The Bayesian Chi-Square Statistic to determinate Kaplan Meier validity

Ahmed Hamimes[1]
M.A.A
École nationale de la statistique et d'économie appliquée
ahmedhamimes@yahoo.com
0779988700

Rachid Benamirouche
professor
École nationale de la statistique et d'économie appliquée
rbena2002@hotmail.com
0550485125

Received date: 24/08/2020 Accepted date: 24/10/2020 Published date:28/10/2020

Abstract:

In this article, we use the Kaplan Meier model in the survival analysis and according to a Bayesian conception, of this context the choice of probabilized the risk of chance whatsoever in the law used (a priori law of beta, the a hierarchical priori,…) or in the structure of the model (to test the foolability,…) requires an adjustment test of the survival data according to the chi-square test of bayesian through the p-value. The test developed in this article is an ideal choice to reinforce the results found through one of the most widely used survival methods, the Kaplan Meier model.

Keywords: Bayesian approach; Kaplan Meier model; p-value; chi-square test.

Jel Classification Codes: **:** C11,C12, C41.

1. Introduction

The term survival time refers to the time that has passed until a particular event occurs. The studied event (commonly called "death") is the irreversible passage (commonly called "alive" and "dead") between two states. The terminal event is not necessarily death: it can be the onset of a disease (for example, the time before a relapse or transplant rejection), a cure (time between diagnosis and recovery), a machine failure (machine operating time, reliability) or a complaint (time between two complaints, actuarial services). Analysis of survival (duration) data is a measure of the pause in the occurrence of this incident. In the field of biomedicine, we study these durations in the

[1] Corresponding author: Hamimes Ahmed, e-mail: ahmedhamimes@yahoo.com.

framework of longitudinal studies such as cohort surveys (monitoring of patients over time) or therapeutic trials (testing of the efficacy of a drug). Therefore, we seek to estimate the distribution of survival times (survival function), to compare multi-group survival functions or to analyze how explanatory variables modify survival functions.

Several authors have discussed the use of Bayesian inference in survival analysis (Ferguson, 1973; Kalbfleisch, 1978), to compare multi-group survival functions or to analyze how explanatory variables modify survival functions. . (Kalbfleisch and Prentice, 1980; Clayton, 1978), in which the prior coefficients of the covariates for the base rate and for the regression coefficients are specified differently. Indeed, one of the advantages of using Bayesian methods to jointly model the regression coefficients of the covariates and the baseline mortality rate is that by using MCMC techniques, we can precisely calculate the posterior distributions of the model. as well as their standard deviations. Rational a priori distribution specification functions as well as the performance of intensive computations remain. The results of these Bayesian proportional hazards studies demonstrated the accuracy of the estimates and the potential benefits of using these methods to analyze survival data (Giorgi, 2002). Parametric models have taken a considerable place in Bayesian survival analysis and parametric modeling offers direct modeling. The statistical literature using a Bayesian parametric approach in analysis and survival tests is very vast, we give here some references in the medical or public health field (Grieve, 1987; Achcar et al., 1987; Achcar et al., 1985; Chen et al., 1985; Dellaportas and Smith, 1993; Kim and Ibrahim, 2001). The book by Ibrahim et al. (2001) gives a good overview of Bayesian models of survival in general, and in particular of parametric models. We can also cite the Bayesian method of breaking point models introduced by Carlin, Gelfand and Smith (1992) in parametric modeling. Florens and Rolin (2001) have mainly demonstrated in nonparametric modeling that estimates by simulation of the Dirichlet process and nonparametric Bayesian inference give good results.These different works have been developed in different methodological contexts, using different a priori and / or by modeling the cumulative risk function or directly the instantaneous risk function (Giorgi 2002). Different parametric, semi-parametric and non-parametric models have been developed in the classical approach, (cite here, KM, Gamma, Log-logistic, Cox (classical version) ... The Kaplan-Meier method makes it possible to estimate survival functions, without requiring regular time intervals, unlike the actuarial method: survival curves are used over time to analyze changes in the size of a given population.

In the context of probabilized, the risk of chance whether in the choice of the law used (a priori law of beta, the hierarchical a priori, ...) or in the structure of the model (to test the potability, ...), the method Kaplan Meier requires a goodness-of-fit test to test the reliability of these results. Our proposed method allows users of this approach and through the Bayesian chi-square test and the p-value, the verification of the results obtained in a fairly simple and straightforward manner.

