Bayesian Estimation of parameters for Survival models using the Cox Proportional model

Daniela Pico Julian Usuga Deivid Zhang

2022-06-19

Abstract

The primary aim of our study was to predict the mortality of the people from African countries, specifically, using data from DHS and functions from package *mortDHS* we were able to create functions for any given country inside the set of DHS surveys, generate a model which can be used to make survival predictions and additional to that, the effects associated with the survival of people are compared, measurement that cannot be done in models with lack of parameters, such as the non-parametric Kaplan-Meier estimator, specifically, we can make a discrimination between the differences in survival depending on the sex and the country to which each person belongs. In order to have Bayesian properties in our models, we used the package *rstanarm*, which uses the Stan software to make his estimations.

Introduction

The problem

In our study we were interested in estimating the mortality on African countries. These countries sometimes lack of a good and reliable registry of deaths and without these estimating mortality rates can be difficult, these rates have prime importance on epidemiological or socio-economic studies.

In our study we used the survey data provided by DHS, these contain a large number of variables, people were asked about the survival status of their siblings so we only include sibling data on our study, we assess siblings survival, expecting to derivate from these the mortality of the population.

Study Population

Many countries don't have good registries management on government institutions and thus they don't have a good record of important events such as population deaths, the Demographic and Health Surveys (DHS) facilitates multiple datasets containing data collected from questionnaires performed in households from a large list of countries.

In our study we will consider the Individual Women's data, which consist of questionnaires performed to women. They were asked about wherever they had siblings, their survival status, age and death date (in case they have died).

Data source

The primary data source was the Demographic and Health Surveys (DHS) Individual Recode (IR) where each row consist of a woman and their responses on multiple questions. There were multiple columns but we selected the ones related to siblings and their survival status, these where:

- Sibling sex (male or female)
- Sibling date of birth
- Sibling survival status (0 = dead, 1 = alive)
- Sibling date of death

Basic concepts

Survival analysis

Survival analysis is a collection of statistical methods for study the time that elapses until an event occurs [1]. The name survival is due to the fact that the applications of this method are mainly study the times of death, some fields of the application of survival analysis They include: medicine, epidemiology, economics, among others. An advantage of models based on survival analysis is that they allow work with censored data, censoring occurs when you have some information on the survival time of a patient but the exact time to failure is not known.

There are three functions of interest for survival analysis, the survival function denoted $S(t)$, the hazard function denoted $h(t)$, and the hazard function accumulated denoted by $H(t)$, these will be the amounts of interest for our study. According to the literature [2], functions are defined as follows:

Let T be a random variable denoting the survival time of a unit or person. The survival function, which is denoted $S(t)$ gives the probability that a unit or person will survives beyond some specific time t. i.e,

$$S(t) = P(T > t)$$

The survival function $S(t)$ is fundamental in AS since for different values of t provide crucial information of survival data In some situations it may be more interest to quantify the risk of failure at a given instant than to estimate survival; a function of interest in survival analysis that allows do this is the danger function

The hazard function usually denoted as $h(t)$ or $\lambda(t)$ is given by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T + \Delta t | T \geq t)}{\Delta t}$$

The numerator represents the conditional probability of the event occurring in an infinitesimal interval $[t, t{+}\Delta t]$ (as $\Delta t \to 0$) given that the unit has survived to t ($T > t$). survived until t (T>t).

The cumulative hazard is defined as:

$$H(t) = \int_0^t h(s)ds$$

Where $h(s)$ is the hazard function. exist a one to-one relationship between the hazard function, the cumulative hazard and the probability of survival, as follows

$$S(t) = exp(-H(t)) = exp(-\int_0^t h(s)ds)$$

Censorship

An observation in a random variable T is said to be right-censored if all that is known about of T is that it is greater than some value c. In AS, T refers at the time of occurrence of some particular event and a case is considered right-censored if it stops observing before for the event to occur[3]. In our work, only the right caesura is presented, since at the time of conducting the survey there were people who had not yet experienced the event of interest (death).

Survival probability can be estimated nonparametrically over temporal observations (censored and uncensored) using the Kaplan-Meyer method. other two Common approaches to modeling survival data consist of modeling the instantaneous rate of the event as a function of time. This includes the class of models known as proportional hazards regression models and non-proportional hazards regression models; the second is to model the time of the event itself. This includes the class of models known as accelerated time to failure (AFT) models.

