A SIMULATION STUDY TO COMPARE IMPUTATION METHODS TO HANDLE GROUPED SURVIVAL DATA

Elizabeth STRAPASSON[1]
Enrico Antonio COLOSIMO[2]
Suely Ruiz GIOLO[3]
Clarice Garcia Borges DEMÉTRIO[4]

▪ *ABSTRACT: Grouped or discrete survival data usually occur when all the experimental units are visited at the same times which can be equidistant or not. To analyze such type of data simple imputation methods are often used by analysts. In this paper, Monte Carlo simulations are performed in order to compare the midpoint, lower and upper limits imputation methods. The Weibull regression model for discrete survival data is also considered. In the simulations, three equidistant visit times, three proportions of censored data, three sample sizes, and the Weibull distribution are used. Also, the situations without and with covariates are considered in the study. The simulation results show that the midpoint is the best imputation method among those considered giving very similar results to those of the Weibull regression model for discrete survival data. A flax data illustrates the comparison of the methods considered in the paper for the analysis of grouped survival data.*

▪ *KEYWORDS: Discrete time; Monte Carlo simulation; Weibull model.*

1 Introduction

Survival time or lifetime is defined as the time to the occurrence of a given event of interest. However, in many studies the exact time of a failure is unknown.

[1]Departamento de Estatística, Universidade Estadual de Londrina – UEL, Caixa Postal 6001, CEP: 86051-990 , Londrina, PR, Brazil. E-mail: *estrapas@uel.br*
[2]Departamento de Estatística, Universidade Federal de Minas Gerais – UFMG, CEP: 31270-901, Belo Horizonte, MG, Brazil. E-mail: *enricoc@est.ufmg.br*
[3]Departamento de Estatística, Universidade Federal do Paraná – UFPR, Caixa Postal 19081, CEP: 81531-990, Curitiba, PR, Brazil. E-mail: *giolo@ufpr.br*
[4]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiroz – ESALQ, Universidade de São Paulo – USP, Caixa Postal 9, CEP: 13418-900, Piracicaba, SP, Brazil. E-mail: *clarice@esalq.usp.br*

Instead, the failure time data is recorded in time intervals and this is known as interval censored data. In some situations the visit times are the same for all the experimental units, leading to many ties and then the survival data can be grouped into mutually exclusive intervals. In this particular case the data are called grouped survival data. Grouped data are common in agronomy studies when the field visits are scheduled for specified dates (Colosimo et al., 2000, Giolo et al., 2009), in animal evaluations when the observations are recorded at relatively long intervals as for instance the time to gain a desirable weight and the length of the productive life of dairy cattle, in enthomology when the death of insects are observed daily (Petkau et al., 1989), and in clinical trials and longitudinal studies when event occurrence is monitored by routine medical visits (Sun, 1996; Kim et al., 1993).

According to Lindsey and Ryan (1998) a common approach to analyze interval-censored data, termed simple imputation, is to assume that the event occurred at the end (or beginning or midpoint) of each interval and then apply common methods for standard time-to-event data. Although no simulation study to compare simple imputation methods were found in the literature, the authors conclude that imputation, when used to analyze interval-censored data (with grouped data being a special case), can lead to biased and misleading results. Law and Brookmeyer (1992) also studied some statistical properties of the midpoint imputation and concluded that they are strongly dependent on the widths of the intervals.

Multiple imputation can also be used for analyzing interval-censored data. It differs from simple imputation in the sense that estimates and measures of uncertainty can also be obtained. A nice overview and discussion of multiple imputation methods is provided by Hsu et al. (2007). In order to handle the data as being grouped, avoiding the use of any imputation, some methods have been proposed in the literature as, for instance, the discrete models (Prentice and Gloeckler, 1978; Collett, 2003; Lawless, 2003). For these models, however, no diagnostics techniques are available for evaluating their goodness-of-fit.

Although multiple imputation seems to be statistically more attractive than simple imputation for analyzing interval-censored data, simplest procedures such as the midpoint imputation are usually prefered by analysts. The objective of this paper is therefore to perform a simulation study to evaluate and compare three methods of simple imputation: midpoint, lower and upper limits of intervals. The remainder of this paper is organized as follows. Section 2 describes four methods that can be used for handling grouped data. Section 3 describes details of the simulation study performed. Section 4 gives results from the simulation study and Section 5 applies the methods in an example that includes covariates.

2 The underlying methods

Suppose that the event times are grouped into k intervals, $I_i = [a_{i-1}, a_i), i = 1, 2, ..., k$, where $0 = a_0 < a_1 < ... < a_k = \infty$ and assume that all censoring takes place at the end of the intervals. Next, three imputation methods and the Weibull discrete model are presented in order to handle this situation.

