Modelo bayesiano basado en splines para predecir la mortalidad en algunos países de África a partir de información de la Demographic and Health Surveys (DHS)

Daniela Pico Arredondo, Julián Úsuga Ortiz, Deivid Zhang Figueroa, Johnatan Cardona Jiménez, Juan Carlos Salazar Uribe, Kene Nwosu
Universidad Nacional de Colombia Sede Medellín, Institución Universitaria Pascual Bravo, University of Geneva

Abstract

In epidemiology, it is important to estimate mortality rates and this frequently depends on the quality of the available registry. For instance, in some African countries, there are no clear or easily accessible mortality records, which is why information from relatives can be used to find out, among other things, dates of death and causes of death. The main objective of this work is to predict the mortality of people in some countries of the African continent. Using data from the "Demographic and Health Surveys" (DHS) [6] and the functions of the mortDHS and rstanarm packages (based on MCMC), a statistical model, based on splines, is formulated to obtain smoothed survival curves corrected for some covariates of interest. With this model, specifically with the posterior distribution, it is also possible to evaluate the size of the effect of these covariates on survival time and obtain credibility intervals. Specifically, a differentiation is made based on the probabilities of survival, according to gender and the country to which they belong, and the riskiest mortality profile is identified.

Objectives

In our study we were interested in estimating the mortality of people 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.

Methodology

The notable steps on the making of this study were:

- Requesting data from the DHS program.
- Transforming and cleaning the data to a sibling-per-row format.
- Model training , selection and verification of the distributions resulting from the simulation carried out.
- Analysis of the obtained results.

Analysis and modeling

Survival analysis is a collection of statistical methods to study the time that elapses until an event occurs. The name survival is due to the fact that the applications of this method are mainly study the times of death, some of the application fields of the survival analysis are 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 unknown.

$$\begin{aligned}\lambda(t|X_i) &= \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) \\ &= \lambda_0(t)\exp(X_i \cdot \beta)\end{aligned} \tag{1}$$

is known as the Cox proportional-hazards model. Where X_i is the vector of p covariates for the person i, β is the vector of coefficients associated with each covariate and $\lambda_0(t)$ represents the hazard base function, which describes how the risk of death changes over time where all the covariates are equal to 0 (only depends on the time t) [3].
As we can see the model (1) can be divided on two parts. The first one is a function and only depends on the time t. The second one only is a factor and only depends on the effects of the covariates X_i.
Cox proposed a method to estimate the β coefficients known as partial likelihood [2]

Bayesian Approach

The main reason to go bayesian in our survival analysis study was the need to make inference, in particular, a vast majority of models used in survival analysis are non-parametric, so inference is not straightforward, hence we added the parametric component in those models and with bayesian estimators we can get easily interpretable conclusions on our parameters and test their significance, without the assumptions of frecuentists approaches.
As stated earlier, our study consists on comparing the mortality between sex and country, so the vector of covariates is defined as follow:

$$\begin{aligned}X_i = &(1, I_{\text{sex}=\text{F}}, I_{\text{country}=\text{GMB}}, I_{\text{country}=\text{RWA}}, \\ &I_{\text{country}=\text{BEN}}, I_{\text{country}=\text{SLE}}, I_{\text{country}=\text{MLI}} \\ &I_{\text{country}=\text{LBR}})\end{aligned} \tag{2}$$

As we can see, the indicators of the sex Man and the country Benin are missing, the reason is that those are the reference levels, and their effect is represented by the *Intercept* level β_0.
In our Bayesian estimation we set different priors for the Cox proportional-hazards model coefficients and estimate them using MCMC, in particular, the hazard base function is estimated using a set of M-Splines with γ_l coefficients and the linear model is estimated using normal priors[1], as follow:

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

$$\beta_0 \sim N(0, 20)$$

$$\beta_i \sim N(0, 2.5) \text{ with } i = 1, \ldots, 7$$

$$\gamma_l \sim Dirichlet(1)$$

The package **rstanarm**[1] handles the calculation and sets a Dirichlet prior with concentration parameter of 1, ensuring an non-informative prior.
The parameters k and δ represent the number of knots of the M-spline and the degree of these splines, respectively. We set the parameter k to 4 and δ to 3, these parameters prevent our model to overfitting.

