Joint model for longitudinal and survival data: an approach in credit risk analysis

Victor Medina-Olivares, Raffaella Calabrese and Jonathan Crook.

The data used in many credit scoring models can be categorized into three groups: (1) time to the event of default, (2) covariates that are time-independent (typically application predictors) and (3) covariates measured repeatedly during the lifetime of the loan (typically behavioural predictors). The interest often dwells in predicting the time to default for each debtor by using all the historical data gathered at a particular moment, thus when new data is collected the prediction can be updated in a dynamical fashion. Methods for dynamically predicting the time to default are extensively described in the literature. One of the approaches commonly used is survival models with time-varying covariates. However, these models can be problematic when the time-varying covariates are internal (endogenous) which is normally the case in credit data. We propose a novel approach in this context, namely joint model for longitudinal and survival data, that allows us to properly account for the association between the repeated measurements and the prediction of the time to default by taking care of endogeneity, measurement errors and non-ignorable dropout mechanisms. As far as we know, this is the first time these models are applied in credit risk analysis. The data used corresponds to the Single Family Loan-Level dataset openly available by Freddie Mac. The prediction performance of this approach is compared with standard methods under different metrics and time-horizon to investigate the circumstances where the joint model can make better use of the historical data.