2.1 Imputation methods

Under the imputation approach the data from the ith interval $I_i = [a_{i-1}, a_i), i = 1, 2, ..., k$, is replaced by a value that is assumed as the exactly observed failure time. The first imputation method ($M1$) considered in this paper assumes that this value is the lower limit of the interval I_i. The second imputation method ($M2$) assumes the midpoint while the last method ($M3$) assumes the upper limit of I_i. Table 1 summarizes these imputation methods.

Table 1 - Summary of the three imputation methods

Interval	Frequency	Imputation methods			Imputed times
		M1	M2	M3	
$[a_0; a_1)$	f_1	a_0	$\frac{a_0+a_1}{2}$	a_1	$t_1, \ldots, t_{f_1}$
$[a_1; a_2)$	f_2	a_1	$\frac{a_1+a_2}{2}$	a_2	$t_{f_1+1}, \ldots, t_{f_1+f_2}$
...	...	...	...	...	...
$[a_{k-1}; a_k)$	f_k	a_{k-1}	$\frac{a_{k-1}+a_k}{2}$	a_k	$t_{\sum_{i=1}^{k-1} f(i)+1}, \ldots, t_{\sum_{i=1}^{k} f(i)}$

The imputed times obtained by using one of the imputation methods ($M1, M2$ or $M3$), are then treated as values of a continuous response, T, and the well known models and diagnostics techniques for survival continuous data (Collett, 2003; Lawless, 2003) are used. Suppose that the continuous random variable lifetime T has distribution with density $f(t)$, survivor function $S(t)$ and hazard function $h(t)$ and consider a vector $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$ of p covariates associated with each individual and a $p+1$ vector of parameters $\boldsymbol{\beta}$. The likelihood function is given by

$$\prod_{i=1}^{k} \prod_{j=1}^{n_i} [f(t_{ij} \mid \mathbf{x}_{ij})]^{\delta_{ij}} \, [S(t_{ij} \mid \mathbf{x}_{ij})]^{1-\delta_{ij}} = \prod_{i=1}^{k} \prod_{j=1}^{n_i} [h(t_{ij} \mid \mathbf{x}_{ij})]^{\delta_{ij}} \, [S(t_{ij} \mid \mathbf{x}_{ij})], \qquad (1)$$

where t_{ij} is defined according to the imputation methods $M1$, $M2$ or $M3$, n_i is the number of events (failures or censorings) at I_i and δ_{ij} is a failure indicator variable for the jth subject's lifetime in the interval I_i ($\delta_{ij} = 0$ if it is censored and $\delta_{ij} = 1$, otherwise). The expression (1) is the standard likelihood function for continuous

right censored data (Lawless, 2003). Supposing that the random variable lifetime T follows a Weibull distribution with

$$f(t) = \frac{\gamma}{\left[\exp(\boldsymbol{\beta}'\mathbf{x})\right]^{\gamma}}\ t^{\gamma-1} \exp\left\{-\left[\frac{t}{\left[\exp(\boldsymbol{\beta}'\mathbf{x})\right]}\right]^{\gamma}\right\},\ t \geq 0,$$

$$S(t) = \exp\left\{-\left[\frac{t}{\left[\exp(\boldsymbol{\beta}'\mathbf{x})\right]}\right]^{\gamma}\right\},\ t \geq 0,$$

and

$$h(t) = \frac{\gamma}{\left[\exp(\boldsymbol{\beta}'\mathbf{x})\right]^{\gamma}} t^{\gamma-1},\ t \geq 0,$$

the likelihood function for the Weibul model is given by

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left\{\frac{\gamma}{\left[\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})\right]^{\gamma}}\ {t_{ij}}^{\gamma-1}\right\}^{\delta_{ij}} \exp\left\{-\left[\frac{t_{ij}}{\left[\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})\right]}\right]^{\gamma}\right\}.$$

Parameter estimation can be obtained by using Newton-Raphson numerical method in a standard statistical package.

