

TESTING HOMOGENEITY AND GOODNESS OF FIT IN SURVIVAL DATA

Jana Timková

Keywords: Survival analysis, counting process, MCMC, goodness of fit, homogeneity test, residual analysis.

Abstract: The present paper deals with the goodness of fit and the two-sample problem testing related to the event-history type data. The proposed graphical methods are related to the bayesian nonparametric approach and take advantage of MCMC estimation of the hazard rate. The first technique of the homogeneity testing is based on the method of Arjas [2] who estimated the hazard rate under the null hypothesis (all individuals are governed by the same rule) and then plotted the estimated cumulative hazard rate against the number of failures in one group and in the other group separately. The second method to detect the inhomogeneity in two groups uses the estimation of the hazard rate based on the first - control - group and compare it to the second one. The article conclude with a concrete application of the methods on a vaginal cancer mortality data.

Abstrakt: Súčasný článok sa zaoberá problémom testu dobrej zhody a dvojvýberovým testom v oblasti dát prežitia. Predložené metódy sú spojené s bayesovským neparametrickým prístupom a MCMC odhadmi pre funkciu hazardu. Prvý test homogeneity vychádza z článku Arjas [2], v ktorom je odhadnutá funkcia hazardu za predpokladu platnosti nulovej hypotézy, čiže všetky objekty sa správajú rovnako nezávisle na rozdelení do skupín. Následne sa vykreslí odhadnutá kumulatívna funkcia hazardu proti kumulatívnemu počtu udalostí v oboch skupinách zvlášť. V druhom prístupe odhadneme funkciu hazardu v jednej skupine a podobne ako v predchádzajúcom porovnáme s druhou. Článok je zakončený konkrétnou aplikáciou na reálnych dátach.

1 Introduction

The aim of the analysis of life history data is to describe the behaviour of the underlying processes what is usually done by assuming a suitable model for the hazard rate. When the model is fitted, the next step is the assessment of quality of the estimation and that leads us to the goodness of fit testing. Over last decades many powerful and convenient methods of detecting, whether such a model is unsatisfactorily specified, arised. Particularly in case of presence of some covariates, the inference is related to the form and significance of the dependence of the observed times on those covariates. The known methods are mostly based on the martingal properties of the differences between observed cumulative number of failures and the estimated cumulative hazard rate.

Let us consider n parallel mutually independent counting processes $N_1(t)$, $N_2(t), \ldots, N_n(t)$ observed in the time interval $[0, T]$. We assume that all processes start from zero, $N_i(0) = 0$, and that $N_i(t)$ increases in $+1$ when the i-th object happens to meet an event of interest. The i-th process is expected to behave according to an intensity process $I_i(t) \cdot \lambda(t)$ where $\lambda(t)$ is a bounded nonnegative continuous hazard function and $I_i(t)$ is an indicator process (indicating whether the i-th individual is at risk of event, i.e. $I_i(t) = 1$ when it is at risk, $I_i(t) = 0$ otherwise). The indicator process has its importance when the censoring is present or when the occurence of event implies the end of the observing of the object. As the result we obtain the multivariate counting process $\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^T$.

The hazard rate related to i-th object, $\lambda_i(t)$, is the instantaneous rate of an event occuring at time t defined as

$$\lambda_i(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(N_i(t + \Delta t) - N_i(t) = 1 | \mathcal{F}_t),$$

where $\mathcal{F}_t := \sigma\{N_i(s), 0 \leq s \leq t\}$ is σ-algebra of the history of the i-th object up to time moment t. In next we will consider a special case of previous setting - a typical survival data with one possible event on each object (either death or censoring).

