Journal of KONBiN 2(5)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
Instytut Badań Systemowych PAN
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

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

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 [2] 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 by Hu and

Lawless [3]. 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 by [3]

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

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 by Cox [1] 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 [5]. 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 [4]:

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

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

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 [4].

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 [5]. 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 [5]

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)

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

Cox [1] proposed a method for the separation of the estimation of the vector of regression coefficient $\boldsymbol{\beta}$ from the estimation of the survivor function $S_0(t)$. 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 [1] proposed to use a pseudo-likelihood function given by

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

Slight modification of (13) has been proposed in Lawless [4]. 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 [4]. 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 [5]), 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 [5]) 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

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

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:

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

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

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.

References

[1] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, ser.B*, 34, 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 Summmer Safety and Reliability Seminars 2007 - SSARS 2007, Gdańsk-Sopot, Poland, 22-29 July 2007, Gdynia Maritime University, 181-192.

[3] Hu, J.X., Lawless, J.F. (1996a). Estimation from truncated lifetime data with supplementary information on covariates and censoring times. *Biometrika*, 83(4), 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.

STATYSTYCZNA ANALIZA DANYCH NIEZAWODNOŚCIOWYCH W PRZYPADKU NIEPRECYZYJNIE OKREŚLONYCH WARUNKÓW EKSPLOATACJI

1. Wstęp

W klasycznych podręcznikach poświęconych analizie danych niezawodnościowych zazwyczaj zakłada się, że wszystkie dane otrzymywane są z przeprowadzonych zgodnie ze z góry zadanym planem badań, takich jak np. badania laboratoryjne. Statystyczna analiza danych pochodzących z takich badań jest względnie prosta, zwłaszcza gdy mamy do czynienia z danymi cenzorowanymi typu I lub typu II. Sytuacja staje się znacznie bardziej skomplikowana, gdy badania poszczególnych obiektów prowadzone są w różnych warunkach, np. tak, jak ma to miejsce w przyspieszonych badaniach niezawodności. Modele matematyczne, które można wykorzystać do opisu wyników takich badań są znane, ale są niezbyt popularne wśród praktyków. Niektóre z nich zostaną przedstawione w drugiej części niniejszej pracy.

Gdy dane niezawodnościowe uzyskiwane są z badań eksploatacyjnych prowadzonych w rzeczywistych warunkach użytkowania sytuacja staje się jeszcze bardziej skomplikowana. Mamy wówczas do czynienia z niepewnymi informacjami różnego rodzaju. Na przykład, informacje o uszkodzeniach docierają z nieznanym opóźnieniem, a wobec tego rzeczywiste wartości czasów do uszkodzenia nie są znane. Także warunki eksploatacji mogą podlegać zmianom w czasie w sposób, który uniemożliwia ich precyzyjne określenie. Wszystkie te zjawiska powodują, że statystyczna analiza rzeczywistych danych niezawodnościowych staje się niezwykle skomplikowana. Precyzyjny opis takich danych przy użyciu rachunku prawdopodobieństwa wymaga przyjęcia wielu upraszczających założeń, które zazwyczaj nie podlegają weryfikacji. W pracy [2] zaproponowana została alternatywna metoda opisu nieprecyzyjnie określonych danych niezawodnościowych, w której wykorzystano teorię zbiorów rozmytych. Zaproponowany w tej pracy model jest modelem rozmyto-losowym, gdyż opisuje on jednocześnie zjawiska o charakterze losowym oraz zjawiska o charakterze rozmytym. W trzeciej części niniejszej pracy przedstawiamy metodę analizy danych niezawodnościowych dla przypadków, gdy informacja o warunkach eksploatacji wyrażana jest w sposób nieprecyzyjny w postaci pewnych przedziałów możliwych wartości. Przedstawione tam wyniki teoretyczne mogą być w prosty sposób uogólnione na przypadek, gdy nieprecyzyjnie określoną informację można modelować przy pomocy zbiorów rozmytych. Jako przykład analizujemy

przypadek modelu proporcjonalnego hazardu ze zmiennymi towarzyszącymi o wartościach podanych w postaci przedziałów. Taki model w przypadku wykorzystania rozkładu Weibulla jest równoważny logarytmicznemu modelowi typu położenie-skala, który jest często wykorzystywany do opisu danych niezawodnościowych (np. pochodzących z przyspieszonych badań niezawodności).