2.2 Weibull discrete model

The discrete model does not consider imputed times as the methods described in Section 2.1. Instead, the grouped data information is taken into account in the likelihood function which is given by

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}[S(a_{i-1} \mid \mathbf{x}_{ij}) - S(a_i \mid \mathbf{x}_{ij})]^{\delta_{ij}}\ [S(a_{i-1} \mid \mathbf{x}_{ij})]^{1-\delta_{ij}}. \tag{2}$$

The expression (2) yields $[S(a_{i-1} \mid \mathbf{x}_{ij}) - S(a_i \mid \mathbf{x}_{ij})]$ for observations grouped into the k intervals $I_i, i = 1, 2, ..., k$, and $[S(a_{i-1} \mid \mathbf{x}_{ij})]$ for right-censored observations. Considering the Weibull regression model $(M4)$ the likelihood function is given by

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left\{\exp\left[-\left[\frac{a_{i-1}}{(\exp(\boldsymbol{\beta}'\mathbf{x}_{ij}))}\right]^{\gamma}\right] - \exp\left[-\left[\frac{a_i}{(\exp(\boldsymbol{\beta}'\mathbf{x}_{ij}))}\right]^{\gamma}\right]\right\}^{\delta_{ij}}$$
$$\left\{\exp\left[-\left[\frac{a_{i-1}}{(\exp(\boldsymbol{\beta}'\mathbf{x}_{ij}))}\right]^{\gamma}\right]\right\}^{1-\delta_{ij}} =$$
$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}\left\{1 - \exp\left[-\exp(\gamma\boldsymbol{\beta}'\boldsymbol{x_{ij}} + \gamma\beta_0 + \log({a_i}^{\gamma} - a_{i-1}^{\gamma}))\right]\right\}^{\delta_{ij}}$$
$$\left\{\exp\left[-\exp(\gamma\boldsymbol{\beta}'\boldsymbol{x_{ij}} + \gamma\beta_0 + \log({a_i}^{\gamma} - a_{i-1}^{\gamma}))\right]\right\}^{1-\delta_{ij}},$$

which can be seen as the likelihood function for n observations δ_{ij} from a Bernoulli distribution (Whitehead, 1989) with probability $p_{ij} = P(a_{j-1} < T_{ij} < a_j \mid T_{ij} \geq a_{j-1})$ with a complementary log-log link function and linear predictor $\eta_{ij} = \log(a_j^{\gamma} - a_{j-1}^{\gamma}) + \gamma\beta_0 + \gamma\boldsymbol{\beta}'\boldsymbol{x}_{ij}$. This fact makes easier the numerical solution of the maximum likelihood estimation.

3 Simulation study

Monte Carlo simulations were performed in order to compare the three imputation methods and the Weibull regression model considered to analyse grouped data. For the simulation study the survival times T_i, $i = 1, 2, \ldots, n$, were assumed to follow a Weibull distribution with shape parameter γ and scale parameter λ. The value assumed for the parameter γ was 2 while the value for the parameter λ was 1 for the no covariate case and $\lambda = \exp(\beta_0 + \beta_1 x)$ with $\beta_0 = 0$ and $\beta_1 = 1$ for the covariate case. The values for the covariate x were generated from a Bernoulli ($p = 0.5$) distribution.

Samples of sizes 50, 100 and 200, using 0%, 20% and 40% of right censoring were simulated, giving a total of 27 simulation runs. For the 27 simulations with 1,000 replications each, three situations with 12, 8 and 4 intervals were used to produce the grouped data and the interval frequencies. The initial time was zero for all the situations.

For each method, number of intervals and combination of simulation scenarios, a Weibull regression model was fitted. The parameter estimates $(\hat{\gamma}, \hat{\beta}_0, \hat{\beta}_1)$, the mean bias $(\hat{\theta} - \theta)$ and the mean square error (MSE) $((\hat{\theta} - \theta)^2 + \hat{Var}(\hat{\theta}))$ over the 1,000 simulation runs, where θ represent any of the parameters, were obtained. The MSE values were used to calculate the efficiency of each estimator related to the true value. All the simulations were carried out using the R environment (R Development Core Team, 2008).

4 Simulation study results

Results obtained from the simulation study for both no covariate and covariate cases are presented in the two next sections.

4.1 No covariate case

Tables 2 to 4 show the parameter estimates of λ and γ and their corresponding mean square error (MSE) across all the 1,000 simulation runs considering the 27 situations described in Section 3. The simulated data was classified into 12, 8 and 4 intervals and analyzed by the three imputation methods and the Weibull discrete regression model. From the tables mentioned, we can see that the estimates obtained from method $M1$ are smaller than those obtained from all other methods for either all proportion of censoring and sample sizes considered. The estimates can also be observed to decrease as the number of intervals decreases.