Now, let us introduce the generalized residual at given time t as

$$X_i(t) = N_i(t) - \int_0^t I_i(u)\lambda_i(u)du$$

for all i. Then the process $(X_i(t), t \in [0, T])$ is an $\mathcal{F}_t$-martingale and is called the residual process for i-th individual. Let us have $t_1, \ldots\ t_n$ observed times of events. By considering $X_i(t)$ at the ordered set of uncensored failure times $t_{(1)},\ t_{(2)}, \ldots, t_{(K)}$, where K is the total number of uncensored failures, we obtain

$$X_i(t_{(k)}) = N_i(t_{(k)}) - \int_0^{t_{(k)}} I_i(u)\lambda_i(u)du, \quad k = 1, \ldots, K.$$

an discrete time martingale of parameter k. In fact, the integrated intensity process approximates the number of uncensored events observed on the individual. Hence, the motivation for the testing technique is straightforward. Let us have the cumulative hazard rate $\Lambda_i(t) = \int_0^t \lambda_i(u)\ du$ and after denoting $\overline{X}(t) = \sum_{i=1}^n X_i(t)$ and $\overline{N}(t) = \sum_{i=1}^n N_i(t)$,

$$\overline{X}(t_{(k)}) = \overline{N}(t_{(k)}) - \sum_{i=1}^n \int_0^{t_{(k)}} I_i(u)\lambda_i(u)du = k - \sum_{i=1}^n \Lambda_i(t_{(k)} \wedge t_i),$$

is close to zero for all $k \in \{1, \ldots, K\}$. Now we can substitute the unknown hazard rate with the appropriate estimation and use it for testing the goodness of fit as well as homogeneity among groups.

Naturally, the martingale property of the residual process is not always preserved after plugging-in the estimated hazard rate but there are cases in which it is true. There is certain number of publications regarding the topic in various models of hazard rate; see e.g. excellent [7] for the generalized residuals in the Cox model, further [11] in Aalen regression model and [6] in Cox-Aalen model. Apparently, once the estimated residual process is a martingale, the central limit theorem for martingales (cf. [4]) gives the weak convergence to zero-mean Gaussian process. This is the way to obtain the confidence bands around the estimated residual process needed for tests, although often tedious and in some cases intractable because of complicated covariance structure of the limiting process.

The alternative which introduces itself is to take the bayesian approach for modelling and estimating, gain the sample of the estimated hazard rate via MCMC algorithms and use the pointwise credibility bands for the estimated residual process. There were developed many bayesian models for hazard rate, for a good overview see [8] or more detailed [9].

Useful graphical methods based on the residual processes were introduced by e.g Andersen [1] who considered various plots mainly focused on the checking the proportional-hazards assumption of Cox model, and later by Arjas [2] who came up with the plot of the estimated cumulative hazard against number of failures in stratas, in which the observed individuals were splitted.

In next sections, a bayesian model for the hazard rate will be introduced and the testing techniques for two-sample problem will be carried out. The results will be presented in example of a real data analysis.

2 Model and estimation of the hazard rate

Here we describe one possible model which can be used in bayesian setting. Let us for now assume that all observed objects are governed by the intensity processes with the same hazard rate $\lambda(t)$ without taking any covariate process into account. To estimate the hazard rate we use nonparametric method firstly proposed by Arjas and Gasbarra [3]. They assumed the piecewise constant function for the hazard rate. All the individuals were expected to behave according to a common hazard rate $\{\lambda_t, t \in [0,T]\}$. Further in their model the hazard function was constant within the $(m+1)$ intervals which came from dividing the whole time interval $[0,T]$ by m jump times $T_1, \ldots, T_m$. The level of hazard function within the interval $(T_{j-1}, T_j]$ was denoted as λ_j. The number of jump times m could vary. To reach the estimates the MCMC were used.

Let us denote $H = (\lambda_1, T_1, \ldots, \lambda_m, T_m, \lambda_{m+1})$. The main idea is to provide as many iterations as necessary in which the piecewise constant hazard function defined by vector of parameters H is upgraded according to the data and prior information. As soon as we possess the Markov chain $(H^{(i)})_{i=1}^{R} = (\lambda_1^{(i)}, T_1^{(i)}, \ldots, \lambda_{m^{(i)}}^{(i)}, T_{m^{(i)}}^{(i)}, \lambda_{m_{(i)}+1}^{(i)})_{i=1}^{R}$, where R is the number

of provided iteration and parameters denotes with $^{(i)}$ are those connected with i-th iteration and we are ensured about the ergodicity of the chain, we could compute the resulting estimate of the hazard function as the mean of the members of the chain. Finally, even though the simulated hazard rates are piecewise constant functions, asymptotically the mean of them yields to a continuous function.

