The Journal of Open Source Software

BayesMFSurv: An R Package to Estimate Bayesian Split-Population Survival Models With (and Without) Misclassified Failure Events

Minnie M. Joo[1], Nicolás Schmidt[2], Sergio Béjar[3], Vineeta Yadav[4], and Bumba Mukherjee[4]

1 Dept. of Political Science, University of Massachusetts Lowell **2** Dept. of Political Science. Universidad de la Republica, UY **3** Dept. of Political Science, San Jose State University **4** Dept. of Political Science, Pennsylvania State University

DOI: 10.21105/joss.02164

Software

- Review
- Repository
- Archive

Editor: Marcos Vital

Reviewers:

- @alletsee
- @andybega

Submitted: 04 February 2020
Published: 30 March 2020

Summary

Social Scientists and Biostatisticians often employ conventional parametric survival or mixture cure models (e.g. Weibull, Exponential) to analyze outcome variables in survival data that focus on the time until an event occurred or "failed" (Box-Steffensmeier & Zorn, 1999; Lee, Chakraborty, & Sun, 2017; Maller & Zhou, 1996). An important assumption underlying these models is that researchers record the date -year, month or day- in which an event or observation of interest failed (i.e. "terminated") *accurately*. Yet events that are recorded as having failed at a given point in time can be inaccurately measured (Bagozzi, Joo, Kim, & Mukherjee, 2019; Clark, Bradburn, Love, & Altman, 2003; Schober & Vetter, 2018). Inaccurate measurement of this sort leads to a subset of *misclassified failure* cases in survival data in which some subjects are recorded as having failed or experienced the event of interest even though they in actuality "live on" past their recorded-failure point.

There are several scenarios where a subset of recorded failure events may persist beyond their recorded failure time, leading to misclassification in event failures. For example, political scientists who analyze the duration of civil wars fought between rebel groups and governments often record end dates ("failures") for specific conflicts based upon 24-month spells with fewer than 25 battle-deaths per year (Balch-Lindsay & Enterline, 2000; Thyne, 2012). The aforementioned threshold is prone to measurement error, especially for lower-intensity civil wars in poor information environments that persist beyond their recorded end date. Other examples include the study of the duration of ancient civilizations (Cioffi-Revilla & Landman, 1999) and the time taken to detect cancer (Schober & Vetter, 2018). In both these latter examples, researchers typically do not have data on the precise time-point of a given failure due to the sands of time or because of inaccurate information. This leads them to -similar to the civil conflict example- underestimate the duration of particularly misclassified event failure cases.

Since these underestimates of duration are non-random, bias will arise in survival estimates of the phenomena mentioned above when using conventional survival models. Hence, the main motivation for developing the Bayesian Misclassified Failure (MF hereafter) split population survival model is to resolve methodological challenges resulting from misclassified event failures by accounting for the possibility that some failure events survive beyond their recorded failure time. The development of the Bayesian MF model is also driven by the fact that it permits researchers to identify when the end date of observations in survival data is misclassified therein providing substantive insights into this process. Further, there does not exist an R package that extracts posterior distribution of estimates from parametric cure (split-population) models,

including the MF model, using Bayesian Markov Chain Monte Carlo (MCMC) methods. `BayesMFSurv` (Joo, Bejar, Schmidt, & Mukherjee, 2019) is an R package (R Core Team, 2019) that contains functions and computationally intensive routines in `C++` to fit the parametric Weibull and Exponential (i) survival model and (ii) Misclassified Failure survival model via Bayesian MCMC methods using slice-sampling (Bagozzi et al., 2019; Neal, 2003).

