Non-homogeneous Poisson process with nonparametric frailty

Vaclav Slimacek
Department of Mathematical Sciences, Norwegian University of Science and Technology,
Trondheim, Norway
email: slimacek@math.ntnu.no

Bo Henry Lindqvist
Department of Mathematical Sciences, Norwegian University of Sciences and Technology,
Trondheim, Norway
email: bo@math.ntnu.no

Abstract

The common way to describe unobserved heterogeneity for repairable systems following a non-homogeneous Poisson process is to multiply the basic rate of occurrences of failures by a random variable with a specified distribution. Since the modeled heterogeneity is unobservable, the choice of the distribution of the unobserved effects is a problematic part of using these models as well as other necessary computations related to the estimation of these models which have to be done numerically in most cases. The main purpose of this paper is to develop a method for estimation of the parameters of a non-homogeneous Poisson process with unobserved heterogeneity without having to make parametric assumptions about the heterogeneity and which avoids the frequently encountered numerical problems associated with using standard models with unobserved heterogeneity. The main idea of the presented approach is that the individual frailties are treated as unknown parameters and are estimated directly from the given data without any restrictive assumption about their distribution. This approach is illustrated on an example with the power law process.
Keywords: non-homogeneous Poisson process, unobserved heterogeneity, non-parametric estimation, power law process.

1 Introduction

In many real life applications there is a substantial heterogeneity between apparently identical repairable systems, which cannot be described by observed covariates. This unobservable heterogeneity, which is often also called frailty in the survival analysis literature, is usually modeled by multiplication of the basic rate of occurrences of failures (ROCOF) by a positive random variable taking independent values across the systems, which has unit mean and is therefore described by variance. In the traditional models with unobserved heterogeneity, it is necessary to define the distribution of the unobserved effects ([1], [2]). Since the modeled heterogeneity is unobservable, the choice of the distribution of the unobserved effects is a problematic part of using these models, similarly to the choice of the prior distribution in Bayes models. Moreover, in fitting the traditional frailty model it is necessary to integrate out the unobserved effects from the likelihood characterizing the model which, together with likelihood optimization, has to be done numerically in many cases. This can cause problems, especially in complicated models and large data sets. Suitable choice of the distribution of the unobserved effects can give interesting general results but generally the main quantity of interest is the variance of the unobserved effects, where a large variance may either indicate deficiencies in choice of the model, or that some important factors or covariates have not been taken into account, which can, for example, also characterize the quality of the model fit.
Another result from models incorporating unobserved effects, which is often overlooked but which is important for the predictions and which can provide important additional information about the properties of the

modeled system, is the ability to estimate the individual frailties. Unobserved effects can be viewed as unobserved covariates and comparison of the estimates of the individual frailties can give indication about important covariates influencing the underlying modeled process.
The main purpose of this paper is to develop a method for estimation of the parameters of a non-homogeneous Poisson process with unobserved heterogeneity which would keep most of the advantages of the traditional models with unobserved heterogeneity and at the same time eliminate the disadvantages of these models. More precisely, we aim at a method for analyzing unobserved heterogeneity without having to make parametric assumptions and which also avoids the frequently encountered numerical problems associated with using standard models with unobserved heterogeneity.
The main idea of the presented approach is that the individual frailties are treated as unknown parameters and are estimated directly from the given data without any restrictive assumption about their distribution. In addition to this, the properties of non-homogeneous Poisson processes are used for obtaining other interesting results, mainly for the estimator of the variance of the unobserved effects.
The approach is illustrated on an example with the power law process.