Table 2 - Estimates (E) of the parameters (P) and mean square error (MSE) considering samples simulated of size (n) 50, 100 and 200 of Weibull distribution with $\lambda = 1$ and $\gamma = 2$, three intervals (G = 12, 8 and 4) and the methods of estimation M1, M2, M3 and M4 for 0% of censoring and no covariate case

n	G	P	M1		M2		M3		M4	
			E	MSE	E	MSE	E	MSE	E	MSE
50	12	λ	0.80	0.045	1.00	0.006	1.16	0.030	1.00	0.006
		γ	1.46	0.355	2.02	0.055	2.41	0.233	2.06	0.067
	8	λ	0.68	0.108	1.00	0.006	1.24	0.062	1.00	0.006
		γ	1.10	0.858	1.99	0.055	2.54	0.368	2.07	0.080
	4	λ	0.29	0.515	1.00	0.006	1.47	0.230	1.00	0.007
		γ	0.53	2.173	1.90	0.040	2.84	0.781	2.24	0.624
100	12	λ	0.80	0.041	1.00	0.003	1.16	0.029	1.00	0.003
		γ	1.43	0.356	2.00	0.025	2.37	0.167	2.03	0.030
	8	λ	0.69	0.103	1.00	0.003	1.24	0.060	1.00	0.003
		γ	1.08	0.871	1.96	0.026	2.50	0.282	2.03	0.035
	4	λ	0.29	0.514	1.00	0.003	1.48	0.230	1.00	0.004
		γ	0.52	2.183	1.87	0.029	2.79	0.662	2.06	0.077
200	12	λ	0.80	0.040	1.00	0.002	1.16	0.027	1.00	0.001
		γ	1.41	0.360	1.98	0.012	2.35	0.135	2.01	0.013
	8	λ	0.69	0.100	1.00	0.002	1.24	0.059	1.00	0.002
		γ	1.07	0.882	1.95	0.014	2.48	0.243	2.01	0.016
	4	λ	0.28	0.516	1.00	0.002	1.48	0.228	1.00	0.002
		γ	0.52	2.190	1.86	0.025	2.77	0.608	2.03	0.029

Table 3 - Estimates (E) of the parameters (P) and mean square error (MSE) considering samples simulated of size (n) 50, 100 and 200 of Weibull distribution with $\lambda = 1$ and $\gamma = 2$, three intervals (G = 12, 8 and 4) and the methods of estimation M1, M2, M3 and M4 for 20% of censoring and no covariate case

n	G	P	M1		M2		M3		M4	
			E	MSE	E	MSE	E	MSE	E	MSE
	12	λ	0.96	0.009	1.13	0.024	1.29	0.090	1.15	0.029
		γ	1.79	0.148	2.35	0.211	2.75	0.672	2.42	0.278
50	8	λ	0.86	0.028	1.13	0.024	1.37	0.142	1.16	0.033
		γ	1.35	0.487	2.28	0.148	2.86	0.849	2.41	0.260
	4	λ	0.46	0.305	1.14	0.029	1.61	0.380	1.22	0.056
		γ	0.60	1.955	2.12	0.064	3.14	1.428	2.47	1.202
	12	λ	0.94	0.007	1.12	0.018	1.28	0.082	1.15	0.025
		γ	1.46	0.318	2.05	0.027	2.43	0.218	2.07	0.030
100	8	λ	0.84	0.028	1.12	0.018	1.36	0.132	1.16	0.030
		γ	1.07	0.888	1.99	0.027	2.54	0.335	2.02	0.031
	4	λ	0.44	0.316	1.14	0.023	1.60	0.370	1.23	0.057
		γ	0.54	2.115	1.95	0.012	2.90	0.840	1.97	0.014
	12	λ	0.95	0.005	1.12	0.017	1.28	0.082	1.15	0.024
		γ	1.47	0.290	2.06	0.015	2.45	0.216	2.09	0.020
200	8	λ	0.85	0.025	1.12	0.017	1.36	0.133	1.16	0.029
		γ	1.10	0.822	2.02	0.014	2.56	0.339	2.06	0.020
	4	λ	0.44	0.317	1.14	0.021	1.60	0.368	1.23	0.055
		γ	0.54	2.128	1.93	0.012	2.87	0.769	1.93	0.018

Table 4 - Estimates (E) of the parameters (P) and mean square error (MSE) considering samples simulated of size (n) 50, 100 and 200 of Weibull distribution with $\lambda = 1$ and $\gamma = 2$, three intervals (G = 12, 8 and 4) and the methods of estimation M1, M2, M3 and M4 for 40% of censoring and no covariate case