Motivation, Description, Applications

Numerous R packages offer functionalities to estimate conventional parametric and semi-parametric survival models via maximum likelihood estimation (MLE) or Bayesian MCMC methods (Diez, 2013; Therneau, 2019; Wang, Chen, Wang, & Yan, 2019; Zhou, Hanson, & Zhang, 2020). Other R packages focus on estimation of parametric or semi-parametric cure survival models using MLE (Amdahl, 2019; Beger, Hill, Metternich, Minhas, & Ward, 2017; Cai, Zou, Peng, & Zhang, 2012; Han, Zhang, & Shao, 2017). To our knowledge, there does not exist an R package that fits parametric mixture cure models, including the MF survival model, via Bayesian MCMC (e.g. slice sampling) methods which offer a powerful yet flexible tool for estimating such models. Further, extant R packages that use Bayesian inference for survival analyses only focus on standard survival models that do not take into account latent misclassified failure events in survival data. Because misclassified failure events are right-censored events, there is a non-zero probability that these misclassified cases persisted beyond their recorded failure time. Failing to account for misclassified failure events in survival data that results from estimating standard survival or cure models will lead researchers to underestimate the duration of time of these events.

Since the underestimates of duration are non-random, bias will arise in survival estimates of these phenomena when researchers use standard survival or cure models. To address this misclassified failure challenge in survival data, our `BayesMFSurv` R package incorporates various functions listed below that fit Bagozzi et al. (2019)'s parametric MF survival model via Bayesian MCMC methods. This model estimates a system of two equations to account for the possibility that some unknown subset of failure events actually "lived on" beyond their recorded failure time. The first is a "splitting" equation that estimates the probability of a case being a misclassified failure, with or without covariates. The second equation is a standard parametric survival model, whose relevant failure and survival probabilities are estimated conditional on a case *not* being a misclassified failure. These features of the model in `BayesMFSurv` account for a heterogeneous mixture of failure cases in survival data and address the non-random underestimates of duration for misclassified failure events. `BayesMFSurv` also incorporates time-varying covariates that are common in panel survival datasets. This model can be applied at least to the following survival datasets where misclassified failure cases are prevalent: civil war termination that determines civil war duration (Thyne, 2012), time taken to detect onset of cancer (Schober & Vetter, 2018), and collapse (and thus duration) of ancient civilizations or political regimes (Cioffi-Revilla & Landman, 1999; Reenock, Bernhard, & Sobek, 2007).

BayesMFSurv R Package

The R package `BayesMFSurv` contains four functions to fit the parametric (Weibull and Exponential) (i) standard survival model and (ii) MF survival model via Bayesian MCMC using a slice-sampling algorithm described in Bagozzi et al. (2019). Bayesian MCMC estimation is conducted by using the Multivariate Normal prior for these models' split and survival stage parameters, and the Gamma prior for the shape parameter. The functions in `BayesMFSurv` are:

- `mfsurv`: Fits a parametric MF model via Bayesian MCMC with slice-sampling to estimate the misclassification failure probability in the split (first) stage and hazard in the second (survival) stage. Slice-sampling, which is conducted by using the univariate slice sampler (Neal, 2003), is employed to draw the posterior sample of the model's split and survival stage parameters.
- `mcmcsurv`: Fits a standard parametric survival model via Bayesian MCMC with slice-sampling employed to draw the posterior sample of the model's survival stage parameters.
- `stats`: Calculates log-likelihood and deviance information criterion (DIC) for fitted model objects of class `mfsurv` *and* `mcmcsurv`.
- `summary`: Summarizes Bayesian MCMC output -this includes the mean, standard deviation, standard error of the mean, and 95% credible interval- of each parameter's posterior distribution from the Bayesian standard and MF parametric survival models.

The ease and speed of estimating the Bayesian standard and MF parametric survival models in `BayesMFSurv` is improved by using `C++` to perform computationally intensive routines (e.g. slice-sampling) before pulling the output into R. Users can also illustrate trace-plots and kernel density plots for each parameter from `mcmcsurv` and `mfsurv` that fits the Bayesian standard and MF parametric models respectively. To illustrate the `BayesMFSurv` package's functionality, all the functions listed above have been tested on a survival dataset of democratic regime failure (Reenock et al., 2007) described and included in this package.

