Journal of KONBiN 5 (8) 2008
ISSN 1895-8281

STATISTICAL ANALYSIS OF RELIABILITY FIELD DATA WHEN WORKING CONDITIONS ARE IMPRECISELY REPORTED

STATYSTYCZNA ANALIZA DANYCH NIEZAWODNOŚCIOWYCH W PRZYPADKU NIEPRECYZYJNIE OKREŚLONYCH WARUNKÓW EKSPLOATACJI

Olgierd HRYNIEWICZ[1]

(1) **Systems Research Institute of PAS**
ul. Newelska 6, 01-447 Warszawa
E-mail: hryniewi@ibspan.waw.pl

Abstract: In contrast to laboratory lifetime tests reliability field tests are usually performed in conditions which vary in time in a random way. We consider the case when users are asked about their description of their vague perceptions of the usage conditions. In the paper we use interval-valued variables for the description of imprecisely known test conditions that may be used as covariates in classical reliability models. We present a simple approximate algorithm for a proportional hazard lifetime model.

Keywords: field reliability data, interval-valued data, proportional hazard

Streszczenie. W przeciwieństwie do badań niezawodnościowych prowadzonych w warunkach laboratoryjnych badania prowadzone w warunkach normalnej eksploatacji prowadzone są w warunkach, które w sposób losowy zmieniają się w czasie. W pracy rozpatrywany jest przypadek, gdy warunki eksploatacji opisane w sposób nieprecyzyjny w postaci przedziałowej. Użyte do tego celu zmienne przedziałowe zostały zastosowane jako zmienne stowarzyszone w klasycznych modelach niezawodnościowych typu proporcjonalnego hazardu.

Słowa kluczowe: dane z eksploatacji, dane przedziałowe, proporcjonalny hazard

1. Introduction

In classical textbooks on lifetime data it is usually assumed that all data are acquired from precisely described tests such as e.g. laboratory tests. Statistical analysis of reliability data acquired from such tests is relatively simple, especially when the test is based on either type-I censoring or type-II censoring schemes. Pertinent statistical methods have been described in numerous papers and textbooks. The situation is more complicated when tests of individual items are performed in different conditions, as in e.g. accelerated life tests. Mathematical models that are useful for the description of such lifetime data are well known but not so popular. We present some of them in the second section of this paper.

When lifetime data are collected from field experiments performed in real conditions the situation is much more complicated. We face in this case imprecise information of different kind. For example, failures are reported with unknown delay and their precise values are not known. Also the working condition may be varying in time in a way which precludes their precise description. All these uncertainties make the analysis of real reliability field data prohibitively complicated. Precise description in terms of probability distributions requires many simplifying assumptions which usually cannot be verified. In (Hryniewicz, 2007) we have proposed an alternative method for modeling imprecise information by using fuzzy sets. The resulting model is fuzzy random as we merge information of random and fuzzy character. In the third section of this paper we present a method for the analysis of lifetime data when information about working conditions is imprecise and is described in terms of intervals. The results presented in that section can be easily generalized to the case of fuzzy-valued imprecise information. As an example, we consider the case of the proportional hazard model with interval-valued covariates. This model in the case of the Weibull distribution is equivalent to the log-location-scale model which is often used in the description of reliability data (e.g. from accelerated life tests).

2. Mathematical model in case of precise information about working conditions

Mathematical models of lifetimes have been developed since the early 1950s. For example, the general mathematical model of lifetime data was presented in (Hu and Lawless, 1996). Following their proposal we consider population $\mathcal{P}$ consisting of n units described by their lifetimes, $t_i, i = 1,\ldots n$, random censoring times, $\tau_i, i = 1,\ldots n$, and q-dimensional vectors of covariates $\mathbf{z}_i, i = 1,\ldots n$, respectively. Triplets $(t_i, \tau_i, \mathbf{z}_i)$ are the realizations of a random sample from a distribution with the joint probability function

$$f(t \mid \theta;\tau,\mathbf{z})dG(\tau,\mathbf{z}), t>0,\tau>0,\mathbf{z}\in R^q, \tag{1}$$

where lifetimes and censoring times are usually considered independent given fixed **z**, and $G(\tau,\mathbf{z})$ is an arbitrary cumulative distribution function. Let *O* be the set of *m* units for whom the lifetimes are observed, i.e. for whom $t_i \le \tau_i, i=1,\ldots,n$. The remaining *n-m* units belong to the set *C* of censored lifetimes for whom only their censoring times τ_i and covariates $\mathbf{z}_i$ are known. The function $S(t \mid \theta;\tau,\mathbf{z})=1-F(t \mid \theta;\tau,\mathbf{z})$, where $F(t \mid \theta;\tau,\mathbf{z})$ is the cumulative distribution function of the lifetime, is called in the literature the survivor function or the survival function. The likelihood function that describes the lifetime data is now given in (Hu and Lawless, 1996)

$$L(\boldsymbol{\theta})=\prod_{i\in O} f(t_i \mid \boldsymbol{\theta};\tau_i,\mathbf{z}_i)dG(\tau_i,\mathbf{z}_i)\times\prod_{i\in C} S(t_i \mid \boldsymbol{\theta};\tau_i,\mathbf{z}_i)dG(\tau_i,\mathbf{z}_i) \tag{2}$$

