Comparison of discrimination ability estimators in the categorisation of continuous predictors in a Cox Proportional Hazards Model

Irantzu Barrio[1], *María Xosé Rodríguez-Álvarez*[2], *Luís Meira-Machado*[3] *and Inmaculada Arostegui*[1,4]

[1] Departamento de Matemática Aplicada y Estadística e I.O, Universidad del País Vasco UPV/EHU

[2] Departamento de Estadística e Investigación Operativa, Universidade de Vigo

[3] Centre of Mathematics and Department of Mathematics and Applications. University of Minho, Portugal

[4] BCAM- Basque Center for Applied Mathematics

The Cox proportional hazards model is the most common survival prediction model for the analysis of time to event data. In medical research, a common strategy when developing survival models is to categorise continuous covariates. However, agreement on criteria to select cut points does not always exist. Previous work on the estimation of optimal cut points with censored data has been done but with the aim of seeking for a unique cut point (Sima and Gönen, 2013). In this work, we extend the proposal by Sima and Gönen (2013) to the categorisation of continuous predictors in more than only two categories. Let X denote a continuous covariate which we want to categorise. Our proposal consists on categorising X in such a way that the best predictive survival model is obtained, considering the maximal discrimination attained. To measure the discrimination ability of the model, we considered the concordance probability index, and two different estimators were studied: the c-index (Harrell, 1982) and the concordance probability estimator (CPE, Gönen and Heller ,2005). The algorithms used to select the optimal cut points, called *Addfor* and *Genetic* have been presented elsewhere (Barrio et al, 2013). We conducted a simulation to evaluate the empirical performance of both the c-index and CPE estimators when it comes to select the optimal cut points for the categorisation of continuous variables. Simulations were performed for different sample sizes (500 and 1000), number of cut points (1,2 and 3) and censoring rates (20%,50% and 70%). The results showed that in general the CPE works better for low censoring rates (20%), while the c-index performs better for high censoring rates (70%). The method performs successfully when it comes to search two or three cut points. However, when the aim is to search a unique cut point, the method's performance depends largely on the location of the theoretical optimal cut point.

Keywords: categorisation, Cox proportional hazards model, prediction.