Availability

`BayesMFSurv` is an open source software made available under the MIT license. It can be installed from its github repository using the `remotes` package: `remotes::install_github("Nicolas-Schmidt/BayesMFSurv")`.

References

Amdahl, J. (2019). *Flexsurvcure: Flexible parametric cure models*. Retrieved from https://CRAN.R-project.org/package=flexsurvcure

Bagozzi, B. E., Joo, M. M., Kim, B., & Mukherjee, B. (2019). A bayesian split population survival model for duration data with misclassified failure events. *Political Analysis*, *27*(4), 415–434. doi:10.1017/pan.2019.6

Balch-Lindsay, D., & Enterline, A. J. (2000). Killing time: The world politics of civil war duration, 1820–1992. *International Studies Quarterly*, *44*(4), 615–642. doi:10.1111/0020-8833.00174

Beger, A., Hill, D. W., Metternich, N. W., Minhas, S., & Ward, M. D. (2017). Splitting it up: The spduration split-population duration regression package for time-varying covariates. *The R Journal*, *9*(2), 474–486. doi:10.32614/RJ-2017-056

Box-Steffensmeier, J. M., & Zorn, C. (1999). Modeling heterogeneity in duration models. In *Summer meeting of the political methodology society, july 15-17*.

Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2012). *Smcure: Fit semiparametric mixture cure models*. Retrieved from https://CRAN.R-project.org/package=smcure

Cioffi-Revilla, C., & Landman, T. (1999). Evolution of maya polities in the ancient mesoamerican system. *International Studies Quarterly*, *43*(4), 559–598. doi:10.1111/0020-8833.00137

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part i: Basic concepts and first analyses. *British Journal of Cancer*, *89*(2), 232–238. doi:10.1038/sj.bjc.6601118

Diez, D. M. (2013). *Survival analysis supplement to openintro guide*. Retrieved from https://CRAN.R-project.org/package=OIsurv

Han, X., Zhang, Y., & Shao, Y. (2017). *Rcure: Robust cure models for survival analysis*. Retrieved from https://CRAN.R-project.org/package=rcure

Joo, M. M., Bejar, S., Schmidt, N., & Mukherjee, B. (2019). *BayesMFSurv: Bayesian misclassified-failure survival model*. Retrieved from https://CRAN.R-project.org/package=BayesMFSurv

Lee, K. H., Chakraborty, S., & Sun, J. (2017). *PsbcGroup: Penalized parametric and semi-parametric bayesian survival models with shrinkage and grouping priors*. Retrieved from https://CRAN.R-project.org/package=psbcGroup

Maller, R. A., & Zhou, X. (1996). *Survival analysis with long-term survivors*. John Wiley & Sons. doi:10.1080/00401706.1998.10485509

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 705–741. doi:10.1214/aos/1056562461

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reenock, C., Bernhard, M., & Sobek, D. (2007). Regressive socioeconomic distribution and democratic survival. *International Studies Quarterly*, *51*(3), 677–699. doi:10.1111/j.1468-2478.2007.00469.x

Schober, P., & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia and Analgesia*, *127*(3), 792. doi:10.1213/ANE.0000000000003653

Therneau, T. M. (2019). *A package for survival analysis in s*. Retrieved from https://CRAN.R-project.org/package=survival

Thyne, C. L. (2012). Information, commitment, and intra-war bargaining: The effect of governmental constraints on civil war duration. *International Studies Quarterly*, *56*(2), 307–321. doi:10.1111/j.1468-2478.2012.00719.x

Wang, X., Chen, M.-H., Wang, W., & Yan, J. (2019). *dynsurv: Dynamic models for survival data*. Retrieved from https://CRAN.R-project.org/package=dynsurv

Zhou, H., Hanson, T., & Zhang, J. (2020). spBayesSurv: Fitting bayesian spatial survival models using R. *Journal of Statistical Software*, *92*(9), 1–33. doi:10.18637/jss.v092.i09