Many other specific models which are comprehensively described in reliability textbooks may be considered as special cases of this general model. Below, we briefly present two families of lifetime data models which are the special cases of (2) and well suited for the description of lifetime tests in field conditions, namely proportional hazard models, and location-scale regression models.

In case of the proportional hazard models, the hazard function, defined as $h(t;\boldsymbol{\theta},\mathbf{z})= f(t;\boldsymbol{\theta},\mathbf{z})/ S(t;\boldsymbol{\theta},\mathbf{z})$, is linked to the test conditions by the following equation

$$h(t \mid \mathbf{z})= h_0(t)g(\mathbf{z}) \tag{3}$$

where $h_0(.)$ is the baseline hazard function which may depend on some unknown parameters and $g(.)$ links reliability with some external variables (covariates) that may describe working conditions. Parameters of these functions have to be estimated from statistical data. Another representation of the proportional hazard model is the following:

$$S(t \mid \mathbf{z})= S_0(t)^{g(\mathbf{z})}. \tag{4}$$

One of the special cases of (3) which is the most frequently used in practice was proposed in (Cox, 1972) and is given by the following general expression

$$h(t \mid \boldsymbol{\theta}, \mathbf{z}) = h_0(t, \boldsymbol{\theta}) e^{\mathbf{z}\boldsymbol{\beta}}, \tag{5}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters, $\mathbf{z}\boldsymbol{\beta} = z_1\beta_1 + \cdots + z_q\beta_q$, and $\beta_1, \ldots, \beta_q$ are unknown regression coefficients. This model was investigated by many authors, and its most comprehensive description can be found in (Lawless, 2003). In case of the Weibull distribution of lifetimes the survivor function is given by the following expression

$$S(t \mid \mathbf{z}) = exp\left[-\left(te^{-\mathbf{z}\boldsymbol{\beta}} \right)^{\delta} \right], \tag{6}$$

where δ>0 is the shape parameter, responsible for the description of the type of failure processes. If we use the transformation $Y = log\, T$, the logarithms of lifetimes are described by a simple linear model

$$Y = \mathbf{z}\boldsymbol{\beta} + \sigma W, \tag{7}$$

where σ=1/δ, and the random variable W is distributed according to the standard extreme value distribution (The Gumbel distribution) with the probability density function $exp[w - exp(w)]$.

When n test units are observed, and independent observations $(y_i, \mathbf{z}_i), i = 1, \ldots, n$ are available, where y_i is either logarithm of lifetime or logarithm of censoring time of the i-th unit, the maximum likelihood estimators of the parameters describing the model can be found by solving the following set of equations (Lawless, 1982):

$$-\frac{1}{\sigma}\sum_{i \in o} z_{il} + \frac{1}{\sigma}\sum_{i=1}^{n} z_{il} e^{x_i} = 0, l = 1, \ldots, q \tag{8}$$

$$-\frac{r}{\sigma} - \frac{1}{\sigma}\sum_{i \in O} x_i + \frac{1}{\sigma}\sum_{i=1}^{n} x_i e^{x_i} = 0, \tag{9}$$

where $x_i = (y_i - \mathbf{z}_i\boldsymbol{\beta})/\sigma$. The solution of q+1 equations given by (8) and (9) yields the maximum likelihood estimators of σ (and hence for the shape parameter δ), and regression coefficients $\beta_1, \ldots, \beta_q$. The formulae for the calculation of the asymptotic covariance matrix of these estimators can be found in (Lawless, 1982).

A second regression model commonly used for the analysis of lifetimes is the location-scale model for the logarithm of lifetime T, also known as the log-location-scale model. In this model the random variable $Y = log\, T$ has a distribution with the location parameter $\mu(\mathbf{z})$, and a scale parameter σ which does

not depend upon the covariates **z**. This model can be expressed as follows:

$$Y = \mu(\mathbf{z}) + \sigma\xi, \tag{10}$$

where $\sigma > 0$ and ξ is a random variable with a distribution that is independent on **z**. Alternative representation of this model can be written as

$$S(t \mid \mathbf{z}) = S_0\left(\frac{t}{\mu(\mathbf{z})}\right). \tag{11}$$

Both families of models, i.e. proportional hazard models and log-location-scale models, have been applied for different probability distributions of lifetimes. The detailed description of those results can be found, for example, in (Lawless, 2003). However, it is worth to note, that only in the case of the Weibull distribution (and the exponential distribution, which is a special case of the Weibull distribution) both models coincide. In such a case the parameters of the model can be found from equations (8) – (9). Similar equations for a general case of any probability distribution can be found in (Lawless, 2003).