2. Kaplan-Meier method[1]

The Kaplan-Meier (KM) estimation method is also called by Anglo-Saxon statisticians "Product Limit Estimations (PLE)". This estimator, which is a generalization of the notion of empirical distribution function, is based on the following idea: to survive after a time t is to be alive just before t and not to die at time t, then the probability of survival at time t_i is equal to the probability of having survived before t_i multiplied by the “conditional” probability of surviving at time t_i . The use of the term "conditional" here means that it is the probability of surviving time t_i knowing that the individuals were survivors in t_i :

$$S(t_i) = P(X > t_i/X \geq t_i) * S(t_{i-1})$$

The probability of survival at t_i then becomes:

$$S(t_i) = S(t_{i-1})(1 - h_i)$$
$$= S(t_{i-1})\frac{n_i - d_i}{n_i}$$

such as

$t_i$represents the follow-up time since inclusion in the study for each individual i.

$d_i$est le nombre de décès au temps t_i;

$n_i$is the number of subjects at risk of presenting the event studied at the instant t_i, i.e. the number of patients who have not yet undergone the event nor the censorship just before t_i.

$(1 - q_i)$ represents the proportion of people who did not experience the event.

The probability of survival in t_ide then becomes:

[1] En 1958, Kaplan et Meier ont introduit l’estimateur de la fonction de survie nommé estimateur de Kaplan-Meier.

$$S(t_i) = S(t_{i-1})(1 - h_i)$$
$$= S(t_{i-1})\frac{n_i - d_i}{n_i}$$

Selon cette équation, la probabilité de survie t_i sachant qu'on était en vie en t_{i-1} est estimée de la manière suivante :

$$\hat{S}(t_i/t_{i-1}) = \frac{n_i - d_i}{n_i}$$

By extension, if we consider $t_1 < t_2 < \cdots < t_n$ the distinct survival times of n individuals, $\hat{S}(t)$ corresponds to the product of all the probabilities of not having known the event since the start of the observation:

$$\hat{S}(t) = \begin{cases} \prod_{t_i \leq t}\left(1 - \frac{d_i}{n_i}\right) = \prod_{t_i \leq t}(1 - h_i) & \text{si } t \geq t_1 \\ 1 & \text{si } t < t_1 \end{cases} \qquad (1)$$

3. The Bayesian conception of the Kaplan Meier estimator

In a Bayesian view it is assumed that the number of deaths in the interval of time is a Binomial distribution given by:

$$d_i \sim \beta in(n_i, q_i) \qquad (2)$$

the parameters $q_{i,}$ in the Bayesian framework are random variables, and when the distribution used in the case of the proportions is that of Beta, we set:

$$q_{i\sim} beta(\alpha, \beta) \qquad (3)$$

the prior law is considered to be weak informative, it provides solutions in the use of algorithms.We ask:

$$q_i \sim \beta(0.01, 0{,}01) \qquad (4)$$

for a binomial distribution and a conjugate prior distribution, we set

$$f_\pi(d_i/\alpha,\beta) = \int_0^1 f(d_i/q_i)\,\pi(q_i/\alpha,\beta)dq_i$$

$$= \int_0^1 [q_i(1-q_i)]^{-1} C_{n_i}^{d_i} {q_i}^{d_i}(1-q_i)^{n_i-d_i}\,dq_i$$

$$= C_{n_i}^{d_i}\frac{1}{B(\alpha,\beta)}\int_0^1 {q_i}^{d_i+\alpha-1}(1-q_i)^{n_i-d_i+\beta-1}\,dq_i$$

$$= C_{n_i}^{d_i}\frac{B(\alpha+d_i, n_i+\beta-d_i)}{B(\alpha,\beta)}$$

which provides a beta – binomial distribution to estimate $\hat{\alpha}, \hat{\beta}$, in order to calculate $\pi\left(q_i/d_i, \hat{\alpha}, \hat{\beta}\right)$.

also let :

$$\begin{cases} n_1 = le\ nombre\ de\ sujet\ dans\ le\ début\ d\ l'étude \\ n_i = n_{i-1} - d_i - c_i \end{cases},$$

4. The Bayesian chi-square and p-value test

To check the quality of the Kaplan Meier Bayesian model in this operation we use the observed values of d_i to form the statistic

$$\chi^2_{obs} = \sum_i \frac{(d_{obs.i} - \mu_i)^2}{\sigma_i^2} = \sum_i \frac{(d_{obs.i} - (q_i * n_i))^2}{(q_i * n_i) * (1 - (q_i)} \qquad (5)$$

