Fuzzy Modeling of Survival Function from Interval or Censored Observations

Andrzej Szymański and Agnieszka Rossa

Department of Demography and Social Gerontology, University of Lodz, Poland

ABSTRACT

In medicine, biology, or actuarial theory, the so-called survival function is often used, i.e., a probability \( P(X > x) \) that the time \( X \) from the beginning of an initial event to a final one will not exceed \( x \). The observed data are mostly right-censored what causes that the observed variable is \( T = \min(X, Z) \) where \( Z \) is the so-called censoring variable. The celebrated and widely used nonparametric estimator of survival function is the Kaplan-Meier estimator which suffers from a disadvantage that it is stepwise and therefore it may happen that in considerably distinct points the values of the estimator may be equal.

Rossa and Zieliński [11] proposed a local smoothing of the Kaplan-Meier estimator based on an approximation by means of the piecewise Weibull survival function. They have shown that Mean Square Error and Mean Absolute Deviation of the smoothed estimator have been significantly smaller. Moreover, the Weibull approximation method has appeared to be a quite simple algorithm based on logarithmic transformations of the data and by applying the standard estimating procedure of the simple regression model \( y = ax + b \). However, the censoring variables introduce uncertainty, which can be treated as the source of fuzziness. Thus, the estimation problem can be transferred into the fuzzy analysis. Another estimator leading to fuzzy model is the semi-parametric model proposed by Rossa [10]. Parametric part of the estimator contains formulae based on estimates of the Weibull parameters, whereas the censored observations yield uncertainty.

The proposed approaches have a general character. As it has been pointed out in Rossa and Zieliński, the Kaplan-Meier survival function can be approximated with a prescribed level of accuracy by a piecewise Weibull survival function. Using double logarithms of that Weibull pieces we get the piecewise linear intervals. All the intervals can have the same slope parameters or can have changing points separating different slopes. Thus, the basic tool in the approach is the fuzzy linear regression. For simplicity, it will be assumed the fuzzy regression coefficients have symmetric triangular membership functions.

Keywords: Survival function estimation; Censored data; Kaplan-Meier estimator; Weibull distribution; Fuzzy linear regression; Diamond distance; Fuzzy model of survival function

*Corresponding author: anszyman@math.uni.lodz.pl
1. Introduction

Let $F(x), x \geq 0$ be the cumulative distribution function (CDF) of time to failure $X$ of an item and $G(z), z \geq 0$ be the CDF of random time to censoring $Z$ of that item.

Let $T = \min(X, Z)$ and $\delta = I(X \leq Z)$. Variables $X, Z$ are unobservable, independent random variables.

Let us consider a sample of the form:

$$(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n), \quad T_1 \leq T_2 \leq \ldots \leq T_n.$$ (1)

The aim is to estimate survival distribution function (SDF) $S_X(x) = P(X > x) = 1 - F(x)$ from the sample (1).

Let $t_1, t_2, \ldots, t_r$ denote the ordered sequence of $r$ distinct failure times observed in the sample such that $0 = t_0 < t_1, t_2, \ldots, t_r < \infty$ constitute a random partition of the half-line $(0, \infty)$. It is assumed that no failure occurs at time zero. Denote by $n_k$ the number of individuals still alive and under observation just after $t_k$ and by $d_k$ the number of failures occurred at a time point $t_k$, where $n_0 = n$. Let $l_k$ be the number of censoring times in the subinterval $(x_k, x_{k+1})$, $k = 0, 1, \ldots, r - 1$. Thus, we have:

- $n_k$ – Number of observations just after a time point $t_k$;
- $l_k$ – Number of censored observations in the subinterval $(t_k, t_{k+1})$;
- $d_k$ – Number of uncensored observations (failures) occurred at the time point $x_k$;

The Kaplan-Meier [3] estimator ($S_{KM}$), also called the product limit estimator, is defined as:

$$S_{KM}(t_k) = \prod_{j=1}^{k} \frac{n_{j-1} - l_j - d_j}{n_{j-1} - l_j}, \quad \text{for} \quad k = 1, 2, \ldots, r,$$

$$1, \quad \text{for} \quad k = 0.$$ (2)

In the case of ties among the $t_k$ the usual convention is that failure times proceed censoring times. By the definition KM estimator is right-continuous.