When the type of the lifetime probability distribution is not known and the proportional hazards model seems to be appropriate we can apply distribution-free methods for the analysis of lifetimes. Let us consider a special case of (4)

$$S(t \mid \mathbf{z}) = S_0(t)^{\mathbf{z}\boldsymbol{\beta}}. \tag{12}$$

A method for the separation of the estimation of the vector of regression coefficient $\boldsymbol{\beta}$ from the estimation of the survivor function $S_0(t)$ has been proposed in (Cox, 1972). Suppose that observed lifetimes are ordered as follows: $t_{(1)} < \cdots < t_{(m)}$. Let $R_i = R(t_{(i)})$ be the set of all units being at risk at time $t_{(i)}$, that is the set of all non-failed and uncensored units just prior to $t_{(i)}$. Note, that in this model censoring times of the remaining $n - m$ units may take arbitrary values. For the estimation of $\boldsymbol{\beta}$, Cox proposed to use a pseudo-likelihood function given by (Cox, 1972)

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m}\left(e^{\mathbf{z}_{(i)}\boldsymbol{\beta}} \Big/ \sum_{l \in R_i} e^{\mathbf{z}_{(i)}\boldsymbol{\beta}}\right) \tag{13}$$

Slight modification of (13) has been proposed in (Lawless, 1982). This modification allows for few multiple failures at times $t_{(i)}, i = 1,\ldots,m$. Formulae for the calculation of the asymptotic covariance matrix of the estimators of $\beta_1,\ldots,\beta_q$

are given in (Lawless, 1982). When the vector of the regression coefficients $\boldsymbol{\beta}$ has been estimated, we can use a distribution-free methods, such as Kaplan-Meier, Breslow or generalized Nelson - Aalen estimators (for more information, see (Lawless, 2003)), for the estimation $S_0(t)$.

3. Mathematical model of lifetime data in case of imprecise information about working conditions

In the models presented in the previous section we assume that the values of covariates **z** are precisely known. Even in this simplest case the analysis of real field data is relatively difficult. In reality, however, the situation is much more complicated. Usually, working conditions are varying time in a random way, and their precise description becomes very difficult (see, e.g. models described in (Lawless, 2003)) or even mathematically intractable. Therefore, there is a need to propose approximate methods that should be simple enough in order to be applied in practice.

In the simplest case the existing partial knowledge about the values of working conditions, described by the vector of covariates **z**, can be presented in terms of intervals representing the values of considered characteristics or quantities. In order to simplify further notation let us denote by Δz a compact interval $[z_{min},z_{max}]$.

Let us consider the case when lifetime data can be described by the proportional hazard model. To be more precise, let us assume that this model has the form proposed by Cox, i.e. the hazard function in this case is given by (6). The log-likelihood function in this case can be expressed as follows

$$l(\boldsymbol{\theta},\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[log\, h_0(t_i;\boldsymbol{\theta}) + \boldsymbol{\beta}'\mathbf{z}_i \right] - \sum_{i=1}^{n} H_0(t_i;\boldsymbol{\theta}) exp(\boldsymbol{\beta}'\mathbf{z}_i). \qquad (14)$$

where δ_i is equal to 1 when we observe a failure at t_i or 0 when t_i is a censoring time. In case of precise information about the times to failures, censoring times, and values of covariates the estimates of the unknown parameters $(\boldsymbol{\theta},\boldsymbol{\beta})$ can be found by maximization of (14). However, in case of imprecise information about covariates $\mathbf{z}_i$ the results of maximization are not unequivocal anymore. The maximum likelihood estimators of $(\boldsymbol{\theta},\boldsymbol{\beta})$ are in this case given as multivariate *intervals* obtained as the solutions of the following optimization problems:

$$(\boldsymbol{\theta}_{min},\boldsymbol{\beta}_{min}) = \inf_{\mathbf{z}_i \in \Delta \mathbf{z}_i} arg\, \max_{(\boldsymbol{\theta},\boldsymbol{\beta})} l(\boldsymbol{\theta},\boldsymbol{\beta}) \qquad (15)$$

$$\left(\boldsymbol{\theta}_{max},\boldsymbol{\beta}_{max}\right)=\sup_{\mathbf{z}_i\in\Delta\mathbf{z}_i}\arg\max_{(\boldsymbol{\theta},\boldsymbol{\beta})} l\left(\boldsymbol{\theta},\boldsymbol{\beta}\right), \tag{16}$$

where $l\left(\boldsymbol{\theta},\boldsymbol{\beta}\right)$ is the log-likelihood function given by (14). The optimization problem defined by (17) – (18) may be, in a general case, difficult, as the interval computations of nonlinear functions are usually time consuming. However, if we use the partial likelihood function (13) the optimization procedure may be simplified.