We then generate predicted values of d_i from its posterior predictive distribution, and construct an analogous statistic:

$$\chi^2_{rep} = \sum_i \frac{(d_{rep.i} - \mu_i)^2}{\sigma_i^2} = \sum_i \frac{(d_{resp.i} - (q_i * n_i))^2}{(q_i * n_i) * (1 - (q_i)} \qquad (6)$$

evaluation of subsequent distributions of $D(d_{obs.i}, q_i)$ and $D(d_{rep.i}, q_i)$provides individual and aggregate measures of goodness-of-fit that can be described graphically or using tail region probabilities called posterior predictive values (Meng, 1994).

$$p - value \equiv P\left[D(d_{rep.i}, q_i)\ \geq D(d_{obs.i}, q_i)/d_i\right]$$

Gelman, Meng and Stern (1996) recommend calculating the "predictive p-value"

$$p - value \equiv P\left[D(d_{rep.i}, q_i) \geq D(d_{obs.i}, q_i)/d_i\right]$$

$$= \int \int I_{\left[D\left(d_{rep.i}, q_{constant}\right) \geq D\left(d_{obs.i}, q_{constant}\right)\right]} f({d_i}^{rep}/q)\, \pi(q_i/d_i) {d_i}^{rep} d\theta \quad (7)$$

where I is the indicator function.

This integral can be approximated by sampling ${q_i}^k$ from the a posteriori distribution of q_i, the same thing fo ${d_i}^{rep}$ from the distribution $f({d_i}^{rep}/q_i)$. In the result we find:

$$p - value = \sum_{k=1}^{T} I\left[D\left({d_i}^k, {q_i}^k\right) \geq D(d_i, {q_i}^k)\right]/T \qquad (8\,)$$

5. Application

In this section, the survival function for two pharmaceutical substances (placebo and prednisolone) is estimated in a clinical study, this example uses the survival times of 42 patients with chronic active hepatitis. These patients were randomized into two equal groups; one received treatment with prednisolone, the other received a placebo (see Held, 2010). Patients with prednisolone are used in this example.

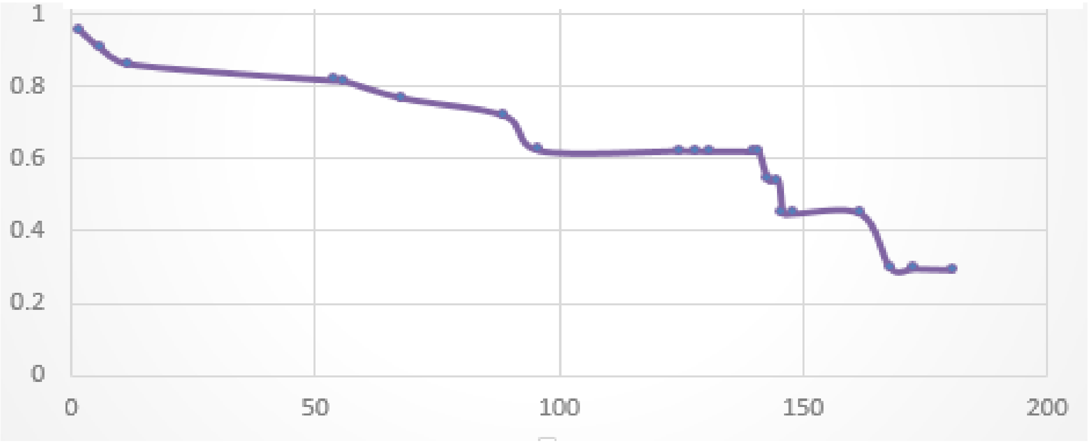
Tab 1 : the survival probabilities by the Bayesian Kaplan Meier method.

Time	Total No of Deaths	Total No of censored	No at risk	Kaplan Meier
2	1	0	21	0.9545
6	1	0	20	0.9082
12	1	0	19	0.8624
54	1	0	18	0.8169
56	0	1	17	0.8164
68	1	0	16	0.7686
89	1	0	15	0.7198
96	1	0	14	0.6233
125	0	1	13	0.6228
128	0	1	12	0.6223
131	0	1	11	0.6218
140	0	1	10	0.6211
141	0	1	9	0.6204
143	1	0	8	0.5414
145	0	1	7	0.5406
146	1	0	6	0.4502
148	0	1	5	0.4492
162	0	1	4	0.4482
168	1	0	3	0.2985
173	0	1	2	0.2969
181	0	1	1	0.294

Source: Developed by us.