Example 1

To illustrate the estimator we shall refer to the well-known example from Freireich et al. [1] – See also Marubini and Valsecchi [8]. The survival times for 21 clinical patients were as follows:

$$6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*$$

where * denotes a censoring observation.
Fuzzy Modeling of Survival Function from Interval or Censored Observations

The Kaplan-Meier estimator is presented in Figure 1 and its values are given in Table 1.

The paper is focused on the improvement of the original Kaplan-Meier estimator by using a fuzzy regression approach. First, we present the Weibull piecewise smoothing of the Kaplan-Maier estimator [10], next the semi-parametric approach proposed by Rossa [15], and the fuzzy model giving the new frames of the former approaches. We have pointed out two uncertainty sources in the survival analysis which lead to such fuzzy models.

2. From local Weibull smoothing to Fuzzy Linear Regression

2.1 Weibull smoothing of the standard Kaplan-Meier estimator

Rossa and Zieliński [10] have proposed local Weibull smoothing of the Kaplan Meier estimator. They have chosen the Weibull survival function mainly because
it gives a simple algorithm of calculating the estimator based on logarithmic transformations of the data and leads to the LS estimating procedure of the parameters in a simple regression model $y = ax + b$.

As it has pointed out by Rossa and Zieliński, each survival probability function may be locally approximated with a prescribed level of accuracy by means of the Weibull survival function of the form:

$$S_{\text{Weil}}(x; \beta, \gamma) = \exp(-\beta x^\gamma), \quad \beta, \gamma > 0, \quad x > 0. \quad (3)$$

Let us denote by $x_1, x_2, \ldots, x_N$ the jump points of KM, and by $P_1, P_2, \ldots, P_N$ the values of KM at these points, i.e:

$$P_i = S_{\text{KM}}(x_i), \quad i = 1, 2, \ldots, N.$$

Let us also define:

$$\bar{P}_i = \frac{P_{i-1} + P_i}{2}, \quad i = 2, 3, \ldots, N.$$

For $i = N$ we put $\bar{P}_N = P_N / 2$ if the last observation is censored and $\bar{P}_N = P_N$ otherwise. For $i = 0$ we have $\bar{P}_0 = 1$.

Now, we use the local smoothing rule which may be expressed as follows. We fit a Weibull survival function which passes through the points $(x_i, \bar{P}_i)$ and $(x_{i+1}, \bar{P}_{i+1})$.

As the smoothed estimator $\hat{S}_X(x)$ of $S_X(x)$ at any $x \in (x_i, x_{i+1})$ we take the value of the fitted Weibull survival probability function. Thus, the values of $\beta_i$ and $\gamma_i$ have to be found by solving the following equations:

$$S_{\text{Weil}}(x_i; \beta_i, \gamma_i) = \bar{P}_i \quad \text{and} \quad S_{\text{Weil}}(x_{i+1}; \beta_i, \gamma_i) = \bar{P}_{i+1}, \quad \text{for} \quad i = 1, 2, \ldots, N.$$

Then we have:

$$\hat{S}_X(x) = S_{\text{Weil}}(x; \beta_i, \gamma_i) \quad \text{for} \quad x \in (x_i, x_{i+1}).$$

Thus, for a fixed $i = 1, 2, \ldots, N$ we have the equations:
Fuzzy Modeling of Survival Function from Interval or Censored Observations

\[
\begin{align*}
\exp(-\hat{\beta}_i \cdot x_i^\gamma) &= \bar{P}_i, \\
\exp(-\hat{\beta}_{i+1} \cdot x_{i+1}^\gamma) &= \bar{P}_{i+1}.
\end{align*}
\] (4)