n	G	P	M1		M2		M3		M4	
			E	MSE	E	MSE	E	MSE	E	MSE
	12	λ	1.13	0.025	1.29	0.090	1.44	0.200	1.33	0.114
		γ	2.11	0.177	2.65	0.558	3.07	1.303	2.72	0.654
50	8	λ	1.06	0.012	1.29	0.092	1.52	0.277	1.35	0.134
		γ	1.60	0.252	2.58	0.441	3.20	1.599	2.72	0.635
	4	λ	0.74	0.091	1.32	0.110	1.77	0.604	1.48	0.238
		γ	0.70	1.683	2.40	0.239	3.51	2.476	2.61	0.438
	12	λ	1.15	0.026	1.29	0.088	1.44	0.198	1.34	0.122
		γ	1.54	0.234	2.20	0.070	2.62	0.422	2.24	0.085
100	8	λ	1.09	0.012	1.29	0.091	1.52	0.276	1.38	0.149
		γ	1.22	0.628	2.20	0.060	2.78	0.648	2.24	0.081
	4	λ	0.74	0.078	1.32	0.107	1.77	0.604	1.52	0.280
		γ	0.59	1.983	2.10	0.020	3.10	1.246	2.10	0.017
	12	λ	1.15	0.026	1.29	0.087	1.44	0.198	1.35	0.126
		γ	1.47	0.292	2.09	0.019	2.49	0.255	2.10	0.020
200	8	λ	1.10	0.013	1.30	0.091	1.52	0.276	1.39	0.157
		γ	1.10	0.812	2.05	0.016	2.62	0.402	2.05	0.017
	4	λ	0.75	0.070	1.32	0.106	1.78	0.603	1.56	0.315
		γ	0.54	2.141	1.94	0.009	2.88	0.795	1.78	0.052

Regarding methods $M2$ and $M4$, it can be observed, for all sample sizes, the estimates of λ increasing and remaining more distant from the true parameter value as the number of intervals decreases and the proportion of censoring increases to 20% and 40%. For 0% of censoring no remarkable changes can be noticed from the results. For method $M3$, the estimates of λ and γ show higher values than those obtained for any other method for all proportion of censoring and sample sizes. The estimates also increase as the number of intervals decreases.

For all methods, we can also see the estimates of λ and γ to increase as the proportion of censoring increases. With regard the mean square error (MSE), it in general decreases as the sample size increases, as well as it increases as the proportion of censoring increases. Exceptions, however, can be observed for method

$M1$. Overall, methods M_2 and M_4 provide very similar results and also the closest estimates to the true parameter values, particularly as the sample size and the number of intervals increase, and the proportion of censoring decreases. Hence, among the three imputation methods evaluated, the midpoint ($M2$) proven to have the best performance, at least for the scenarios considered.

4.2 Covariate case

Results obtained when a dichotomic covariate is considered in the same simulations scenarios evaluated in Section 4.1 are displayed in Tables 5 to 7. From these tables, the estimates of β_1 for both methods $M2$ and $M4$ can be observed to be similar for all proportion of censoring considered. Also, when compared with $M2$, estimates of γ obtained from $M4$ are in general larger for 0% of censoring, and smaller for 20% and 40% of censoring. These results can be noticed for all sample sizes considered. In addition, estimates of β_1 and γ when obtained from $M2$ are observed to increase as the number of intervals decreases for all proportion of censoring. In opposite, for method $M4$, estimates of these same parameters decrease as the number of intervals decreases for both 20% and 40% of censoring and for all three sample sizes.

With regard to method $M1$, the estimates of β_1 and γ for all sample sizes are smaller than those obtained from all other methods. They also decrease as the number of intervals decreases for all proportion of censoring and sample sizes. Concerning the method $M3$, estimates of β_1 are larger than those obtained from all other methods. They also increase as the number of intervals decreases. In addition, for all proportion of censoring and sample sizes considered, estimates of γ are observed to decrease as the number of intervals decreases.

When looking at the results for all four methods together, we can see that the estimates of β_1 increase with the increasing in the proportion of censoring. In addition, for all sample sizes, estimates of γ for 20% and 40% of censoring decrease when obtained from $M3$ and $M4$. In general, estimates are less closer to the true parameter values as the number of intervals decreases. On the other hand, for all methods in the three proportion of censoring, the MSE usually increases as the number of intervals decreases. MSE also decreases as the sample size increases, particularly for methods $M2$ and $M4$.

Overall, and as in the no covariate case, the midpoint imputation method $M2$ gives similar results as the discrete model $M4$. Also, among the three simple imputation methods evaluated in the simulation study, $M2$ seems to be the one presenting the best performance in the sense that, in general, the estimates obtained from it were closer to the true parameter values, especially as the sample size and the number of intervals increase, and the proportion of censoring decreases.

