Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Analysis of discrete lifetime data under middle-censoring and in the presence of covariates

S. Rao Jammalamadaka[a] & Elvynna Leong[a]
[a] Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA
Published online: 22 Dec 2014.

Journal of Applied Statistics, 2014
http://dx.doi.org/10.1080/02664763.2014.993364

Taylor & Francis
Taylor & Francis Group

Analysis of discrete lifetime data under middle-censoring and in the presence of covariates

S. Rao Jammalamadaka and Elvynna Leong*

Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA

(*Received 23 February 2014; accepted 26 November 2014*)

'Middle censoring' is a very general censoring scheme where the actual value of an observation in the data becomes unobservable if it falls inside a random interval (L, R) and includes both left and right censoring. In this paper, we consider discrete lifetime data that follow a geometric distribution that is subject to middle censoring. Two major innovations in this paper, compared to the earlier work of Davarzani and Parsian [3], include (i) an extension and generalization to the case where covariates are present along with the data and (ii) an alternate approach and proofs which exploit the simple relationship between the geometric and the exponential distributions, so that the theory is more in line with the work of Iyer *et al.* [6]. It is also demonstrated that this kind of discretization of life times gives results that are close to the original data involving exponential life times. Maximum likelihood estimation of the parameters is studied for this middle-censoring scheme with covariates and their large sample distributions discussed. Simulation results indicate how well the proposed estimation methods work and an illustrative example using time-to-pregnancy data from Baird and Wilcox [1] is included.

Keywords: middle censoring; EM algorithm; accelerated failure time model; exponential distribution; geometric distribution; discrete censoring

1. Introduction

Middle censoring is a very general censoring scheme introduced in [7]. It refers to situations where some of the observations become unobservable because they happen to fall within a random censoring interval. For some individuals, the exact values are available while for others, the corresponding intervals are observed. If a subject is temporarily absent or withdrawn from study and the event of interest occurs during this time interval, the exact time of occurrence cannot be observed but instead only the censoring interval is observed. The more commonly studied

*Corresponding author. Emails: leong@pstat.ucsb.edu, elvynna.leong@ubd.edu.bn

right-censoring and left-censoring can be viewed as special cases of this middle censoring by suitable choice of this censoring interval. Some examples where middle-censoring may occur are (i) equipment failure that could occur during a period where observation is not possible or is not being made (ii) in biomedical studies, a patient under observation may be absent from study for a short period during which time the event of interest may occur, for example, in the study of African infant precocity by Leiderman *et al.* [9] where the time from birth to the learning time was the variable of interest. If the observation was not possible during a fixed time interval (a random interval relative to the individual's lifetime) such as the temporary closure of the clinic due to an outbreak of war say, and some children developed the skill during this time, the exact age of these children at the time of skill development are not observed but rather only the information that the event of interest occurred during this time interval. Other authors have studied this situation under the labels *partially interval-censored data* and *mixed interval-censored data* (see Huang [5] and Yu et al. [11]).

The middle-censoring scheme can be described in the following notations. Assume that there is a random sample of individuals of size n from a specific population with lifetimes $T_1, T_2, \ldots, T_n$, where not all these T_i are observable. Corresponding to the ith individual in the sample, there is a random censoring interval (L_i, R_i) which is independent of the lifetime so that the observed data X_i's are given by

$$X_i = \begin{cases} T_i & \text{if } T_i \notin (L_i, R_i), \\ (L_i, R_i) & \text{if } T_i \in (L_i, R_i) \end{cases}$$

for $i = 1, 2, \ldots, n$.

Jammalamadaka and Mangalam [7] developed a self-consistent estimator and the nonparametric maximum likelihood estimator for the middle-censored data. Jammalamadaka and Iyer [8], then suggested an approximate self-consistent estimator and established its weak convergence. Iyer *et al.* [6] considered middle-censoring scheme in a parametric set-up when the lifetime distributions are exponentially distributed. They consider maximum likelihood and Bayes estimates of the relevant parameters. Middle-censoring in a discrete set-up was discussed by Davarzani and Parsian [3] (DP[3] from now on) where the lifetimes as well as the lower limit and length of censoring interval are assumed to have geometric distributions.

In this paper, we consider lifetimes that follow a geometric distribution as in DP [3] but we generalize their set-up to the important case where covariates are present as well as provide alternate results and proofs by exploiting the elegant relationship between the exponential and geometric distributions. In Section 2.1, we discuss this connection and use it in Section 2.2 to find the maximum likelihood estimates (MLEs) under middle-censoring in the presence of covariates using the accelerated failure time model for the geometric case and discuss the EM algorithm for obtaining them. The novelty of our approach, contrasted with that in DP [3], is to adapt the methods of Iyer *et al.* [6] to the geometric case. Simulation studies are carried out in support of the theory to indicate how well the proposed estimation methods work. We also consider the asymptotic distribution of the MLE in terms of Fisher information. Section 3 illustrates the application of the proposed model to time-to-pregnancy study from Baird and Wilcox [1].

2. An alternate approach to discrete lifetimes and with covariates

In this section, first we show how one can utilize the connection between the geometric and exponential distributions so that the results in DP can be subsumed by what has been done in [6] for exponential data. This connection will also allow us to more readily extend the results to the case of covariates, as we do in Section 2.2.

