TEST
https://doi.org/10.1007/s11749-023-00862-1

DISCUSSION

Comments on: Nonparametric estimation in mixture cure models with covariates

César Sánchez-Sellero[1] · **Wenceslao González-Manteiga**[1]

Received: 4 April 2023 / Accepted: 10 April 2023

Mathematics Subject Classification 62E05 · 62E08 · 62N01 · 62N02

We are very grateful to the editors of TEST for the opportunity to discuss on this interesting invited paper. Our congratulations to the authors for a nice revision on the topic of nonparametric estimation in mixture cure models with covariates and for a relevant contribution in the case of different covariates in the latency and incidence parts.

The main principle that guides this paper is to avoid parametric assumptions. In particular, the logistic model that is usually assumed in the literature for the effect of covariates on the incidence part is not required here. Then, the probability of cure as a function of covariates is estimated by the conditional survival at the largest observation, where the Beran estimator is used to estimate the survival conditionally to the covariates. This is a natural solution to estimate the incidence in a nonparametric framework, but it has the weakness of depending a lot on the largest observations, particularly when the probability of cure is small and there is high censoring. Note also that the effective sample size, defined as the "number of neighbours" used in the Beran local estimation, could be small, making the estimations more unstable. This could be the reason why a parametric assumption on the incidence part, usually a logistic model, is so common in the literature of mixture cure models. Anyway, the problems of model misspecification are good reasons to consider nonparametric estimators, as the authors claim.

This comment refers to the invited paper available at https://doi.org/10.1007/s11749-022-00840-z.

✉ Wenceslao González-Manteiga
wenceslao.gonzalez@usc.es

César Sánchez-Sellero
cesar.sanchez@usc.es

[1] Research group MODESTYA, CITMAga, Department of Statistics, Mathematical Analysis and Optimization, Facultad de Matemáticas, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain

Published online: 17 May 2023

In order to estimate the latency part, it is considered whether the latency and incidence parts depend on the same covariate or not. If yes, two estimators are reviewed. One of them, due to López Cheda et al. (2017), is obtained by just reweighting the Beran estimator using its value at the largest observation, which was precisely the estimation of the probability of cure. Alternatively, Patilea and Van Keilegom (2020) proposed an estimation of the Beran form with an adjustment of the set at risk to take into account the possibly cured observations.

A new latency estimator is given for the situation with different covariates in the latency and incidence parts. It is obtained by an EM algorithm, which contains a formula to adjust the set at risk in the Beran estimator using the expected value of the indicator of cure for censored observations.

A simplified version of this estimator is proposed where the probability of cure is estimated unconditionally. This is actually a simplification of the estimator, but we have doubts about the consistency of the resulting estimator, particularly when the two covariates, X and Z in the paper, are not independent.

An interesting simulation study is given to compare several estimators, including the new method and its simplified version. In general, the estimator by López Cheda et al. (2017), denoted by NPSXX in the paper, shows a worse performance than the other estimators, particularly than that of Patilea and Van Keilegom (2020) (PVK estimator) and the new proposal (NPSXZ estimator). Note that the NPSXZ estimator is introduced with the purpose of estimating a model with different covariates in the latency and incidence parts, but it can also be used when both covariates are the same. The authors suggest that the better performance of NPSXZ compared to NPSXX can be due to the flexibility of choosing different bandwidths for the incidence and latency parts. Although this could have an effect, we rather think that adjusting the set at risk is more efficient to estimate the latency part than reweighting the Beran estimator by its value at the largest observation. This could also be a reason for the relatively good behaviour of the PVK estimator.

Now we observe that any of these estimators of the latency part that do not assume a parametric model on the incidence part require using an estimator of the cure probability that is taken as the survival at the largest observation, maybe conditionally to the covariates. In the case of the NPSXX, this probability of cure is used to reweight the Beran estimator while NPSXZ and PVK estimators use the estimated probability of cure to adjust the sets at risk. Since the Kaplan–Meier and Beran estimators suffer from a certain bias that is much related to the difficulties of estimation at the right tail of the lifetime distribution, we could expect an important bias in the nonparametric estimation of the probability of cure based on the estimated survival at the largest observation, especially in the case of high censoring. For this reason, we suggest to include in these estimators some modification to correct for the bias. One proposal that proved to be good in reducing the bias and even the variance of the Kaplan–Meier estimator was given by Stute (1994). One adaptation of this technique, or other proposals available in the literature, could be useful to improve the estimation of both the incidence and the latency parts in a mixture cure model.

Another improvement of the estimation in mixture cure models could be obtained if it is assumed that the censoring variable does not depend on the covariates. In this case, global Kaplan–Meier weights can be considered ignoring the covariates. Later

on, these weights could be used to estimate joint distributions of the lifetime with any of the covariates or conditional distributions by using a certain smoothing technique. See (Stute 1996) for details on this type of models and estimators with censored data. Note that this approach would require much less computations.

Finally, we have observed that in the mean integrated squared error computed both in the bootstrap methods to obtain the bandwidths and also to evaluate the latency estimators in the simulation study, the integral was defined with respect to a weighting function giving low weight to the right tail of the latency distribution. In the simulation study, this weight was zero above the 90th percentile. We understand that this is due to the difficulties to estimate the latency distribution at its right tail, but then it could be interesting to consider quantile methods. These methods could be useful to assess the effect of covariates on different quantiles of the latency distribution, and even in a conditional approach where the researcher is just interested in the latency distribution for certain values of the covariates, some crucial quantiles could be the most relevant information both from a practical point of view and with the goal of providing reliable estimations.

Acknowledgements This work was supported by the project PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021/2024 (ED431C2021/24) from the Xunta de Galicia.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

López Cheda A, Jácome MA, Cao R (2017) Nonparametric latency estimation for mixture cure models. TEST 26:353–376

Patilea V, Van Keilegom I (2020) A general approach for cure models in survival analysis. Ann Stat 48:2333–2346

Stute W (1994) Improved estimation under random censorship. Commun Stat Theory Methods 23:2671–2682

Stute W (1996) Distributional convergence under random censorship when covariables are present. Scandinavian J Stat 23:461–471