Table 5 - Estimates (E) of the parameters (P) and mean square error (MSE) considering samples simulated of size (n) 50, 100 and 200 of Weibull distribution with $\gamma = 2$, $\beta_0 = 0$ and $\beta_1 = 1$, three intervals (G = 12, 8 and 4) and the methods of estimation M1, M2, M3 and M4 for 0% of censoring

n	G	P	M1		M2		M3		M4	
			E	MSE	E	MSE	E	MSE	E	MSE
	12	β_1	0.55	0.227	0.91	0.019	1.09	0.013	0.92	0.019
		γ	0.78	1.521	1.96	0.033	2.33	0.145	2.05	0.061
50	8	β_1	0.21	0.680	0.91	0.019	1.12	0.021	0.91	0.021
		γ	0.51	2.218	1.93	0.021	2.12	0.056	2.03	0.089
	4	β_1	-1.16	4.944	0.97	0.015	1.22	0.057	0.90	0.025
		γ	0.35	2.707	2.37	0.207	1.32	0.458	3.20	7.330
	12	β_1	0.72	0.087	1.03	0.006	1.15	0.026	1.04	0.007
		γ	0.79	1.487	1.93	0.017	1.90	0.036	2.06	0.026
100	8	β_1	0.53	0.242	1.04	0.008	1.20	0.046	1.06	0.010
		γ	0.52	2.200	1.97	0.006	1.77	0.064	2.27	0.121
	4	β_1	-0.50	2.408	1.15	0.030	1.32	0.109	1.06	0.013
		γ	0.35	2.724	2.99	1.138	1.18	0.674	2.51	1.411
	12	β_1	0.65	0.126	0.99	0.003	1.13	0.018	1.00	0.003
		γ	0.77	1.524	2.04	0.008	2.18	0.045	2.17	0.041
200	8	β_1	0.43	0.340	1.00	0.003	1.17	0.031	1.00	0.003
		γ	0.51	2.221	2.01	0.004	1.97	0.005	2.21	0.071
	4	β_1	-0.82	3.378	1.08	0.009	1.27	0.076	0.98	0.004
		γ	0.35	2.728	2.64	0.437	1.24	0.584	2.12	0.166

Table 6 - Estimates (E) of the parameters (P) and mean square error (MSE) considering samples simulated of size (n) 50, 100 and 200 of Weibull distribution with $\gamma = 2$, $\beta_0 = 0$ and $\beta_1 = 1$, three intervals (G = 12, 8 and 4) and the methods of estimation M1, M2, M3 and M4 for 20% of censoring

n	G	P	M1		M2		M3		M4	
			E	MSE	E	MSE	E	MSE	E	MSE
	12	β_1	0.70	0.109	0.98	0.012	1.12	0.022	0.98	0.011
		γ	0.98	1.072	2.10	0.046	2.26	0.116	2.06	0.040
50	8	β_1	0.46	0.342	0.97	0.012	1.16	0.033	0.96	0.012
		γ	0.63	1.874	2.04	0.019	1.98	0.045	1.88	0.062
	4	β_1	-0.79	3.497	1.00	0.015	1.30	0.097	0.80	0.076
		γ	0.39	2.584	2.23	0.116	1.15	0.729	0.95	1.396
	12	β_1	0.83	0.037	1.07	0.010	1.18	0.036	1.06	0.009
		γ	0.99	1.036	1.98	0.020	1.77	0.081	1.87	0.038
100	8	β_1	0.69	0.117	1.08	0.013	1.23	0.060	1.06	0.010
		γ	0.63	1.874	1.98	0.007	1.60	0.167	1.70	0.093
	4	β_1	-0.24	1.681	1.15	0.031	1.38	0.149	0.95	0.022
		γ	0.39	2.600	2.67	0.562	1.00	0.989	0.95	1.358
	12	β_1	0.80	0.043	1.05	0.006	1.17	0.031	1.05	0.005
		γ	0.88	1.257	2.10	0.019	2.05	0.019	1.98	0.009
200	8	β_1	0.66	0.128	1.06	0.006	1.22	0.051	1.04	0.004
		γ	0.60	1.965	2.06	0.007	1.81	0.040	1.73	0.076
	4	β_1	-0.43	2.122	1.10	0.015	1.36	0.134	0.94	0.010
		γ	0.39	2.605	2.49	0.267	1.07	0.863	0.96	1.073