Except few changes we adopt the similar structure of the model for the hazard rate. The chosen prior distribution of the levels of hazard rate λ_i, $i = 1,\ldots,m+1$ is gamma distribution with particular setting of parameters dependent on parameter α to fold in the martingale structure. The jump times are distributed as the homogeneous Poisson process in $[0,T]$ with unknown constant rate μ. Further, we extended Arjas and Gasbarra's model by accompanying the hierarchical structure into model. For details about the estimation and the discussion on the ergodicity of the chain see [10].

3 Testing

As it was announced before the main focus of the paper is in a simple two-sample problem which occurres when the observed objects come from two different environments. Also more complicated cases could be transformed to this situation by dividing the individual processes in two stratas according to certain values of covariates or the observed times themselves (as it was done in [2, sec. 3.3]).

3.1 Arjas's plot

Arjas [2] suggested to provide the estimation under the assumption that the individuals are divided in two or more stratas according to some covariates observed alongside (more clearly, in his case he worked with the Cox model with p covariates and wanted to check whether $(p+1)$-th covariate should by included in model). We accommodate his approach in our simple case of two stratas. The null hypothesis for us is that the observed individuals (i.e. the counting processes) behave in the same way in both stratas. We provide the estimation of the hazard rate according to the null hypothesis what means to use all observations from all objects together, as if they were replicates of each other, to estimate the hazard rate. Once we have the estimated hazard rate $\hat{\lambda}_i(t) \equiv \hat{\lambda}(t), 0 \le t \le T$, for all i, we can compute the estimated cumulative hazard function for $0 \le t \le T$ in the first strata as follows

$$\hat{H}_{(strata1)}(t) = \sum_{i \in strata1} \int_0^t I_i(s)\hat{\lambda}(s)\, ds \tag{1}$$

and analogically in the second one. Now we base plotting technique on the zero-mean martingale property of difference between observed numbers of failures and the real cumulative hazard function. Let $t_{(1)} \le t_{(2)} \le \cdots \le t_{(N)}$ be ordered observed times of failures. The idea is to plot $\hat{H}_{(strata1)}(t_{(i)})$

and $\hat{H}_{(strata2)}(t_{(i)})$ against the observed cumulative number of failures in the first strata and the second strata respectively. In case the null hypothesis of homogeneity of both groups is correct, both lines should be placed near the diagonal line $f(x) = x$. If the lines are far from each other and from the diagonale then the null hypothesis is rejected. The significant departure from the null hypothesis is proposed to be detected by the pointwise 95% credibility bands created from the sample of $\hat{H}_{strata1}$ calculated from the posterior sample of the hazard rate.

3.2 Two-sample test

The basis of the second proposed approach is to provide the estimation of the hazard rate from the observations belonging to one - control - group of objects. Then we calculate the estimated cumulative hazard function $\hat{H}_{(strata1)}$ accordingly to (1) and plot $\hat{H}_{(strata1)}(t_{(i),2})$ against the observed cumulative number of failures in the second strata. In the previous we used the notation $t_{(1),2} \leq t_{(2),2} \leq \cdots \leq t_{(N_2),2}$ for the observed times of failures occurring in the second group of objects with the total number of observations denoted as N_2. If both groups of objects possess the same behaviour, the line is near the imaginary line $f(x) = x$. Again, to detect a significant departure from the null hypothesis we use the simulated pointwise confidence bands drawn from the MCMC simulation output.

Apart from the homogeneity testing, the introduced idea is plausible for testing of overall goodness of fit. Similarly, as in (1) for stratas, we can calculate $\hat{H}(t) = \sum_{i=1}^{n} \int_0^t I_i(s)\hat{\lambda}(s)\, ds$ for whole dataset. Plotting the values of $\hat{H}(\cdot)$ in observed uncensored failure times versus number of observed failures is close to diagonale $f(x) = x$ if the fit is sufficient.