After taking logarithms of both sides in the equations we receive:

\[
\begin{align*}
\hat{\beta}_i x_i^\gamma &= -\ln \bar{P}_i, \\
\hat{\beta}_{i+1} x_{i+1}^\gamma &= -\ln \bar{P}_{i+1}.
\end{align*}
\] (5)

Taking again logarithms of both sides we obtain:

\[
\begin{align*}
\ln \hat{\beta}_i + \gamma_i \ln x_i &= \ln(-\ln \bar{P}_i), \\
\ln \hat{\beta}_{i+1} + \gamma_{i+1} \ln x_{i+1} &= \ln(-\ln \bar{P}_{i+1}).
\end{align*}
\] (6)

The solution of (6) is:

\[
\gamma_i = \frac{\ln(-\ln \bar{P}_i) - \ln(-\ln \bar{P}_{i+1})}{\ln x_{i+1} - \ln x_i},
\] (7)

and

\[
\ln \hat{\beta}_i = \ln(-\ln \bar{P}_i) - \gamma_i \ln x_i.
\] (8)

The last expression can be rewritten as:

\[
\ln(-\ln \bar{P}_i) = \gamma_i \ln x_i + \ln \hat{\beta}_i.
\] (9)

2.2 The Weibull Plot

Let us assume the coordinate system \((u, v)\) such that:

\[
u_i = \ln x_i \quad \text{and} \quad v_i = \ln(-\ln \bar{P}_i).
\] (10)
Now we can write (9) in the form:

\[ v_i = \tilde{\gamma}_i u_i + \tilde{b}_i \]

Where \( \tilde{b}_i = \ln \tilde{\beta}_i \). Thus, the estimates \( \tilde{\gamma}_i \) and \( \tilde{b}_i \) can be expressed also as follows:

\[
\tilde{\gamma}_i = \frac{v_{i+1} - v_i}{u_{i+1} - u_i} \quad \text{and} \quad \tilde{b}_i = v_i - \tilde{\gamma}_i u_i.
\]  

(11)

Now, we can use the Weibull plot concept introduced by Nelson [9]. The plot has the coordinates \( u \) and \( v \), as given in (10). If the data follow Weibull distribution then points on a plot will be organized in a linear or nearly linear form. The Weibull locally smoothed Kaplan-Meier estimator, given by Rossa and Zieliński, constitutes a piecewise linear plot. We can see it on the Figure 2 using the data given in the Table 2.

There is remarkable influence of the considered data on the values of Kaplan-Meier estimator and therefore one can treat \( \bar{P}_i, i = 1, 2, ..., N \) as fuzzy numbers. The censoring times can change coordinates on the Weibull Plot, as it is shown in the Table 3 (see Example 2).

![Figure 2. The Weibull Plot for the smoothed estimator.](image-url)
Example 2

Let us consider the survival times for 17 clinical patients from the Example 1. However, in this example we change the censoring times to be shorter than the original ones. Then we have:

6, 6, 6, 6*, 7, 9*, 10, 10*, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*.

The locally-smoothed Kaplan-Meier estimates are presented in the Table 3. The problem of estimation of survival functions under right-censoring model can be considered from the fuzzy linear regression perspective, where independent variables $u_k$ are crisp and dependent variables $v_k$ are fuzzy. Hence, the survival function estimation can be solved by means of fuzzy linear regression.

Table 2. Coordinates and $\bar{P}_k$ for the Weibull locally smoothed estimator (calculations based on the censored sample given in Example 1).