Table 7 - Estimates (E) of the parameters (P) and mean square error (MSE) considering samples simulated of size (n) 50, 100 and 200 of Weibull distribution with $\gamma = 2$, $\beta_0 = 0$ and $\beta_1 = 1$, three intervals (G = 12, 8 and 4) and the methods of estimation M1, M2, M3 and M4 for 40% of censoring

n	G	P	M1		M2		M3		M4	
			E	MSE	E	MSE	E	MSE	E	MSE
	12	β_1	0.86	0.037	1.06	0.014	1.18	0.042	1.06	0.015
		γ	1.08	0.860	2.14	0.053	2.11	0.072	1.80	0.069
50	8	β_1	0.71	0.118	1.05	0.014	1.24	0.064	1.07	0.015
		γ	0.78	1.500	2.08	0.028	1.79	0.097	1.49	0.280
	4	β_1	-0.32	2.006	1.06	0.019	1.46	0.223	1.06	0.034
		γ	0.46	2.360	2.15	0.096	0.96	1.089	0.82	3.258
	12	β_1	0.91	0.016	1.10	0.014	1.20	0.044	1.08	0.012
		γ	1.38	0.411	1.95	0.030	1.57	0.210	1.63	0.155
100	8	β_1	0.80	0.056	1.10	0.017	1.26	0.073	1.09	0.014
		γ	0.95	1.118	1.97	0.018	1.39	0.384	1.46	0.317
	4	β_1	-0.02	1.185	1.14	0.027	1.44	0.202	0.94	0.040
		γ	0.46	2.368	2.23	0.111	0.80	1.447	0.64	2.492
	12	β_1	0.91	0.011	1.10	0.013	1.22	0.048	1.10	0.013
		γ	1.10	0.815	2.12	0.029	1.85	0.040	1.71	0.093
200	8	β_1	0.81	0.043	1.10	0.014	1.28	0.078	1.11	0.015
		γ	0.78	1.488	2.06	0.009	1.59	0.174	1.44	0.320
	4	β_1	-0.09	1.257	1.12	0.020	1.49	0.241	1.09	0.016
		γ	0.46	2.388	2.22	0.074	0.87	1.280	0.65	2.003

5 Illustrative example

This application is part of an experimental study performed with a flax cultivar, susceptible to pathogen *Fusarium oxysporum* (Höper et al., 1995). A sample of 286 susceptible cultivars of flax were randomly selected for the three treatments (different soil-clay mixtures: MI, MK and MM) and a natural soil (M). The response was the time taken until the occurrence of the wilt caused by the pathogen. Each cultivar was evaluated twice a week during the period of 52 days, resulting in 10 time intervals. The Weibull model was fitted to these data, considering the covariate substratum.

Table 8 presents the results obtained for the five estimation methods. The corresponding values of shape parameter (γ) are 4.93, 5.35, 5.78 and 5.95 for the methods $M1$, $M2$, $M3$ and $M4$, respectively, or rather, the failure rate is increasing. The methods $M2$ and $M4$ present similar results, and method $M3$ presents smaller standard deviation than method $M1$. Those results are in agreement with the discussions in Section 4. That is, $M2$ is the best approximation method since its results are very closer to the exact one ($M4$).

Table 8 - Parameter (P) estimates (E) and standard deviation (SD) for the covariate substratum for four estimation methods

P	M1		M2		M3		M4	
	E	SD	E	SD	E	SD	E	SD
β_0	3.570	0.025	3.620	0.023	3.667	0.021	3.620	0.023
β_1	0.115	0.035	0.105	0.032	0.097	0.030	0.106	0.032
β_2	0.050	0.035	0.045	0.032	0.040	0.030	0.045	0.032
β_3	0.133	0.034	0.124	0.032	0.116	0.029	0.124	0.032

Conclusions

In this paper we considered four methods to analyze grouped survival data: the discrete model and three simple imputation methods commonly used by analysts. Simulations to compare these methods were performed by considering different sample sizes, proportion of censoring and number of intervals.

From the simulation results, as well as from the illustrative example in Section 5, the main conclusions were: i) method $M2$ (midpoint imputation) gives very similar results as the method $M4$ (discrete model); ii) methods $M2$ and $M4$ were those which, in general, provide parameter estimates closest to the true parameter values, particularly as the sample size and the number of intervals increase, and the proportion of censoring decreases; iii) method $M1$ seems usually to underestimate the parameters and $M3$ overestimate; iv) the mean square error (MSE) decreases as the sample sizes increase, and also increases as the proportion of censoring increases; v) in the covariate case, MSE were, in general, larger suggesting that covariates can produce some changes in the parameters estimates; vi) MSE also increases as the number of intervals decreases, which can be a consequence of the increase in the number of ties and in the width of the intervals.

