Multiple Imputation and Model Selection in Cox Regression: Estimating the Risk of Dementia in Elderly

M.Schipper, M.M.B. Breteler, Th. Stijnen
Dept. of Epidemiology & Biostatistics,
Erasmus MC, Rotterdam, the Netherlands

Missing values in the observed predictor variables complicate the derivation of an adequate prognostic time to event model. We suggest an approach with multiple imputation via van Buurens MICE library followed by stepwise model selection. In every step we pool the coefficients (and its variance-covariance matrix) of the Cox proportional hazard models, each based on one of a (small) number of imputed datasets. A generalized Wald-test by Rubin gives suitable diagnostics for the next step in the model selection procedure.

Once we know how to do the model selection on a series of imputed datasets using a proportional hazards approach, we can repeat this procedure any number of times in a bootstrap setting. In this way it is possible to assess the calibration and discrimination abilities of our final model following the approach of Harrell et al. The bootstrap also enables to estimate the shrinkage factors for the final model to get a better calibration.

Because of its flexibility and extent R is an excellently suitable environment to program the whole model selection procedure. Choosing an object-based approach, we can even use default R model selection functions.

In an example we apply this modeling strategy to derive a model that assesses the risk of developing dementia over time. The example is based on The Rotterdam Study.

A population based prospective cohort study in Rotterdam of about 8,000 people of 55 years and above. During 10 years of follow up, over 400 cases of dementia were recorded. Initially there are 39 covariates of interest. Although only 9missing, deletion of incomplete cases leads to a loss of over 70the cohort. Therefore multiple imputation and application of the aforementioned model selection procedure seems an appropriate approach here.

References:

[1] Harrell FE Jr, Lee KL, Mark DB. Tutorial in Biostatistics. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Stat Med 15:361-387, 1996

[2] Rubin DB. Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York, 1987

[3] van Buuren S, Oudshoorn CGM. Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054, 1999