Master thesis proposal
Effects of missing values on survival time prediction

Krzysztof Bartoszek

November 11, 2016

Background — models of cancer evolution, survival analysis

Progressive diseases like cancer are today modelled by so called compartmental models [4, 7], see Fig. . The times between the compartments are random and depend on the patient's characteristics. A medical doctor is often interested in predicting the survival time (i.e. time till death) of a patient given that the patient is observed in a certain compartment with given clinical or genetic characteristics. This prediction of this survival time is based on databases where patients' survival times and characteristics have been collected. There are many possible ways to predict the survival time, e.g. regression approaches, artificial neural networks [ANNs, 3, 5]. However, these databases often have many missing values, making any analysis and prediction difficult.

Thesis project

The project is to explore and compare the predictive power of ANNs and regression methods when data is simulated under a compartmental model. The exploration is to be done with various probabilities of missingness and ways of dealing with missingness. The model of disease will be a simple one: maximum 4 stages before death, exponential (with rates dependent

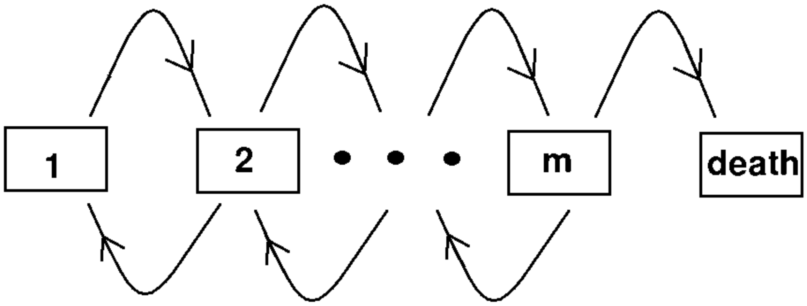

Figure 1: A compartmental model of disease.

on parameters) times between stages, no improvement in health. One can consider prediction of time till death (continuous response) or whether a patient survives longer than a certain period (binary response). ANNs handle both types of responses, for the first one a usual regression can be tried for the second e.g. a logistic regression.

Goals

The below general goals are for an “ideal” thesis. Depending on the student they will be made more specific in the direction of the student’s interests. In particular the focus of the work will not be on the mathematical models (these will be “provided”) but on implementing and putting together software to do simulations, inference and explore the statistical aspects of the models.

1. Become acquainted with compartmental models of diseases and be able to simulate them [4, 7].

2. Become acquainted with missing value models [1].

3. Become acquainted with survival time estimation for compartmental models of diseases [2].

4. Explore the effects of missing values.

Data

The topic will be predominantly illustrated with simulated data. Furthermore, the Pomeranian breast cancer dataset will be available [6].

References

[1] P. D. Allison. *Missing Data.* Sage Publications, 2001.

[2] K. Bartoszek, M. Krzemiński, and J. Skokowski. Survival time prognosis under a Markov model of cancer development. In *Proceedings of the XVI National Conference on Applications of Mathematics in Biology and Medicine*, pages 6–11, 2010.

[3] V. Bourdès, S. Bonnevay, P. Lisboa, M. Aung, S. Chabaud, T. Bachelot, D. Perol, and S. Negrier. Breast cancer predictions by neural network analysis: a comparison with logistic regression. In *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, pages 5424–5426, 2006.

[4] D. Faissol, P. Griffin, and J. Swann. Bias in markov models of disease. *Mathematical Biosciences*, 220:143–156, 2009. Only for model description.

[5] Ripley R. *Neural Network Models for Breast Cancer Prognosis.* PhD thesis, Oxford University, Oxford, 1998.

[6] J. Skokowski. *Wartości rokownicze wybranych czynników klinicznych i patomorfologicznych w raku piersi (Predictive value of certain clinical and patomorphological parameters in breast cancer).* PhD thesis, Medical University of Gdańsk, Gdańsk, 2001.

[7] A. Yashin, I. Akushevich, K. Arbeev, L. Akushevich, A. Kulminski, and S. Ukraintseva. Studying health histories of cancer: A new model connecting cancer incidence and survival. *Mathematical Biosciences*, 218: 88–97, 2009. Only for model description.