Similarly to Law and Brookmeyer (1992), we could notice from our simulation study that the midpoint imputation can be a reasonable procedure to analyze grouped survival data. Our study, however, also indicated that some cautions are needed concerning the use of this imputation method because in some cases it might lead to biased estimates. For instance, high values of MSE can be obtained when we have few intervals. In these cases, intervals are usually wide and a large proportion of ties can fall just into one or two intervals leading to biased estimates. For several other situations, however, our study suggested that the midpoint imputation can be a reasonable procedure for analyzing grouped survival data. Compared with method $M4$, a feature that makes the midpoint imputation attractive for situations where it can be appropriate is that techniques for evaluating model adequacy are available for standard survival models while for $M4$ additional studies remain needed for such purpose.

Acknowledgments

This work was supported by CNPq-Brazil and is part of the PhD thesis of the first author at the Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, Brazil. The second, third and fourth authors also thank the CNPq, a Brazilian science funding agency, for the partial financial support.

STRAPASSON, E.; COLOSIMO, E. A.; GIOLO, S. R.; DEMÉTRIO, C. G. B. Um estudo de simulação para comparar métodos de imputação para dados de sobrevivência agrupados. *Rev. Bras. Biom.,* São Paulo, v.27, n.2, p.210-224, 2009.

- RESUMO: Dados de sobrevivência agrupados ou discretos ocorrem quando todas as unidades experimentais são visitadas nos mesmos tempos, que podem ser equidistantes ou não. Para analisar dados dessa natureza, métodos de imputação simples são frequentemente utilizados pelos analistas. Nesse artigo, simulações de Monte Carlo são realizadas para comparar métodos de imputação que usam o ponto médio, o limite inferior e o limite superior dos intervalos. O modelo de regressão Weibull para dados de sobrevivência discretos também é considerado. Nas simulações, são usados três tempos de visita equidistantes, três proporções de censura, três tamanhos amostrais e a distribuição Weibull. Também são consideradas no estudo, as situações sem e com covariável. Os resultados das simulações mostram que o ponto médio é o melhor método de imputação dentre aqueles considerados, apresentando resultados similares aos do modelo de regressão Weibull para dados de sobrevivência discretos. Dados de um estudo realizado com cultivar de linho ilustra a comparação dos métodos considerados no artigo para a análise de dados de sobrevivência agrupados.

- PALAVRAS-CHAVE: Tempo discreto; simulação de Monte Carlo; modelo Weibull.

References

COLLETT, A. *Modelling survival data in medical research*. London: Chapman and Hall, 2003. 391p.

COLOSIMO, E. A.; CHALITA, L. V. A. S.; DEMÉTRIO, C. G. B. Tests of proportional hazards and proportional odds models for grouped survival data. *Biometrics*, Washington, v.56, p.1233-1240, 2000.

GIOLO, S. R.; COLOSIMO, E. A.; DEMÉTRIO, C. G. B. Different approaches for modelling grouped survival data: A mango tree study. *J. Agric. Biol. and Environ. Stat.*, Alexandria, v.14, p.154-169, 2009.

HÖPER, H.; STEINBERG, C.; ALABOUVETTE, C. Importance of physical and chemical soil properties in the suppressiveness of soils to plant diseases. *Eur. J. Soil Biol.*, Gauthier Villars, v.32, p.41-58, 1995.

HSU, C-H.; TAYLOR, J. M. G.; MURRAY, S.; COMMENGES, D. Multiple imputation for interval censored data with auxiliary variables. *Stat. Med.*, Chichester, v.26, p.769-781, 2007.

KIM, M. Y.,; DE GRUTTOLA, V. G.; LAGAKOS, S. W. Analyzing doubly censored data with covariates, with application of AIDS. *Biometrics*, Washington, v.49, p.13-22, 1993.

LAW, C. G.; BROOKMEYER, R. Effects of mid-point imputation on the analysis of doubly censored data. *Stat. Med.*, Chichester, v.11, p.1569-1578, 1992.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New Jersey: John Wiley & Sons, 2003. 630p.

LINDSEY, J. C.; RYAN, L. M. Tutorial in biostatistics methods for interval-censored data. *Stat. Med.*, Chichester, v.17, p.219-238, 1998.

PETKAU, A. J.; SITTER, R. R. Models for quantal response experiments over time. *Biometrics*, Washington, v.45, p.1299-1308, 1989.

PRENTICE, R. L.; GLOECKLER, L.A. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, Washington, v.34, p.57-67, 1978.

R DEVELOPMENT CORE TEAM. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em nov. 2008.

SUN, J. A. Nonparametric test for interval-censored failure data with application to AIDS studies. *Stat. Med.*, Chichester, v.15, p.1387-1395, 1996.

WHITEHEAD, J. The analysis of relapse clinical trials with application to a comparison of two ulcer treatments. *Stat. Med.*, Chichester, v.8, p.1439-1454, 1989.

Received in 21.04.2009.

Approved after revised in 25.08.2009.