to estimate effects of independent variables, respectively. Theory of statistical modeling is well-established if the set of independent variables to consider is fixed and small. In this case we can assume that effect estimates are unbiased and the usual methods for confidence interval estimation yield valid results. In practice, however, we are often confronted with 10 to 30 candidate variables and we do not know a priori which independent variables to include in a model. Such numbers of variables are often considered too large for a practically useful statistical model. Encouraged by the availability of many different automatic selection procedures in statistical software packages, many researchers believe that the task of variable selection can simply be shifted to data analysis. We will discuss this and other existing myths about variable selection (Heinze and Dunkler, 2017).

Moreover, we will provide an overview of various available variable selection methods which are based on significance or information criteria, penalized likelihood, the change-in-estimate criterion, background knowledge, or combinations thereof. These methods were developed in the context of a linear regression model and then transferred to generalized linear models or models for censored survival data. Variable selection compromises stability of a final model, unbiasedness of regression coefficients, and validity of p-values or confidence intervals. This is particularly relevant in explanatory modeling, where regression coefficients are of central interest. A simulation study illustrates how the reliability of those quantities is associated with sample size.

We will give pragmatic recommendations for the practicing statistician on applying variable selection methods in low-dimensional modeling problems and on performing stability investigations. These recommendations mainly depend on the amount of information provided by the data relative to the number of modeling questions of interest.

Software packages offering automated variable selection procedures usually present results from the finally estimated model ignoring all preceding data-driven selection steps. This has led to the practice of reporting model-based estimates and standard errors from that final model in research articles. We propose some additional quantities that should be routinely reported by software packages (Heinze, Wallisch and Dunkler, 2018). These quantities can be computed for any variable selection method. The first of them, relative conditional bias (RCB), estimates the bias of regression

[*]Corresponding author: e-mail: georg.heinze@meduniwien.ac.at

coefficients that is incurred by variable selection. The second, termed root mean squared difference ratio (RMSD ratio), expresses the increase or decrease in standard errors of regression coefficients caused by variable selection relative to fitting a global model without variable selection. These quantities can be easily interpreted by users of variable selection procedures and they are obtained by resampling techniques that repeat the entire variable selection process in each resample. We evaluated the performance of these quantities in a simulation study. Resampling can also be used to assess model stability by estimating inclusion frequencies of individual variables or of sets of variables (model selection frequencies). In line with De Bin et al (2016), we show that such quantities are better estimated by subsampling than by the bootstrap. However, subsampling severely underestimates RMSD. We propose to extend a summary table of a variable selection procedure by variable inclusion frequencies estimated by subsampling, and the RCB and RMSD ratio estimated by the bootstrap. These quantities should be reported for all variables, not just the selected ones. The global regression coefficients should be added as well.

A still highly debated and only partly solved problem is valid inference after variable selection (Taylor and Tibshirani, 2015). We will also discuss this issue and give some recommendations. Application of our recommendations in practice will be exemplified by means of analysis of some real studies.

References:

Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*. 2018;60:431–449. https://doi.org/10.1002/bimj.201700067

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Heinze, G., & Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, *30*, 6–10.

De Bin, R., Janitza, S., Sauerbrei, W., & Boulesteix, A. L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*, *72*, 272–280.

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 7629–7634.