XGBoost-Based vs. AFT Model Imputation: Addressing Interval Censoring in Time-to-Event Data

Gustavo Soutinho[1,a)] and Luís Meira-Machado[2]

[1]*Research on Economics, Management and Information Technologies (REMIT) - Portucalense University, Rua Dr. António Bernardino de Almeida, 541 4200-072 Porto, Portugal*
[2]*Centre of Mathematics, University of Minho, Campus de Azurém, 4800 - 058 Guimarães, Portugal*

[a)]gustavo.soutinho@upt.pt

Abstract. Interval-censored data pose significant challenges in survival analysis, as the exact timing of events is unknown and only known to fall within observed intervals. This study explores imputation-based strategies for regression modeling under interval censoring, including traditional midpoint and Accelerated Failure Time (AFT) model imputations, as well as a machine learning-based approach using XGBoost. We further introduce the Scaled Linear Redistribution Method, a novel rescaling mechanism that adjusts model-based imputations to respect censoring intervals while preserving their relative variability. Using real clinical data, we illustrate how these methods influence the estimation of survival curves. Since true event times are not observed, direct evaluation of the accuracy of imputed times is not possible. Instead, we assess the resulting survival estimates by comparing them with the Turnbull estimator, a nonparametric method that fully accounts for interval censoring without requiring imputation. The analysis demonstrates that midpoint, AFT, and XGBoost-based imputations yield survival curves that are broadly consistent with the Turnbull curve in this dataset.