Results

After modeling, we checked for convergence, visually the MCMC procedure converges very well in each of the 4 chains initialized at a diverse set of starting points placed randomly.

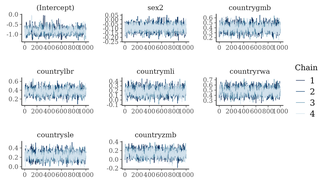

Figure 1:Trace plots for the coefficients Intercept, Sex and Country

The *Figure 1.* can evidence that chains converge to the same target coefficient; also each chains have attained stationarity.[4]

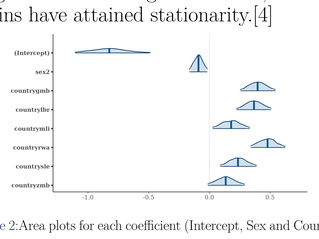

Figure 2:Area plots for each coefficient (Intercept, Sex and Country), with 90% and 95% probability intervals.

From *Figure 2.* we can see the effects that covariates have in mortality, from these we see that the second level of the factor (Sex = Woman) reduces mortality, compared with the Intercept level. Secondly, being a person from Rwanda increases the mortality when it is compared with being from the country Benin (Intercept)[4].

Results

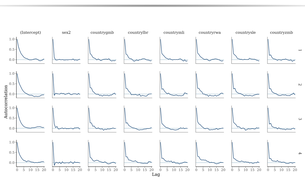

Figure 3:Autocorrelation plots (ACF) for the coefficients Intercept, Sex and Country

It is important that the model doesn't show signs of correlation between chains, the training process takes care of this problem doing a *warm-up* training, in the *Figure 3* we can see that the correlation for each chain and coefficient is around 0, this is a good sign of convergence.

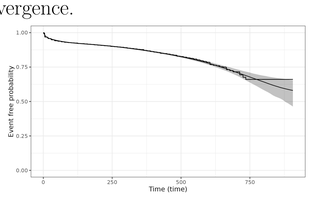

Figure 4:The non-parametric Kaplan-Meier estimator compared with the estimated M-spline model

The *Figure 4* shows the popular non-parametric Kaplan-Meier model, and on top of it our Bayesian model based on M-Splines. We can visually conclude that our model is not so distant from the non-parametric model.

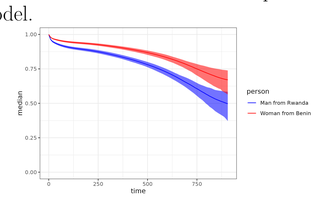

Figure 5:Survival function for two groups, Women from Benin and Men from Rwanda, with a blue line representing the median probability and a 95% interval.

With our model trained and with evidence that our model coefficients have converged we can use it to predict, in *Figure 5* we plot the survival function from man from Rwanda (blue) and woman from Benin (red).
If we are interested in the mortality at an exact time, we fix our time t, in the table we fix t to 219 months (18.25 years) and get the probabilities.

Time	Median	L	U	Person
219	0.9307	0.923	0.938	Woman from Benin
219	0.8820	0.871	0.893	Man from Rwanda

Table 1: Model evaluated at $t = 219$ (18.25 years)

In Table 1 we see the probabilities of surviving past 219 months, or more generally, the probabilities that the death has not yet occurred by 219 months or 18.25 years for each one of the two groups. L and U represent the lower and upper bound of the 95% probability interval [5].

Conclusions

In this study we created a model based on Bayesian statistics to estimate mortality, obtaining point estimations and probability intervals.

References

[1] Samuel L Brilleman et al. "Bayesian survival analysis using the rstanarm R package". In: *arXiv preprint arXiv:2002.09633* (2020).
[2] David R Cox. "Partial likelihood". In: *Biometrika* 62.2 (1975), pp. 269–276.
[3] David R Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.
[4] Andrew Gelman et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
[5] Brandon George, Samantha Seals, and Inmaculada Aban. "Survival analysis and regression models". In: *Journal of nuclear cardiology* 21.4 (2014), pp. 686–694.
[6] DHS Program. *Demographic and health surveys (DHS)*. 2021.