2 The power law process with non-parametric frailty

Let us consider m independent power law processes with unobserved heterogeneity, i.e. the rate of occurrence of failures (ROCOF) of the jth system ($j = 1, \ldots, m$) is given by

$$\lambda_j(t|z_j) = z_j a b t^{b-1}$$

while the cumulative ROCOF is

$$\Lambda_j(t|z_j) = z_j a t^b$$

where $a > 0$ and $b > 0$ are the standard parameters of the power law process, while the z_j represents the unobserved heterogeneity for the jth system. It is assumed that the z_j's are independent and identically distributed realizations of an unobserved positive random variable Z with unit mean. Note, that the classical model (i.e. model without unobserved heterogeneity) is included in this settings as a special case with $\mathrm{P}(Z = 1) = 1$.
Let the jth system be observed for a time τ_j, which is assumed to be a realization of a positive random variable τ independent of the random variable Z, denote the number of events in this system by n_j, and let the event times in this system be denoted as t_{ij} ($j = 1, \ldots, m$, $i = 1, \ldots, n_j$). Then the likelihood of the jth system, conditional on the z_j, is given by ([1])

$$\begin{aligned} L_j &= \left(\prod_{i=1}^{n_j} \lambda_j(t_{ij}|z_j) \right) \exp(-\Lambda_j(\tau_j|z_j)) \\ &= \left(\prod_{i=1}^{n_j} z_j a b t_{ij}^{b-1} \right) \exp(-z_j a \tau_j^b) \end{aligned} \tag{1}$$

The total loglikelihood for the m systems is then given by (conditionally on z_j's, $j = 1, \ldots, m$)

$$\begin{aligned} l &= \sum_{j=1}^{m} n_j \log(z_j) + n_j \log(a) + n_j \log(b) + (b-1) \sum_{i=1}^{n_j} \log(t_{ij}) - z_j a \tau_j^b \\ &= \sum_{j=1}^{m} n_j \log(z_j) + n \log(a) + n \log(b) + (b-1) S - a \sum_{j=1}^{m} z_j \tau_j^b \end{aligned} \tag{2}$$

where $n = \sum_{j=1}^{m} n_j$ and $S = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log(t_{ij})$.

If we treat the individual frailties as unknown parameters, then the score equations are given by

$$\frac{\partial l}{\partial a} = \frac{n}{a} - \sum_{j=1}^{m} z_j \tau_j^b = 0 \tag{3}$$

$$\frac{\partial l}{\partial b} = \frac{n}{b} + S - a \sum_{j=1}^{m} z_j \tau_j^b \log(\tau_j) = 0 \tag{4}$$

$$\frac{\partial l}{\partial z_j} = \frac{n_j}{z_j} - a\tau_j^b = 0, \qquad j = 1, \ldots, m \tag{5}$$

The individual frailties can be expressed from the score equations (5) as

$$z_j = \frac{n_j}{a\tau_j^b}, \qquad j = 1, \ldots, m \tag{6}$$

Solution of the score equation (4) (with use of (6)) gives the well known estimator of b ([2])

$$\hat{b} = \frac{n}{\sum_{j=1}^{m} n_j \log(\tau_j) - S} \tag{7}$$

Substitution of (6) into equation (3) will transform this equation to the form $0 = 0$, which indicates a well known identification problem ([2]). This identification problem can be solved by transformation of the original model to the equivalent model where parameters

$$a_j = z_j a, \qquad j = 1, \ldots, m \tag{8}$$

are used instead of the parameters a and z_j's.
The loglikelihood of this equivalent model is then given by

$$l = \sum_{j=1}^{m} n_j \log(a_j) + n \log(b) + (b-1)S - \sum_{j=1}^{m} a_j \tau_j^b \tag{9}$$

The score equations for this equivalent model are

$$\frac{\partial l}{\partial a_j} = \frac{n_j}{a_j} - \tau_j^b = 0, \; j = 1, \ldots, m \tag{10}$$

$$\frac{\partial l}{\partial b} = \frac{n}{b} + S - \sum_{j=1}^{m} a_j \tau_j^b \log(\tau_j) = 0 \tag{11}$$

Solution to these score equations gives the estimator of the parameters a_j's

$$\hat{a}_j = \frac{n_j}{\tau_j^{\hat{b}}}, \qquad j = 1, \ldots, m \tag{12}$$

The estimator $\hat{b}$ of the parameter b is the same as the estimator (7).
We then introduce an estimator $\hat{a}$ of the parameter a given by

$$\hat{a} = \frac{1}{m} \sum_{j=1}^{n} \hat{a}_j = \frac{1}{m} \sum_{j=1}^{m} \frac{n_j}{\tau_j^{\hat{b}}} \tag{13}$$