Bayesian approach and why Bayesian

The main reason to go Bayesian in our study was to be able to make inference, in particular, to make easy to interpretative conclusions on our parameters and their significance, without the assumptions of frecuentists approaches.

Models

Under this modeling framework, it is proposed to implement a of Bayesian Cox proportional hazard to later make a comparison versus the frequentest approach and the kaplan Meier curve.

Kaplan–Meier estimator

The Kaplan-Meier estimator, also known as the limit product estimator, is a nonparametric method for estimating the survival function. survival function by maximizing the sample likelihood function. Suppose one has k different failure times $t_1 < t_2 < ... < t_k$, in each time $t_j (j = 1, 2, ...k)$ there are n_j subjects that are under observation and at risk of an event of interest.

The K-M estimator is defined as

$$\hat{S}_{KM}(t) = \prod_{j:t_j \leq t} [1 - \frac{d_j}{n_j}]$$

$$para \quad t_1 \leq t \leq t_k \quad y \quad d_j = \#faults$$

Cox proportional hazards model

Under a hazard scale formulation, we model the hazard of the event for individual i at time t using the regression model:

$$h(t|X_i) = h_0(t) exp(X_i \beta)$$

Where h_0 is called the hazard base, that is, $h0(t)$ is the risk when all X_i variables are 0. $h_0(t)$ characterizes the way in which hazard changes as a function of survival time, while the second term characterizes the way the hazard changes as function of the covariates at the same time guarantees that this is positive, it is also called the linear predictor.

The formulation of the Cox model in terms of the hazard and the survival function are given by: From (2) we note that

$$S(t;x) = exp(-H(t;x))$$

Where $H(t;\vec{x})$ is the cumulative hazard for a subject with covariates $\vec{x} = (x_1, x_2, ..x_k)$ Assuming survival time is continuous

$$H(t;\vec{x}) = \int_0^t h(s;\vec{x}) ds = \int_0^t h_0(s) exp(\beta^T \vec{x}) ds$$

$$= exp(\beta^T \vec{x}) \int_0^t h_0(s) ds = exp(\beta^T \vec{x}) H_0(T)$$

Cox model in terms of cumulative hazard. In this expression $H_0(t)$ is the cumulative baseline hazard. This relationship can be thought of as a baseline cumulative risk measure which is modified according to the function

From the above relationship, the Cox model can be formulated in terms of survival:

$$S(t;\vec{x}) = exp(-H(t;\vec{x})) = exp(e^{\beta^T \vec{x}}) H_0(t) =$$

$$[exp(-H_0(t)]^{e^{\beta^T \vec{x}}} = [S_0(t)]^{e^{\beta^T \vec{x}}}$$

Cox model in terms of survival. In this expression $S_0(t)$ is the survival baseline.

Estimation

In our study we used an M-Spline model approximation for the hazard baseline h_0, this approximation uses for the γ_l coefficients, the package rstanarm handles this using a Dirichlet prior with concentration parameter of 1.

So the hazard of dying for the individual i at time t is given by:

$$h_i(t) = \sum_{l=1}^{L} \gamma_l \mathrm{M}_l(t \mid k, \delta) * \exp(\beta \cdot X_i)$$

Where $k = \{k_1, ..., k_J\}$ is a set of knots given by the user, we observed that a decent number of knots is around 6 (including the 2 boundary knots, k_1 located at the earliest entry time and k_J located at the lastest event), $\beta = (\beta_0, \beta_1, ..., \beta_n)$ is the vector of parameters to estimate in our linear model and $X_i = (1, x_1, ..., x_n)$ is the vector of covariables. We will leave the default degree of the splines δ at 3.

References

[1] Allison, PD. (1995). Survival Analysis Using the SAS System: A practical guide, Cary, NC:SAS Institute Inc., 292 pp.

[2] COX, D. R. Regression Models and Life Tables (with Discussion), Journal of The Royal Statistical Society, Series B, 34, 187-220, 1972.

[3] ALLISON, P.D. Discrete-Time Methods for the Analysis of Event Histories, In Sociological Methodology 1982, ed. S. Leinhardt, San Francisco, CA: Jossey-Bass, 1982.