PyDTS: A Python Package for Discrete-Time Survival Analysis with Competing Risks

Tomer Meir[1], Rom Gutman[1*] and Malka Gorfine[2]*
[1] Technion - Israel Institute of Technology
[2] Department of Statistics and Operations Research, Tel Aviv University
[*] These authors contributed equally
[*]Corresponding Authors: tomer1812@gmail.com, rom.gutman1@gmail.com

Background.
Time-to-event analysis (survival analysis) is used when the outcome or the response of interest is the time until a pre-specified event occurs. Sometimes, events can only occur at regular, discrete points in time. In other situations, events may occur at any point in time, but available data record only the particular interval of time in which each event occurs. This type of analysis is very common when analyzing healthcare data, for example, death from cancer measured by months since time of diagnosis, or length of stay in hospital recorded on a daily basis. Competing events arise when individuals are susceptible to several types of events but can experience at most one event. For example, competing risks for hospital length of stay are discharge and in-hospital death. This work focuses on discrete-time regression with competing events. To the best of our knowledge, currently, there is no publicly available Python implementation of any regression model of discrete-time survival data with competing events.

Methods.
We adopt the collapsing approach of Lee et al. (2018), but provide a new estimation technique that saves substantial computation time with no substantial loss of accuracy or efficiency, as we demonstrate by simulations. Additionally, we present PyDTS, an open-source Python package which implements the estimation methods of Lee et al. (2018) and ours.

Results.
PyDTS is an open-source Python package which implements tools for discrete-time survival analysis with competing risks. The code is available at https://github.com/tomer1812/pydts under GNU GPLv3. Documentation, with details about main functionalities, API, installation and usage examples, is available at https://tomer1812.github.io/pydts/. The package was developed following best practices of Python programming, automated testing was added for stability, Github Issues can be opened by any user for maintainability and open source community contributions are available for ongoing improvement. The package is published using the Python Package Index (PyPI), can be installed (using the "pip" installer) and used with a few simple lines of code, as shown in the Quick Start section of the documentation. Yet, it is flexible enough to support more advanced pre-processing, regularization and other fitting options, as demonstrated in the Hospitalization Length of Stay and in the Estimation with TwoStagesFitter examples of the documentation. As shown in the Comparing the Estimation Methods section of the documentation, both estimation methods provide highly similar results in terms of point estimators and their standard errors. However, the computational running time of our proposed approach is 1.5-3.5 times shorter depending on the number of discrete time points (d), with more significant performance improvement for higher values of d.

Conclusion.
Discrete-time survival analysis with competing risks is required in many practical cases when analyzing healthcare data. This work provides a new estimation procedure that could be highly useful in large datasets with a large number of covariates. We provided a simple to use Python package which implements the estimation procedure of Lee et al. (2018) and our proposed faster estimation method. We expect PyDTS to be adopted by Python users, and specifically by healthcare data-scientists, and to be further improved by comments and contributions from the open-source Python community.