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_k$</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>$P_k$</td>
<td>0.857</td>
<td>0.807</td>
<td>0.753</td>
<td>0.690</td>
<td>0.627</td>
<td>0.538</td>
<td>0.448</td>
</tr>
<tr>
<td>$\bar{P}_k$</td>
<td>0.929</td>
<td>0.832</td>
<td>0.780</td>
<td>0.722</td>
<td>0.659</td>
<td>0.583</td>
<td>0.493</td>
</tr>
<tr>
<td>$u_k$</td>
<td>1.792</td>
<td>1.946</td>
<td>2.303</td>
<td>2.565</td>
<td>2.773</td>
<td>3.091</td>
<td>3.135</td>
</tr>
<tr>
<td>$v_k$</td>
<td>-2.601</td>
<td>1.693</td>
<td>-1.383</td>
<td>-1.120</td>
<td>-0.873</td>
<td>-0.615</td>
<td>-0.346</td>
</tr>
</tbody>
</table>

Table 3. Coordinates and $\bar{P}_k$ for the Weibull locally smoothed estimator (calculations based on the censored sample given in Example 1, after removing 4 last censoring times).

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_k$</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>$n_k$</td>
<td>18</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>$d_k$</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$l_k$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>$P_k$</td>
<td>0.824</td>
<td>0.760</td>
<td>0.691</td>
<td>0.605</td>
<td>0.518</td>
<td>0.346</td>
<td>0.173</td>
</tr>
<tr>
<td>$\bar{P}_k$</td>
<td>0.912</td>
<td>0.880</td>
<td>0.846</td>
<td>0.802</td>
<td>0.759</td>
<td>0.673</td>
<td>0.586</td>
</tr>
<tr>
<td>$u_k$</td>
<td>1.792</td>
<td>1.946</td>
<td>2.303</td>
<td>2.565</td>
<td>2.773</td>
<td>3.091</td>
<td>3.135</td>
</tr>
<tr>
<td>$v_k$</td>
<td>-2.382</td>
<td>-2.058</td>
<td>-1.785</td>
<td>-1.513</td>
<td>-1.289</td>
<td>-0.925</td>
<td>-0.628</td>
</tr>
</tbody>
</table>
3. The semi-parametric model

3.1 Censoring in the semi-parametric model

Rossa proposed in 2002 a new semi-parametric estimator of survival function under right-censorship (see [11]). The model assumes independent random censoring, as it was described in the Introduction. In the model an essential assumption concerns the distribution of the random variable \( X \). She assumed that \( X \) has the Weibull distribution with unknown parameters \( \beta \) and \( \gamma \), i.e.:

\[
X \sim \text{Wei}(\beta, \gamma).
\]

3.2 Estimation of survival function in the semiparametric model

Let us define the theoretical distribution functions \( G_0(x) \), \( H(x) \) and their empirical estimates \( \hat{G}_0(x) \), \( \hat{H}(x) \) as:

\[
G_0(x) \equiv P(Z \leq x, \delta = 0), \quad \hat{G}_0(x) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i \leq x, \delta = 0),
\]

\[
H(x) \equiv P(T > x), \quad \hat{H}(x) = \frac{1}{n} \sum_{i=1}^{n} I(T_i > x).
\]

Then the semi-parametric estimator \( \text{SPE}(x) \) of \( S_X(x) \) can be written in the form:

\[
\text{SPE}(x) = \prod_{i=1}^{k} \left[ \frac{\hat{H}(x_i)}{\hat{H}(x_{i-1})} + \frac{1}{\hat{H}(x_{i-1})} \int_{x_{i-1}}^{x_i} \exp\left(\beta (z^\gamma - x_i^\gamma)\right) d\hat{G}_0(z) \right]. \tag{12}
\]

It can be seen that the estimator (12) is right-continuous, non-increasing and defined on the positive half-line. It has a smaller magnitude of jumps than the original KM estimator. The last property is important especially for small or moderate samples under heavy censoring.