4 Application

In following the application on a real data is presented. The data set was analyzed in [5]. The observations give the times from insult with the carcinogen DMBA to mortality from vaginal cancer in rats. Two group were distinguished by a pretreatment regimen. We want to compare the impact of different regimens on the waiting times to death. The data represents the usual time-to-event setting with one observation (failure or censored event) belonging to each individual.

First, the hazard rate was estimated separately for each group as well as for the whole set of observations. The estimated hazard rates are plotted in figure 1. In the bottom part of the figure the observed times of events for both groups are displayed, the failures are denoted with $\circ$ and censored events with $+$. The upper row of observations is from the first group and bellow is the row of the second group's observations. Figure 2 shows the plots related to the estimated cumulative hazard rate in the first group which acts as the control group here. In the left side the goodness of fit of the estimated hazard

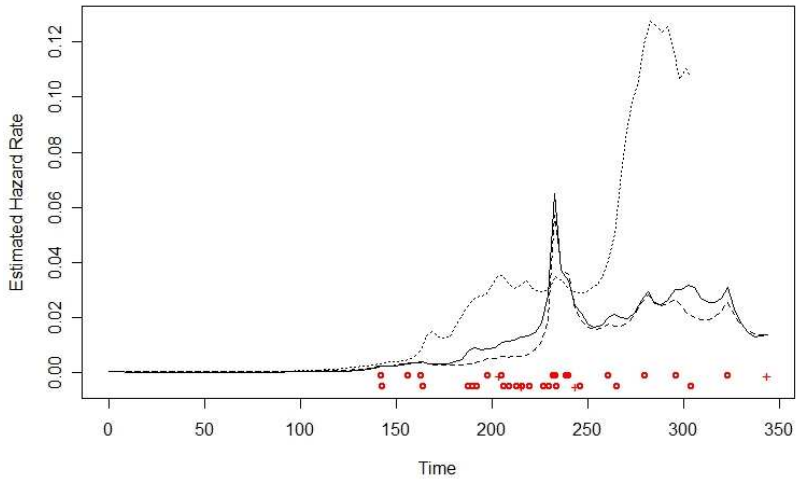

Figure 1: The plot of the estimated hazard rate in the whole set of observations (solid line) and in the two groups separately (dashed lines). At the bottom of the figure the data are displayed in two lines - the observations from the first group in the upper level and from the second group below.

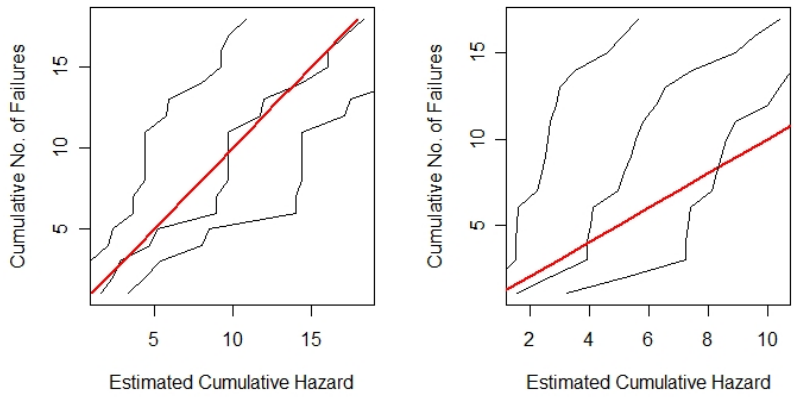

Figure 2: Plots of the cumulative estimated hazard rate against the cumulative number of failures; in the left side the goodness of fit test related to the control group, the right-hand plot displayes the test of homogeneity of both groups. The 95% pointwise credibility bands are included.