Hence the individual frailties z_j's can be estimated as

$$\hat{z}_j = \frac{\hat{a}_j}{\hat{a}} = \frac{n_j}{\hat{a}\tau_j^{\hat{b}}} = \frac{\frac{n_j}{\tau_j^{\hat{b}}}}{\frac{1}{m} \sum_{j=1}^{m} \frac{n_j}{\tau_j^{\hat{b}}}}, \qquad j = 1, \ldots, m \tag{14}$$

which is in agreement with (6) and (8) and these estimators also solve the score functions (3), (4) and (5). Note, that

$$\frac{1}{m}\sum_{j=1}^{m}\hat{z}_j = 1 \tag{15}$$

which means that the defined estimator $\hat{a}$ is also with agreement in (8).
The asymptotic variance of these estimators can be computed with use of the observed information matrix. Note, that these estimated asymptotic variances are computed under the assumption that the individual frailties are viewed as parameters , i.e. conditionally on z_j's, $j = 1, \ldots, m$.
The estimator of the parameter b can also be derived by using the well known property of the non-homogeneous Poisson processes which states, that the event times in the nonhomogeneous Poisson process with cumulative ROCOF $\Lambda(t)$ are, conditionally on the number of events within time τ, distributed as order statistic from a sample with cumulative distribution function $F(t) = \frac{\Lambda(t)}{\Lambda(\tau)}$. It is therefore possible to derive the distribution of $\hat{b}$ and it follows, that this distribution does not depend on z_j's, $j = 1, \ldots, m$, and it is also the same for model without unobserved heterogeneity. The estimators of $\hat{b}$ and its estimated variance are the same for both approaches and hence the variance computed with use of the observed information matrix represents also the unconditional estimator of the variance of $\hat{b}$.
The unconditional variance of the estimator $\hat{a}$ can be computed as

$$\mathrm{Var}(\hat{a}) = \mathrm{E}(\mathrm{Var}(\hat{a}|\, z_k, k = 1 \ldots, m)) + \mathrm{Var}(\mathrm{E}(\hat{a}|\, z_k, k = 1 \ldots, m)) \tag{16}$$

Thus the unconditional variances of the derived estimators can be estimated as

$$\widehat{\mathrm{Var}}\left(\hat{b}\right) = \frac{\hat{b}^2}{n} \tag{17}$$

$$\widehat{\mathrm{Var}}(\hat{a}) = \frac{1}{m^2}\sum_{j=1}^{m}\frac{\hat{a}_j^2}{n_j} + \frac{\widehat{\mathrm{Var}}\left(\hat{b}\right)}{m^2}\left(\sum_{j=1}^{m}\hat{a}_j^2\log(\tau_j)^2 + 2\sum_{1\le i<j\le m}\hat{a}_i\hat{a}_j\log(\tau_i)\log(\tau_j)\right) + \hat{a}^2\frac{\widehat{\mathrm{Var}}(Z)}{m} \tag{18}$$

where $\widehat{\mathrm{Var}}(Z)$ denotes an estimator of the variance of the unobserved effects which can be obtained with the use of properties of the non-homogeneous Poisson process.
The conditional expectation and variance of the number of events N within given time τ and with given frailty Z in the system following the power law process is equal to

$$\mathrm{E}(N|Z,\tau) = Za\tau^b \tag{19}$$

$$\mathrm{Var}(N|Z,\tau) = Za\tau^b \tag{20}$$

Note, that

$$\mathrm{E}(N) = \mathrm{E}(\mathrm{E}(N|Z,\tau)) = a\mathrm{E}\left(\tau^b\right) \tag{21}$$

$$\mathrm{E}\left(\frac{N}{\tau^b}\right) = \mathrm{E}\left(\mathrm{E}\left(\frac{N}{\tau^b}\middle|\, Z,\tau\right)\right) = \mathrm{E}\left(\frac{1}{\tau^b}\mathrm{E}(N|Z,\tau)\right) = a \tag{22}$$

which gives

$$a = \frac{\mathrm{E}(N)}{\mathrm{E}(\tau^b)} = \mathrm{E}\left(\frac{N}{\tau^b}\right) \tag{23}$$