3.3 Parameters estimation in the semi-parametric model

Let be given a sample (1) of size \( n \). The sequence \( T_1, T_2, \ldots, T_n \) consists of both true failure and censoring times according to the values of \( \delta \). We rearrange this sequence into two subsequences \( T_1, T_2, \ldots, T_r \) and \( T_{r+1}, \ldots, T_n \), corresponding to the values \( \delta = 1 \) and \( \delta = 0 \), being the values of \( X \) and \( Z \), respectively.
We write the likelihood functions $L_X$ and $L_Z$ for both sequences. The likelihood function for the sample is thus the product of the two likelihood functions, i.e.:

$$L(t_1, t_2, \ldots, t_n; \beta, \gamma) = L_X(t_1, t_2, \ldots, t_r; \beta, \gamma) \times L(t_{r+1}, t_{r+2}, \ldots, t_n; \beta, \gamma).$$

The likelihood function $L_X$ for failures times has the form:

$$L_X(t_1, t_2, \ldots, t_r; \beta, \gamma) = \beta^r \gamma^{t_r-1} \exp\left(-\beta t_r^\gamma\right).$$

The likelihood function $L_Z$ for censoring times is the following:

$$L_Z(t_{r+1}, t_{r+2}, \ldots, t_n; \beta, \gamma) = \prod_{i=r+1}^n S_X(t_i) = \prod_{i=r+1}^n \exp\left(-\beta t_i^\gamma\right). \quad (13)$$

From (13) the maximum likelihood equations for determining the estimators of $\beta$ and $\gamma$ are as follows:

$$\begin{align*}
\beta \left( \frac{r}{\beta} - \frac{\sum t_i^\gamma}{\sum t_i^\gamma} \right) &= 0, \\
\gamma \left( \frac{r}{\gamma} + \frac{\sum \ln t_i - \beta \sum t_i^\gamma \ln t_i}{\sum t_i^\gamma \ln t_i} \right) &= 0,
\end{align*} \quad (14)$$

or equivalently:

$$\begin{align*}
\hat{\beta} &= \frac{r}{\sum t_i^\gamma}, \\
\hat{\gamma} &= \left( \frac{\sum \ln t_i - \frac{r}{\beta} \frac{\sum t_i^\gamma \ln t_i}{\sum t_i^\gamma}}{\sum t_i^\gamma \ln t_i} \right) = 0.
\end{align*} \quad (15)$$
Example 3 (Application of the semi-parametric model)

Let us consider data of Klein et al. [4]. The dataset represents times to death (in days), post transplant, of 42 patients in the Ohio State University Bone Marrow Transplant Program (* denotes a censored observation).

Dataset (in days):

\[
\begin{array}{cccccccccccccccccccccccccccccc}
& 2 & 27 & 32^* & 43^* & 50 & 55^* & 62 & 82^* & 102^* & 103^* & 122 \\
145^* & 148 & 158 & 162 & 194^* & 250^* & 251 & 267^* & 271 & 276 & 284^* & 292^* \\
319^* & 326^* & 346^* & 365^* & 404^* & 417 & 418 & 423^* & 438^* & 491 & 584^* \\
595^* & 613^* & 642^* & 649^* & 693^* & 707^* & 746^* & 755^* & 826^* \\
\end{array}
\]

Figure 3. Kaplan-Meier estimator for data from 42 patients from Ohio State University [11].

Estimates $\hat{\beta}$ and $\hat{\gamma}$ of the parameters are the following:

$\hat{\beta} = 0.00333728$, $\hat{\gamma} = 0.789556483$.

It is worth noting that the estimates depend on the length $T_n$ of the observation period (see Table 4). Thus, it can be treated as a source of uncertainty, what is illustrated in the Table 4.

The data, transformed into the Weibull Plot, show that the distribution of the random variable is nearly Weibull (see Figure 4). However, one can see a single outlier. After removing it we get a well-adjusted linear regression line and the problem of estimation reduces to the linear fuzzy regression analysis (see Figure 5).
3.4 Theoretical backgrounds of the fuzzy numbers

Definition 1
We define a fuzzy number \( A \) as a set \( A = \{ (t, \mu_A(t)) : t \in \mathbb{R} \} \), where \( \mu_A : \mathbb{R} \rightarrow [0, 1] \) is a membership function of \( A \).