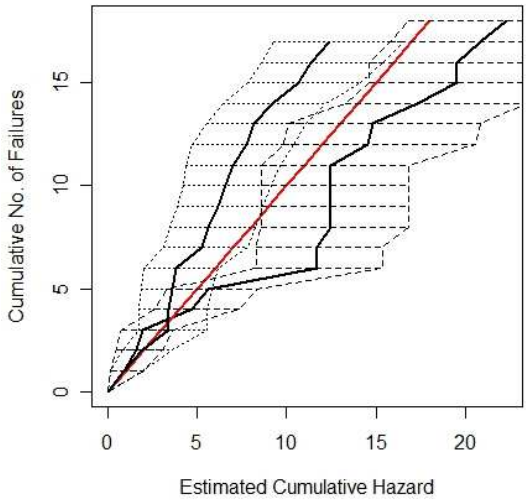

Figure 3: Arjas's plot of the homogeneity of two groups. The 95% pointwise credibility bands are included.

rate to the control group is shown and the right-hand plot displayes the test of homogeneity described in the section 3.2. Further, in figure 3 Arjas's plot is presented. Obviously, we can reject the hypothesis of homogeneity of mortality in two groups of rats.

The parameters chosen for the MCMC estimation of the hazard rate was $\mu_a = 10$, $\mu_b = 100$, $\alpha_a = 1000$, $\alpha_b = 100$, $\alpha_0 = 1$ and $\beta_0 = 10000$. The number of iterations was 5000 and first 1000 were erased.

5 Conclusion and discussion

The main idea of the previous pages was to think of alternative way of assessing the goodness of fit and homogeneity of two samples in survival data. The proposed methods are related to bayesian methodology and work with a sample of simulated hazard rates from MCMC procedure. The advantage is concentrated in the fact that existence of the sample allows us to create the testing machinery based on the calculated pointwise credibility bands instead of the rather cumbersome asymptotics. This type of approach can also be a good starting point for testing in more complicated models in which the key martingale property after plugging-in the estimator is lost.

The disadvantage of the method is that the result of the test depends also on the precision of the estimation which is connected with the choice

of parameters. Theoretically, as long as the graphical goodness of fit test approves the fit of the estimated hazard rate to the observed data of the control group, the test of the homogeneity of the second group with the control group should be valid. Nevertheless, as the results of test may differ among the group of the approved fits, it is important to decide which from those estimators suits the data and situation the best and accordingly to this provide the test.

References

[1] Andersen P.K. (1982). *Testing goodness of fit of Cox's regression and life model.* Biometrics **38**, 67–77.

[2] Arjas E. (1988). *A graphical method for assessing goodness of fit in Cox's proportional hazard models.* Journal of the American Statistical Association **83**, 401, 204–212.

[3] Arjas E., Gasbarra D. (1994). *Nonparametric Bayesian inference from right censored survival data, using Gibbs sampler.* Statist. Sinica **4**, 505–524.

[4] Fleming T.R., Harrington D.P. (2005). *Counting processes and survival analysis.* Wiley Series in Probability and Science.

[5] Kalbfleisch J.D., Prentice R.L. (2002). *The statistical analysis of failure time data.* Wiley Series in Probability and Science.

[6] Kraus D. (2004). *Testing goodness of fit of hazard regression models.* WDS 2004 Proceedings of contributed papers, Part I: Mathematics and Computer Sciences, 6–12.

[7] Marzec L., Marzec P. (1997). *Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models.* The Annals of Statistics **25**, 2, 683–714.

[8] Sinha D., Dey D.K. (1997). *Semiparametric Bayesian analysis of survival data.* Journal of the American Statistical Association **92**, 439 , 1195–1212.

[9] Ibrahim J.G., Chen M., Sinha D. (2001). *Bayesian survival analysis.* Springer-Verlag.

[10] Timková J. (2008). *Bayesian nonparametric estimation of hazard rate in survival analysis using Gibbs sampler.* WDS 2008 Proceedings of contributed papers, Part I: Mathematics and Computer Sciences, 80–87.

[11] Volf P. (1996). *Analysis of generalized residuals in hazard regression models.* Kybernetika **32**, 5, 501–510.

Poděkování: The present work was supported by the grants GA ČR 201/05/H007 and GA AV IAA101120604.

Adresa: J. Timková, MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín
ÚTIA AV ČR, Pod vodárenskou věží 4, 182 08 Praha 8

E-mail: timkova@karlin.mff.cuni.cz