2. Model matematyczny dla przypadku precyzyjnie określonych warunków eksploatacji

Matematyczne modele opisujące czasy do uszkodzenia były rozwijane od wczesnych lat pięćdziesiątych ubiegłego stulecia. Na przykład, jeden z najbardziej ogólnych takich modeli został zaproponowany w pracy Hu i Lawlessa [3]. Zgodnie z ich propozycją rozpatrzmy zbiór $\mathcal{P}$ składający się z *n* jednostek opisanych odpowiednio czasami poprawnej pracy (czasami do uszkodzenia), $t_i, i = 1,\ldots n$, losowymi czasami cenzorowania, $\tau_i, i = 1,\ldots n$ oraz *q*-wymiarowymi wektorami zmiennych towarzyszących $\mathbf{z}_i, i = 1,\ldots n$. Trójki $(t_i, \tau_i, \mathbf{z}_i)$ są realizacjami próbki losowej opisanej rozkładem prawdopodobieństwa danym łączną funkcją gęstości

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

gdzie o czasach poprawnej pracy oraz czasach cenzorowania zazwyczaj zakłada się, że są wzajemnie niezależne dla ustalonej wartości wektora **z**, a $G(\tau, \mathbf{z})$ jest dowolną dystrybuantą. Niech *O* oznacza zbiór *m* jednostek, dla których zaobserwowano czasy uszkodzeń, tzn. dla których zachodzi $t_i \le \tau_i, i = 1,\ldots,n$. Pozostałe *n-m* jednostek należy do zbioru *C* jednostek podległych cenzorowaniu, dla których znane są jedynie czasy cenzorowania τ_i oraz zmienne towarzyszące $\mathbf{z}_i$. Funkcja $S(t \mid \theta; \tau, \mathbf{z}) = 1 - F(t \mid \theta; \tau, \mathbf{z})$, gdzie $F(t \mid \theta; \tau, \mathbf{z})$ jest dystrybuantą czasu poprawnej pracy nazywana jest w literaturze funkcją przeżycia lub funkcją niezawodności. Funkcja wiarogodności opisująca dane niezawodnościowe jest wobec tego dana wzorem [3]

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

Wiele innych szczególnych modeli, które są szeroko opisywane w podręcznikach niezawodności może być rozpatrywanych jako przypadki tego ogólnego modelu. Poniżej przedstawimy dwie rodziny modeli opisujących dane niezawodnościowe, które są szczególnymi przypadkami modelu (2) i dobrze nadają się do opisu danych niezawodnościowych uzyskanych w wyniku przeprowadzenia badań w warunkach normalnej eksploatacji. Rozpatrzymy mianowicie modele proporcjonalnego hazardu oraz regresyjne modele typu położenie – skala.

W przypadku modeli proporcjonalnego hazardu funkcja hazardu (zwana też funkcją intensywności uszkodzeń), definiowana zależnością $h(t;\boldsymbol{\theta},\mathbf{z}) = f(t;\boldsymbol{\theta},\mathbf{z})/S(t;\boldsymbol{\theta},\mathbf{z})$, jest powiązana z warunkami, w których przeprowadzano badanie przy pomocy następującej zależności

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

gdzie $h_0(.)$ jest bazową funkcją hazardu, która może zależeć od pewnych nieznanych parametrów, a funkcja $g(.)$ opisuje zależność niezawodności od pewnych zmiennych zewnętrznych (zmiennych towarzyszących), które mogą opisywać warunki eksploatacji. Parametry tych funkcji należy wyestymować na podstawie danych statystycznych. Inną formą przedstawienia modelu proporcjonalnego hazardu jest

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

Jeden ze szczególnych przypadków (3), który jest najczęściej stosowany w praktyce, został zaproponowany przez Coxa [1], jest opisany zależnością