In this article, we use a conjugate prior law (due to the absence of any a priori information on the estimated survival model) such that $\alpha = \beta = 0.01$.

Fig 1- The survival curve estimated according to the Bayesian Kaplan-Meier method with an a priori of beta.



Source: Developed by us.

From figure (1), we notice that the treatment study includes 100% of the individuals in the sample at the start of the curve. 50% of patients died after more than 146 days of using the sample drug. But for the majority of people in the study, medication failure lasts a long time, for some it reaches 180 days.

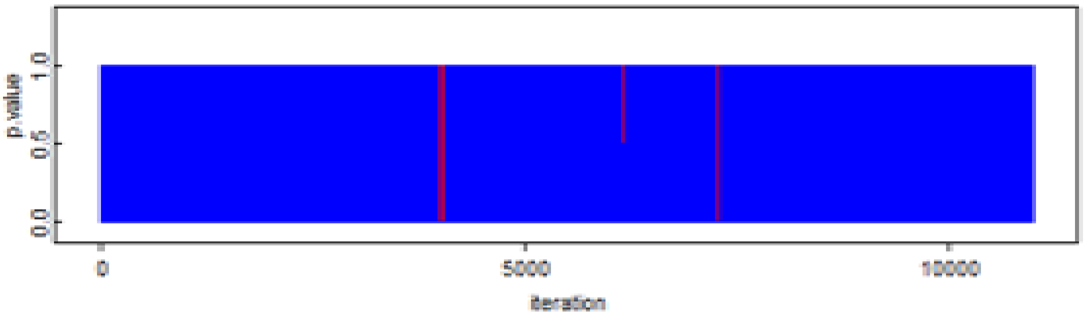
Tabl2 : Estimated Parameters of the Kaplan Meier model.

	mean	**sd**	**MC_error**	**val2.5pc**	**val97.5pc**
p.value	0.555	0.497	0.01294	0.0	1.0

Source: Developed by us, using OpenBUGS program.

The predictive p-value a posteriori or else known under the name of PPP-value (p-value = 0.555) this value is close to 0.5 so the distributions of the repeated and real data are close. Also this value is directly interpreted as the probability of observing in future samples with $D(d_i, q_i{}^k)$ higher than that already observed. In another way the Bayesian Kaplan Meier model and on this choice of the a priori is a significant model.

Fig 2: The trace of the posterior distribution for the p-value parameters..

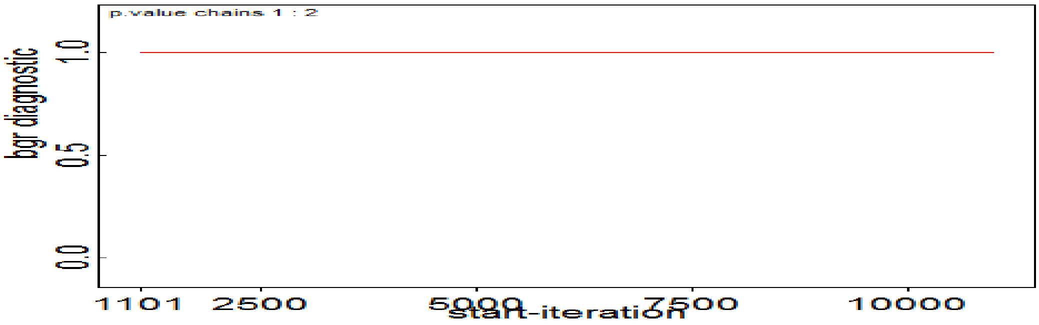


Source: Developed by us, using OpenBUGS program.

In Figure 2, each color denotes an MCMC chain. The two chains mix well: convergence is achieved (see code in appendix).

Brooks and Gelman in 1998 proposed a generalization of the method of Gelman and Rubin which was introduced in the year 1992, it is a method of validating the ergodic sequences of the MCMC algorithms.

FIg .3: The Brooks and Gelman graph "convergence - diagnosis - graph"



Source: Developed by us, using OpenBUGS program.