Consider the partial likelihood function given by (13) as a function of covariates. The partial derivatives with respect to *k*-th element of the vector of covariates **z** are expressed as follows:

$$\frac{\partial \ln L\left(\boldsymbol{\beta}\right)}{\partial z_{(i),k}}=\frac{\partial l}{\partial z_{(i),k}}\sum_{i=1}^{n}\left[\mathbf{z}_{(i)}\boldsymbol{\beta}-\ln\sum_{l\in R_{(i)}}e^{\mathbf{z}_{(l)}\boldsymbol{\beta}}\right]=\beta_k\left[\frac{\sum_{l\in R_{(i)}}e^{\mathbf{z}_{(l)}\boldsymbol{\beta}}-e^{\mathbf{z}_{(i)}\boldsymbol{\beta}}}{\sum_{l\in R_{(i)}}e^{\mathbf{z}_{(l)}\boldsymbol{\beta}}}\right] \tag{17}$$

The sign of (17) is the same as the sign of β_k. Hence, for a covariate described by a positive regression coefficient the maximum value of (13) is attained for the maximal possible value of this covariate. In case of negative coefficients this maximum is attained for the minimal possible value of the covariate. The conditions for obtaining a minimum value of the likelihood function are just opposite. This result suggests a simple algorithm for finding the interval estimate of $\boldsymbol{\beta}$**.** We can start with any set of values $z_{(i),k}, i=1,\ldots,n, k=1,\ldots,q$ such that $z_{(i),k}\in\left[z_{(i),k,min},z_{(i),k,max}\right]$, and calculate initial estimators of $\beta_1,\ldots,\beta_q$. Then, depending of the signs of the elements of this vector we replace the values of $z_{(i),k}$ with their respective minimal or maximal values. For these new values of covariates we find upper limits for the interval estimates of $\beta_1,\ldots,\beta_q$. If the signs of the elements of this vector have not been changed the estimation procedure stops. Otherwise, we continue this iterative procedure. When the procedure does not converge after a few steps we have to use a general estimation procedure. The same is repeated, with appropriate changes of the procedure, for the lower limits.

After having obtained the interval estimates of the regression coefficients $\beta_1,\ldots,\beta_q$ we can estimate the remaining parameters of the model. For example, in case of the Weibull distribution we have to solve a nonlinear equation (9). Note, however, that the values x_i in (9) are now interval-valued. Therefore, the solution of

(9) with respect to σ has to bee interval-valued too. The exact solution is not simple, and requires extensive computations. However, reasonable approximations can be found if for the calculation of the lower limit of σ we will use lower limits of x_i that correspond to small values of times t_i, and upper limits of x_i that correspond to large values of times t_i. In case of the calculation of the upper limit of σ we should proceed in an opposite way.

4. Conclusion

Statistical analysis of reliability field data is quite complicated, as it is necessary to take into account different test conditions. These conditions are usually not well defined or vary in time in an unknown way. The methodology proposed in this paper let to take into account those uncertainties in the description of test conditions that cannot be described by precise probabilistic models.

References

1. Cox D.R. (1972). *Regression models and life tables (with discussion)*. Journal of the Royal Statistical Society, ser. B, v.34, p.187 – 202.
2. Hryniewicz O. (2007). *Statistical analysis of interval and imprecise data - applications in the analysis of reliability field data*. In: Proceedings of the conference: The First Summer Safety and Reliability Seminars 2007 - SSARS 2007, Gdańsk-Sopot, Poland, 22-29 July 2007, Gdynia Maritime University, p.181-192.
3. Hu J.X., Lawless J.F. (1996). *Estimation from truncated lifetime data with supplementary information on covariates and censoring times.* Biometrika, v.83, no.4, p.747-761.
4. Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, New York.
5. Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*. *Second Edition*. John Wiley and Sons, New York.

Prof. dr hab. Olgierd HRYNIEWICZ. Systems Research Institute of the Polish Academy of Sciences, Warsaw. Specializes in reliability, quality control, statistics and fuzzy sets. Author of more than 170 papers from these and related fields. His main area of interest is the statistical analysis of imprecisely reported data, and the optimization of quality and reliability test procedures.