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

gdzie $\boldsymbol{\theta}$ jest wektorem nieznanych parametrów, $\mathbf{z}\boldsymbol{\beta} = z_1\beta_1 + \cdots + z_q\beta_q$, zaś $\beta_1,\ldots,\beta_q$ są nieznanymi współczynnikami modelu regresji. Model ten był badany przez wielu autorów, a jego wyczerpujący opis można znaleźć w pracy Lawlessa [5]. W szczególnym przypadku, gdy czasy do uszkodzenia opisane są rozkładem Weibulla, funkcja przeżycia dana jest następującym wzorem:

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

gdzie δ>0 jest parametrem kształtu, opisującym rodzaj procesu uszkodzeń. Jeżeli zastosujemy przekształcenie $Y = log\,T$, to logarytmy czasów poprawnej pracy opisane są przy pomocy prostego modelu liniowego

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

gdzie σ=1/δ, a zmienna losowa W ma rozkład wartości skrajnych (rozkład Gumbela) opisany funkcją gęstości prawdopodobieństwa $exp[w - exp(w)]$.

Jeśli obserwowanych jest *n* badanych obiektów i dysponujemy informacjami o niezależnych obserwacjach $(y_i, \mathbf{z}_i), i = 1,\ldots,n$, gdzie y_i jest albo logarytmem czasu poprawnej pracy albo logarytmem czasu cenzorowania *i*-tego obiektu, to estymatory największej wiarogodności parametrów rozpatrywanego modelu mogą być znalezione jako rozwiązania następującego układu równań [4]:

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

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

gdzie $x_i = (y_i - \mathbf{z}_i\boldsymbol{\beta})/\sigma$. W wyniku rozwiązania układu *q*+1 równań danych zależnościami (8) i (9) uzyskujemy oszacowania estymatorów największej wiarogodności parametru σ (a stąd parametru kształtu δ) oraz współczynników regresji $\beta_1,\ldots,\beta_q$. Wzory pozwalające obliczyć asymptotyczną macierz kowariancji tych estymatorów można znaleźć w pracy Lawlessa [4].

Drugim modelem regresyjnym, który powszechnie wykorzystywany jest w analizie danych niezawodnościowych jest model typu położenie-skala dla logarytmu czasu poprawnej pracy *T*, znany także jako logarytmiczny model typu położenie-skala. W przypadku tego modelu zmienna losowa $Y = \log T$ ma rozkład prawdopodobieństwa o parametrze położenia $\mu(\mathbf{z})$ i parametrze skali σ, który nie zależy od wartości zmiennych towarzyszących **z**. Model ten opisany jest następującą zależnością:

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

gdzie $\sigma > 0$, a ξ jest zmienną losową o rozkładzie niezależnym od **z**. Alternatywny zapis tego modelu jest następujący

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

Obie przedstawione powyżej rodziny modeli, tzn. modele proporcjonalnego hazardu oraz logarytmiczne modele typu położenie – skala znalazły zastosowanie dla różnych rozkładów prawdopodobieństwa czasów poprawnej pracy. Szczegółowy opis tych zastosowań można znaleźć na przykład w pracy [5]. Warto przy okazji podkreślić, że tylko w przypadku rozkładu Weibulla (a także rozkładu wykładniczego, który jest szczególnym przypadkiem rozkładu Weibulla) oba

modele mają tę samą postać. W takim przypadku parametry modelu mogą być znajdywane w wyniku rozwiązania równań (8) – (9). Podobne równania dla ogólnego przypadku, dotyczącego dowolnego rozkładu prawdopodobieństwa, podane są w pracy [5].

Gdy typ rozkładu czasu poprawnej pracy nie jest znany, a model proporcjonalnego hazardu wydaje się właściwy, możemy stosować nieparametryczne metody analizy danych niezawodnościowych. Rozpatrzmy następujący szczególny przypadek zależności (4)