The green curve indicates the width of the 80% inter-chain credibility interval. The blue curve indicates the average width of the within-chain 80% credibility intervals. The red curve indicates the Brooks and Gelman statistic (i.e., the ratio of the green / blue curves). The Brooks and Gelman statistic tends towards 1, which means that there is convergence.

6. Conclusion

In this article, Bayesian p-value was used to measure the goodness of fit of Kaplan Meier's Bayesian model in survival analysis. In the proposed example and we find that this method allows to verify the validity of the Bayesian model of Kaplan Meier in a fairly simple and straightforward way.

7. References

- Achcar, J.A., Bolfarine, H., Pericchi, L.R.(1986), Transformation of survival data to an extreme value distribution, the statistician, 36, 229-234.
- Achcar, J.A., Brookmeyer, R., Hunter, W.G (1985), An application of Bayesian analysis to medical follow-up data, Statistics in Medicine, 4, 509-520.
- Carlin, B.P., Gelfand, A.E., Smith, A.F.M (1992), Hierarchical Bayesian Analysis of Change point Problems, Appl. Statist. 41(2), 389-405.
- Chen, W.C., Hill, B.M., GREENHOUSE, J.B., FAYOS, J.V (1985), Bayesian analysis of survival curves for cancer patients following treatment, In Bayesian Statistics 2 (Eds. J.O. Berger, J. Bernardo. A.F.M. Smith), Amsterdam: North-Holland, 299-328.

- Clayton, D.G., A (1978), Model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika, 65, 141 – 151.
- Dellaportas, P., Smith, A.F.M (1993), Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, Applied Statistics, 42, 443-459.
- Dezfuli, H., Kelly, D., Smith, C., Vedros, K., & Galyean, W (2009), Bayesian inference for NASA probabilistic risk and reliability analysis.
- Ferguson, T.S (1973), A Bayesian analysis of some nonparametric problems, The Annals of Statistics, 1(2), 209-230.
- Florens, J.P., Rolin, J.M (2001), Simulation of posterior distributions in nonparametric censored analysis, International Statistical Review, 67(2), 187-210.
- Giorgi, R (2002), Analyses comparatives des méthodes de survie et extensions d'un modèle régressif de survie relative : prise en compte de la non-proportionnalité des risques par des fonctions B-splines et développement d'une méthode d'analyse bayésienne, Thèse de doctorat, Université Aix-Marseille II, France.
- Grieve, A.P (1987), Applications of Bayesian software: Two examples, The Statistician, 36, 283-288.
- Held, U (2010), Représentation graphique et comparaison de courbes de survie. Forum Med Suisse, 10(33), 548-550.
- Ibrahim, J. G., Chen, M, H., Sinha, D (2001), Bayesian survival analysis, Springer.
- Kalbfleisch, J.D.(1978). Nonparametric bayesian analysis of survival time data, Journal of the Royal Statistical Society, Series B, 40(2), 214-221.
- Kim, S.W., Ibrahim, J.G (2001), On Bayesian inference for proportional hazards models using noninformative priors, Lifetime Data Analysis, to appear.

8. Appendices (OpenBUGS code)

```
model {
for (i in 1 : m) {
d[i] ~ dbin(q[i], n[i]) #Binomial model for d
q[i] ~ dbeta(alpha, beta)
d.rep[i] ~ dbin(q[i], n[i]) #Replicate from posterior predictive distribution
diff.obs[i] <- pow(d[i] - q[i]*n[i], 2)/(n[i]*q[i]*(1-q[i])) #Difference between observed and
expecte d
diff.rep[i] <- pow(d.rep[i] - q[i]*n[i], 2)/(n[i]*q[i]*(1-q[i])) #Difference between replicated
and expected d
}
n[1]<- 22
alpha ~ dgamma(0.0001, 0.0001)
beta ~ dgamma(0.0001, 0.0001)
chisq.obs <- sum(diff.obs[])
chisq.rep <- sum(diff.rep[])
p.value <- step(chisq.rep - chisq.obs) #Value should be near 0.5 for homogeneous data
for(i in 2:m){
n[i]<-n[i-1]-d[i-1]-ce[i-1]
}
}
list(m=21,d=c(1,1,1,1,0,1,1,2,0,0,0,0,
0,1,0,1,0,0,1,0,0),
ce=c(0,0,0,0,1,0,0,0,1,1,1,
1,1,0,1,0,1,1,0,1,1))
list(alpha=0.1,beta=0.5)
list(alpha=0.5,beta=0.9)
```