Definition 2
A triangular fuzzy number \( A = (a, l_A, r_A) \) with a center \( a \in \mathbb{R} \), and left and right spreads \( (l_A, r_A) \) is defined by its triangular membership function (see Figure 6).

Table 4. Parameter estimates and \( \gamma \) for different observation length.

<table>
<thead>
<tr>
<th>( T_n )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>850</td>
<td>0.003337</td>
<td>0.789556483</td>
</tr>
<tr>
<td>800</td>
<td>0.003128</td>
<td>0.809314792</td>
</tr>
<tr>
<td>750</td>
<td>0.002409</td>
<td>0.828171624</td>
</tr>
<tr>
<td>700</td>
<td>0.002409</td>
<td>0.871872897</td>
</tr>
<tr>
<td>650</td>
<td>0.002202</td>
<td>0.897099204</td>
</tr>
<tr>
<td>600</td>
<td>0.001388</td>
<td>0.986406727</td>
</tr>
<tr>
<td>550</td>
<td>0.000899</td>
<td>1.072002064</td>
</tr>
<tr>
<td>500</td>
<td>0.000899</td>
<td>1.072002064</td>
</tr>
</tbody>
</table>

Figure 4. The Weibull Plot for the Kaplan-Meier estimator based on the data from Ohio State University [10].
Definition 3

The $\lambda$-cuts of the fuzzy number $A$ is a set $A_\lambda = \{ t \in R : \mu_A(t) \geq \lambda ; 0 \leq \lambda \leq 1 \}$.

Definition 4

A symmetric triangular fuzzy number (STFN) $A = (a, s_a)$ with a center $a \in R$ and the spread $s_a$ is defined by its triangular membership function with $l_a = r_a = s_a$.
Definition 5

For two fuzzy numbers \( A \) and \( B \) with \( \lambda \)-cuts \( A_\lambda = [A_\lambda^L, A_\lambda^U] \) and \( B_\lambda = [B_\lambda^L, B_\lambda^U] \) the Diamond metric is expressed as:

\[
d^2(A, B) = \int_0^1 (A_\lambda^L - B_\lambda^L)^2 d\lambda + \int_0^1 (A_\lambda^U - B_\lambda^U)^2 d\lambda
\]

Property 1. (Diamond metric for triangular numbers)

Let \( A = (a, l_A, r_A) \) and \( B = (b, l_B, r_B) \) be two triangular fuzzy numbers. Then the Diamond metric is equal:

\[
d^2(A, B) = (a - b)^2 + ((a - l_A) - (b - l_B))^2 + ((a + r_A) - (b + r_B))^2
\]

Property 2

Let \( A = (a, s_A) \) and \( B = (b, s_B) \) be two triangular symmetric fuzzy numbers. Then the Diamond metric can be written in the form:

\[
d^2(A, B) = 2(a - b)^2 + \frac{2}{3}(s_A - s_B)^2
\]

4. Fuzzy linear regression

4.1 Parameters estimation

Fuzzy linear regression problem may be expressed as a problem of identification of parameters \( a, b \) of a fuzzy linear model \( V = aU + b \).

Let \( 7(\mathbb{R}) \) be the set of all symmetric triangular numbers and let \( \mathbb{R} \) be the real space. There are four simple regression models described in the literature:

(F1) \( V = aU + b, \ a, b \in \mathbb{R}, U, V \in 7(\mathbb{R}) \);

(F2) \( V = aU + E, \ a \in \mathbb{R}, E, U, V \in 7(\mathbb{R}) \);

(F3) \( V = Au + B, \ u \in \mathbb{R}, A, B, V \in 7(\mathbb{R}) \);
our problem belongs to (F3), i.e.:

\[ V = Au + B, \quad u \in \mathbb{R}, \quad A, B, V \in \mathbb{R}. \]