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

Cox [1] zaproponował metodę, w której estymatory wektora współczynników regresji $\boldsymbol{\beta}$ znajduje się niezależnie od estymatora funkcji przeżycia $S_0(t)$. Przyjmijmy, że zaobserwowane czasy poprawnej pracy można uporządkować w następujący sposób: $t_{(1)} < \cdots < t_{(m)}$. Niech $R_i = R(t_{(i)})$ będzie zbiorem wszystkich badanych jednostek, o których mówimy, że "podlegają ryzyku" w momencie $t_{(i)}$. Jest to zbiór wszystkich jednostek, które nie uszkodziły się lub nie podlegały cenzorowaniu w chwili bezpośrednio poprzedzającej czas $t_{(i)}$. Zauważmy, że w tym modelu czasy cenzorowania pozostałych $n - m$ jednostek mogą przyjmować dowolne wartości. Do estymacji wektora $\boldsymbol{\beta}$ Cox [1] zaproponował wykorzystanie funkcji pseudo-wiarogodności, zwanej także cząstkową funkcją wiarogodności, daną wzorem

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

Niewielka modyfikacja wzoru (13) została też zaproponowana w książce Lawlessa [4]. W modyfikacji tej dopuszcza się występowanie kilku uszkodzeń (lub cezorowań) wielokrotnych w chwilach $t_{(i)}, i = 1, \ldots, m$. Wzory pozwalające wyznaczyć asymptotyczną macierz kowariancji estymatorów parametrów $\beta_1, \ldots, \beta_q$ podane są w pracy [4]. Gdy wektor współczynników regresji $\boldsymbol{\beta}$ zostanie wyestymowany, możemy skorzystać z nieparametrycznych estymatorów funkcji przeżycia, takich jak estymator Kaplana-Meiera, estymator Breslowa lub uogólniony estymator Nelsona – Aalena (więcej informacji na ten temat można znaleźć w [5]) i wyznaczyć ocenę bazowej funkcji przeżycia $S_0(t)$.

3. Matematyczne modele niezawodności w przypadku nieprecyzyjnych informacji o warunkach eksploatacji

W modelach przedstawionych w poprzedniej części zakładaliśmy, że wartości zmiennych towarzyszących **z** są dokładnie znane. Okazuje się, że nawet w tym najprostszym przypadku statystyczna analiza rzeczywistych danych pochodzących z eksploatacji jest trudna. W rzeczywistość, niestety, sytuacja jest jeszcze bardziej skomplikowana. Warunki eksploatacji zmieniają się zazwyczaj w czasie w sposób losowy, a ich precyzyjny opis staje się niezwykle trudny (patrz np. modele opisane w [5] lub nawet niemożliwy. Wynika stąd potrzeba zaproponowania metod przybliżonych, które byłyby wystarczająco proste, tak by dały się zastosować w praktyce.

W najprostszym przypadku istniejąca częściowa wiedza o wartościach parametrów opisujących warunki pracy badanych obiektów, przedstawiona przy pomocy wektora zmiennych towarzyszących **z,** może być zapisana w postaci przedziałów wartości rozpatrywanych charakterystyk lub innych wielkości. W celu uproszczenia zapisu będziemy w dalszej części pracy oznaczali przez Δz zwarty przedział $[z_{min},z_{max}]$.

Rozpatrzmy teraz przypadek, gdy dane niezawodnościowe opisane są modelem proporcjonalnego hazardu. Dokładniej, będziemy zakładać, że model ten ma postać zaproponowana przez Coxa, tzn. że funkcja hazardu jest w tym przypadku dana zależnością (6). Logarytmiczna funkcja wiarogodności jest w tym przypadku następująca:

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

gdzie δ_i jest równe 1 gdy w chwili t_i obserwujemy uszkodzenie i 0, gdy czas t_i jest czasem cenzorowania. W przypadku dokładnej informacji o czasach poprawnej pracy, czasach cenzorowania oraz wartościach zmiennych towarzyszących estymatory nieznanych parametrów $(\boldsymbol{\theta},\boldsymbol{\beta})$ mogą być znajdowane w drodze maksymalizacji wyrażenia (14). Jednakże w przypadku istnienia nieprecyzyjnej informacji o wartościach zmiennych towarzyszących $\mathbf{z}_i$ rezultaty takiej maksymalizacji przestają być jednoznaczne. Estymatory największej wiarogodności parametrów $(\boldsymbol{\theta},\boldsymbol{\beta})$ są w takim przypadku wielowymiarowymi *przedziałami* uzyskanymi jako rozwiązania następujących zagadnień optymalizacyjnych:

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

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

gdzie $l(\boldsymbol{\theta},\boldsymbol{\beta})$ jest logarytmiczną funkcja wiarogodności dana wzorem (14).