This equality can be interpreted as that the ratio of the mean number of events and the mean time of the observation of a process is the same as the mean of the ratios of number of events and observation time in the individual processes and both ratios can be used for the estimation of the parameter a. The empirical version of the third term in (23) is the derived estimator (13).
The unconditional variance of the number of events can be then computed as

$$\begin{aligned}\mathrm{Var}(N) &= \mathrm{E}(\mathrm{Var}(N|Z,\tau)) + \mathrm{Var}(\mathrm{E}(N|Z,\tau)) \\ &= a\mathrm{E}\left(\tau^b\right) + \mathrm{E}\left(Z^2\right)a^2\mathrm{E}\left(\tau^{2b}\right) - a^2\mathrm{E}\left(\tau^b\right)^2\end{aligned} \tag{24}$$

which together with (23) gives

$$\mathrm{E}\big(Z^2\big) = \frac{\mathrm{Var}(N) - a\mathrm{E}\big(\tau^b\big) + (a\mathrm{E}\big(\tau^b\big))^2}{a^2\mathrm{E}(\tau^{2b})} = \frac{\mathrm{E}\big(N^2\big) - \mathrm{E}(N)}{\mathrm{E}(N)^2} \cdot \frac{\mathrm{E}\big(\tau^b\big)^2}{\mathrm{E}(\tau^{2b})} \tag{25}$$

The variance of the unobserved effects can be then computed as

$$\mathrm{Var}(Z) = E[Z^2] - 1 = \frac{\mathrm{E}\big(N^2\big) - \mathrm{E}(N)}{\mathrm{E}(N)^2} \cdot \frac{\mathrm{E}\big(\tau^b\big)^2}{\mathrm{E}(\tau^{2b})} - 1 \tag{26}$$

Using the empirical form of the expectation in the previous expression gives the estimator $\widehat{\mathrm{Var}}(Z)$ of the variance of the unobserved effects.
Natural choice for the estimator of the variance of unobserved effects is the empirical variance of the estimated $\hat{z}_j$'s. Since the estimated $\hat{z}_j$'s are defined by (14), the empirical variance of the estimated $\hat{z}_j$'s, which will be denoted as $\widehat{\mathrm{Var}}(\hat{z}_j)$, estimates the variance $\mathrm{Var}\big(\frac{N}{a\tau^b}\big)$, which can be computed as

$$\mathrm{Var}\left(\frac{N}{a\tau^b}\right) = \mathrm{E}\left(\mathrm{Var}\left(\frac{N}{a\tau^b}\middle| Z,\tau\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\frac{N}{a\tau^b}\middle| Z,\tau\right)\right) = \frac{\mathrm{E}\big(\tau^{-b}\big)}{a} + \mathrm{Var}(Z) \tag{27}$$

That means, that the empirical variance of the estimated $\hat{z}_j$'s overestimates the true variance of the unobserved effects by the factor $\frac{\mathrm{E}\big(\tau^{-b}\big)}{a}$. The formula (27) defines another estimator of the unobserved effects (with use of the derived estimators and empirical mean and variance).

2.1 Special case with all τ_j's equal

Let us consider the special case of the process described above, which is often used in real life applications and which is also often discussed in the literature, and in which $\tau_j = \tau$ for all $j = 1, \ldots, m$, i.e. each process is observed for the same given time τ.
Using the same notation as introduced above, the estimators of the parameters of this process are then given by

$$\hat{a} = \frac{n}{m\tau^{\hat{b}}} \tag{28}$$

$$\hat{b} = \frac{n}{n\log(\tau) - S} \tag{29}$$

The variance of these estimators can be estimated as

$$\widehat{\mathrm{Var}}(\hat{a}) = \frac{\hat{a}^2}{n} + \hat{a}^2\mathrm{Var}(\hat{b})\log(\tau)^2 + \hat{a}^2\frac{\widehat{\mathrm{Var}}(Z)}{m} \tag{30}$$