Suppose that our data set consists of pairs \((u_i, V_i)\), \(i = 1, 2, \ldots, n\), where:

\begin{align*}
V_i &= (y_i, e_i), \quad (16) \\
A &= (a, s_A), \quad (17) \\
B &= (b, s_B). \quad (18)
\end{align*}

We will consider the best fit of the model with respect to the Diamond metric. The parameters \(a, b, s_A, s_B\) will be evaluated by minimizing the error -- Measured by means of the Diamond metric -- Between the actual observations and the estimates obtained from the model. The corresponding least square optimization problem is as follows:

\[
\text{minimize } r(A, B) = \sum_{i=1}^{n} d^2(V_i, Au_i + B). \quad (19)
\]

First, crisp data (i.e., non-fuzzy data) \(v_i, i = 1, 2, \ldots, n\) resulting from the Kaplan-Meier procedure should be fuzzified before calculations. One of the fuzzification methods is shown in [5]. According to this method, fuzzy numbers \(V_i\) can be obtained from fuzzy regression. The task is to find symmetric fuzzy numbers \(A, B\) and \(V_i\) with centers \(a, b, v_i\) and spreads \(s_A, s_B, e_i\) such that:

\[
\begin{align*}
(v_i, e_i) &= (a, s_A) \times u_i + (b, s_B) \\
&= (au_i, s_A u_i) + (b, s_B) \\
&= (au_i + b, s_A u_i + s_B) \quad (20)
\end{align*}
\]

From the Property 2 we have:

\[
d^2(V_i, Au_i + B) = 2(v_i - au_i - b)^2 + \frac{2}{3}(e_i - s_A u_i - s_B)^2
\]
Thus, the functional (19) takes the form:

$$r(A, B) = 2 \sum_{i=1}^{n} (u_i - au_i - b)^2 + \frac{2}{3} \sum_{i=1}^{n} (e_i - s_A u_i - s_B)^2$$

(21)

Hence, denoting:

$$f(a, b) = 2 (u_i - au_i - b)^2 \quad \text{and} \quad g(s_A, s_B) = \frac{2}{3} (e_i - s_A u_i - s_B)^2$$

We can write (21) as:

$$r(A, B) = f(a, b) + g(s_A, s_B)$$

(22)

After differentiating and solving the least-square equations we get:

$$\hat{a} = \frac{\overline{uv} - \overline{u} \cdot \overline{v}}{\overline{u}^2 - \overline{u}^2} \quad \text{and} \quad \hat{b} = \overline{v} - \hat{a} \cdot \overline{u}$$

(23)

Where $\overline{z}$ denotes an arithmetic average.

4.2 Defuzzification of the resulting regression

Very often in practice we are not satisfied with fuzzy results and need a precisely described crisp counterparts. In such cases an adequate defuzzification method should be used. However, the problem arises, how to choose the optimal defuzzification operator. Many defuzzification methods are described in the literature, especially in the series of papers of Kosiński and co-workers (see: [6, 7]). In Grzegorzewski [2] two fuzzification operators are described: the maximum value operator and the randomized operator. They are simple and natural. By the definition [8], the maximum value operator selects the crisp decision which maximizes a membership function of the considered fuzzy decision. In the case of symmetric triangular fuzzy observations with crisp input, the maximum value of defuzzification operator gives the central value.

Example 4

Let us call the survival times from the Example 1. The fuzzy regression method gives the following results:
\[ \hat{a} = 1.229 \text{ and } \hat{b} = -4.3718. \]

The fitted and defuzzified survival function is shown on the Figure 7.

**Figure 7.** Defuzzified estimator derived from the data [8].

**Example 5**

Let us use the survival times from the **Example 2** concerning survival times after heart transplantation. In this example we get:

\[ \hat{a} = 0.8807 \text{ and } \hat{b} = -6.1845. \]

The fitted and defuzzified survival function has been shown on the Figure 8.

**Acknowledgement**

This research was supported for both authors by a grant from the National Science Center under contract DEC-2011/01/B/HS4/02882, for which the authors are indebted.
References