Zagadnienie optymalizacji zdefiniowane zależnościami (17) – (18) może być w ogólnym przypadku trudne, gdyż obliczenia przedziałowe są w przypadku funkcji nieliniowych złożone i wymagają dużych nakładów obliczeniowych. Jednakże w przypadku wykorzystania cząstkowej funkcji wiarogodności (13) procedura optymalizacyjna może być znacznie uproszczona.

Rozpatrzmy cząstkową funkcję wiarogodności daną wzorem (13) jako funkcję zmiennych towarzyszących. Pochodne cząstkowe tej funkcji, obliczone względem *k*-tego elementu wektora zmiennych towarzyszących **z,** są wyrażone w następujący sposób:

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

Zauważmy, że znak (17) jest taki sam jak znak β_k. Stąd w przypadku zmiennej towarzyszącej opisanej dodatnim współczynnikiem regresji, maksymalną wartość wyrażenia (13) uzyskuje się w przypadku możliwie maksymalnej wartości tej zmiennej. W przypadku ujemnego współczynnika regresji taka wartość maksymalna uzyskiwana jest dla minimalnej możliwej wartości zmiennej towarzyszącej. Warunki, dla których uzyskuje się minimalne wartości funkcji wiarogodności są dokładnie przeciwne. Wynik ten sugeruje wykorzystanie prostego algorytmu mającego na celu znajdywanie przedziałowych wartości estymatorów parametrów **β.** Obliczenia możemy rozpocząć przyjmując dowolny zbiór wartości $z_{(i),k}, i = 1,\ldots,n, k = 1,\ldots,q$ takich, że $z_{(i),k} \in \left[z_{(i),k,min}, z_{(i),k,max}\right]$, a następnie wyznaczyć początkowe wartości estymatorów $\beta_1,\ldots,\beta_q$. Następnie w zależności od znaków elementów tego wektora należy zamienić wartości $z_{(i),k}$ ich, odpowiednio, minimalnymi lub maksymalnymi wartościami. Dla tych nowych wartości zmiennych towarzyszących znajdujemy górne ograniczenia przedziałowych estymatorów parametrów $\beta_1,\ldots,\beta_q$. Jeżeli znaki elementów tego wektora nie uległy zmianie, to procedurę zatrzymujemy. W przeciwnym przypadku powyższą procedurę iteracyjną kontynuujemy. Gdy po kilku iteracjach procedura obliczeniowa nie jest zbieżna musimy zastosować ogólną procedurę estymacji. Podobna procedura, po zastosowaniu odpowiednich modyfikacji, może być stosowana do wyznaczenia dolnych granic przedziałów wartości współczynników regresji.

Po wyznaczeniu przedziałowych estymatorów współczynników regresji $\beta_1,\ldots,\beta_q$ możemy estymować pozostałe parametry modelu. Na przykład, w przypadku rozkładu Weibulla musimy rozwiązać nieliniowe równanie (9). Zauważmy jednak, że w takim przypadku występujące w równaniu (9) wartości x_i mają teraz wartości przedziałowe. Wynika stąd, że rozwiązanie równania (9) ze względu na σ musi być także w postaci przedziałowej. Wyznaczenie takiego dokładnego rozwiązania nie jest proste i może wymagać wielu obliczeń. Jednakże rozsądne rozwiązanie przybliżone można znaleźć jeżeli do obliczenia dolnej granicy przedziału wartości parametru σ przyjmiemy dolne granice przedziałów wartości tych wielkości x_i które odpowiadają małym wartościom czasów t_i, zaś górne granice przedziałów wartości tych wielkości x_i, które odpowiadają dużym wartościom czasów t_i. W przypadku wyznaczanie górnej granicy przedziału wartości parametru σ powinniśmy postępować w sposób przeciwny.

Prof. dr hab. Olgierd HRYNIEWICZ. Systems Research Institute of the Polish Academy of Sciences, Warsaw, Director. Specializes in reliability, quality control, statistics and fuzzy sets. Author of more than 170 papers from these and related fields.