Prediction and explanation in studies with rare events: problems and solutions

SPEAKER: Georg Heinze, Medical University of Vienna, georg.heinze@meduniwien.ac.at

We discuss problems arising in the analysis of logistic regression models with sparse data sets, considering the case where interest lies in both prediction and effect estimation. The first problem is separation, where perfect separability of the outcomes in a data by the covariates set hampers effect estimation (Mansournia et al, 2018). Another problem is low accuracy despite a high sample size. Several approaches to deal with these problems have been proposed, e.g. methods based on bias reduction such as Firth's logistic regression, or the use of weakly informative priors in a Bayesian framework. Because of many advantageous properties, Firth's logistic regression has become a standard approach for the analysis of binary outcomes with small samples. It is implemented in the R package logistf and in many other statistical software systems. Whereas it reduces the bias in maximum likelihood estimates of coefficients, bias towards one-half is introduced in the predicted probabilities. The stronger the imbalance of the outcome, the more severe is the bias in the predicted probabilities. We propose a simple modification of Firth's logistic regression resulting in unbiased predicted probabilities (Puhr et al, 2017). While this method introduces a little bias in the regression coefficients, this is compensated by a decrease in their mean squared error. We demonstrate the properties of our proposed method in a comparative simulation study, including also other methods, and exemplify its use with real data.

References:

Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences and control. American Journal of Epidemiology 2018, doi:10.1093/aje/kwx299

Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates AND predictions? Statistics in Medicine 2017 Jun 30;36(14):2302-2317, doi:10.1002/sim.7273

ABOUT THE SPEAKER:

Georg Heinze is head of the Section for Clinical Biometrics at the Center for Medical Statistics, Informatics and Intelligent Systems of the Medical University of Vienna since 2015. He received his PhD Degree in Statistics from the University of Vienna in 1998, and was appointed Associate Professor in 2004. Primary research focuses on biostatistical modeling strategies for prediction and estimation of effects of exposures on outcomes, particularly when sample sizes are small or outcome events are rare. Secondary focus is the re-use of health data for medical research, particularly when sample sizes are very large, as in nationwide studies on health insurance claims. For more information please visit the website https://cemsiis.meduniwien.ac.at/en/kb