$$\widehat{\mathrm{Var}}\big(\hat{b}\big) = \frac{\hat{b}^2}{n} \tag{31}$$

The individual frailties can be estimated as $\hat{z}_j = \frac{n_j}{\frac{n}{m}}$.
Note, that the classical model, i.e. model without any unobserved heterogeneity, can be easily obtained from (2) by putting all z_j's equal to 1. In general, it is not possible to compute the analytical expressions for the estimators of the parameters a and b in this model, but in the special case in which $\tau_j = \tau$ for all $j = 1, \ldots, m$, the analytical estimators can be obtained and they are identical to the estimators (28) and (29). The estimated variance of the estimator $\hat{b}$ is also the same as the corresponding estimated asymptotic variance (31) but the estimated asymptotic variance of the estimator $\hat{a}$ is smaller by the amount $\hat{a}^2\frac{\widehat{\mathrm{Var}}(Z)}{m}$ than the estimated asymptotic variance (30). That means, that if there is unobserved heterogeneity present ($\mathrm{Var}(Z) > 0$), which is often the case in real data, then the wrong asymptotic variance is obtained by fitting the model which does not consider unobserved heterogeneity, which can lead to wrong conclusions about the

value of the estimated parameter. Therefore in this case will also be observed differences in the variance computed by the inversion of the observed Fisher information matrix and the variance computed by some other method, e.g. by bootstrapping.
The likelihood in the traditional frailty models for the jth process is the same as (1) but the distribution of the z_j's, usually described by one parameter representing the variance of the unobserved effects, is specified. One of the most commonly used models for modeling the unobserved heterogeneity is the gamma frailty model, i.e. z_j's are assumed to be distributed as a gamma random variable with mean 1 and variance α. Since the z_j's are unobservable, they have to be integrated out from the likelihood (1) by taking expectation with respect to the frailty variable. As can be easily checked, it is generally not possible to compute the analytical expressions for the estimators of the unknown parameters. Nevertheless, there exist analytical solution to the special case mentioned above, in which $\tau_j = \tau$ for all $j = 1, \ldots, m$. The estimators of the unknown parameters a and b in this case are the same as the estimators (28) and (29) respectively. The estimated asymptotic variances of the estimators $\hat{a}$ and $\hat{b}$ correspond to estimated variances (30) and (31) respectively (where $\widehat{\mathrm{Var}}(Z)$ is substituted by the estimator of the parameter α). Note, that it is also possible to compute the estimates of individual frailties z_j's in the gamma frailty model, $\hat{z}_{g,j} = \frac{n_j + \frac{1}{\hat{\alpha}}}{\frac{n}{m} + \frac{1}{\hat{\alpha}}}, j = 1, \ldots, m$. Note, that $\frac{1}{m} \sum_{j=1}^{m} \hat{z}_{g,j} = 1$.

3 Conclusions

A new approach to unobserved heterogeneity in NHPP, which does not make restrictive assumptions about the distribution of the unobserved effects, was introduced in this paper in an example with the power law process. This approach can easily be generalized to a general NHPP. The approach allowed furthermore to compute analytical expressions for the estimators of the parameters of the process as well as for standard errors in general case. The classical and gamma frailty model, on the other hand, have to be solved only numerically. It is also possible to estimate the individual frailties and the variance of the unobserved effects It was also shown, that if the classical model is used for the data with unobserved heterogeneity, then although correct estimators can be obtained, the wrong variance of these estimators is computed by the the commonly used method of negative inversion of the hessian of the loglikelihood function, which can lead to wrong conclusions about the true values of these parameters.

References

[1] O. Aalen, O. Borgan, and H. Gjessing. *Survival and event history analysis: a process point of view.* Springer, 2008.

[2] J. F. Lawless. Regression methods for poisson process data. *Journal of the American Statistical Association*, 82(399):808–815, 1987.