2.1 *An important link*

As is known, the geometric distribution is the discrete analogue of the exponential distribution and the following well-known lemma provides the elegant connection between the two. We add the short proof for completeness and to describe the notations:

LEMMA 1 *If X is an exponentially distributed random variable with parameter λ, then $Y = \lfloor X \rfloor$ where $\lfloor\ \rfloor$ is the floor function (the integer part of x), is a geometrically distributed random variable with parameter $p = 1 - e^{-\lambda}$.*

Proof Let $X \sim \exp(\lambda)$ with p.d.f. $f(x) = \lambda e^{-\lambda x}$. Suppose we have $Y = \lfloor X \rfloor$. Then,

$$P(\lfloor X \rfloor = a) = P(a \leq X < a + 1) = e^{-a\lambda}(1 - e^{-\lambda}) = (1 - p)^a p$$

Therefore, $Y = \lfloor X \rfloor \sim$ geometric(p) where $p = 1 - e^{-\lambda}$, $a = 0, 1, 2, \ldots$, $0 \leq p \leq 1$ and $\lambda > 0$. ■

The geometric distribution also inherits the interesting property known as the memoryless property which the exponential distribution has. For integers $s > t$, it is the case that

$$P(X > s | X > t) = P(X > s - t),$$

that is, the geometric distribution 'forgets' what has occurred. The probability of getting an additional $s - t$ failures, having already observed t failures is the same as the probability of observing $s - t$ failures at the start of the sequence.

Applying the property of memorylessness and using the relationship of exponential and geometric distributions from Lemma 1 to middle censoring, the geometric lifetimes can be generated from the exponentially distributed lifetime, $T_i \sim Exp(\lambda)$.The geometric distributed lifetimes is $Y_i = \lfloor T_i \rfloor \sim$ geometric(p) with probability function

$$P(Y_i = y_i) = p(1 - p)^{y_i}$$

for $y_i = 0, 1, 2, \ldots$ and $p = 1 - e^{-\lambda}$.

The left point of the censored interval is $U_i = \lfloor L_i \rfloor \sim$ geometric(p_u) with probability function

$$P(U_i = u_i) = p_u(1 - p_u)^{u_i},$$

where $L_i \sim \text{Exp}(\alpha)$, $p_u = 1 - e^{-\alpha}$ and $u_i = 0, 1, 2, \ldots$ while the length of the censored interval is $W_i = \lfloor S_i \rfloor \sim$ geometric(p_w) with probability function

$$P(W_i = w_i) = p_w(1 - p_w)^{w_i - 1},$$

where $S_i \sim \text{Exp}(\beta)$, $S_i = R_i - L_i$, $p_w = 1 - e^{-\beta}$, $w_i = 1, 2, \ldots$ and $W_i = V_i - U_i$, where V_i is the right point of the censored interval. The lifetimes Y_i, U_i and W_i are independent for all i.

2.2 *Geometric model in the presence of covariates*

In this section, we consider a geometric lifetime with middle censoring in the presence of covariates. The covariates that we consider here are fixed, that is, known at baseline or entry to the study. The relationship between an exponential distribution and the geometric distribution discussed in Section 2.1 can be applied here for a geometric lifetime in the presence of covariates. Here, each person has a survival time, T_i and covariates specific to that individual $\boldsymbol{Z}_i$.

It may be recalled that when the baseline distribution is an exponential, the Cox proportional hazard assumption is equivalent to the accelerated failure time assumption. See, for example, Cox and Oakes [2, p. 70–72] who show that the exponential regression model is an example of an accelerated failure time model with proportional hazards. Hence, the lifetimes, T_i are first generated from an exponential accelerated failure model or a Cox PH model when the baseline distribution is the exponential distribution, that is, $T_i \sim \text{Exponential}(\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i})$ with p.d.f.

$$f(t|\boldsymbol{Z}_i) = \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} \exp(-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} t)$$

where $\boldsymbol{\theta}$ is the effect of each covariate $\boldsymbol{Z}$ and the superscript T stands for transpose operation. Hence, one can generate geometric lifetimes from the generated exponential lifetime, that is, $Y_i = \lfloor T_i \rfloor \sim \text{geometric}(p_i)$, where $p_i = 1 - e^{-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}$. We take the left end point of the censored interval $\boldsymbol{U}_i \sim \text{geometric}(p_u)$ while the width of the censored interval is taken to be $\boldsymbol{W}_i \sim \text{geometric}(p_w)$, where $\boldsymbol{W}_i = \boldsymbol{V}_i - \boldsymbol{U}_i$ and $\boldsymbol{V}_i$ is the right-censored point of the censored interval.

Since the model is completely parametric, the likelihood can be written down and the MLE of p can be solved. Suppose that there are $n_1 > 0$ uncensored observations and $n_2 > 0$ censored observations, where $n = n_1 + n_2$. After re-ordering the data, without loss of generality, it is assumed that the first n_1 are the uncensored observations while the remaining n_2 are the censored observations. Hence, the observed data are

$$\{Y_1, Y_2, \ldots, Y_{n_1}, [U_{n_1+1}, V_{n_1+1}], [U_{n_1+2}, V_{n_1+2}], \ldots, [U_{n_1+n_2}, V_{n_1+n_2}]\}.$$

Similar to the methods used in DP [3] in Section 2, the likelihood function of the observed data is written as

$$L(p_i) = c p_i^{n_1} (1-p_i)^{(\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_1+n_2} u_i)} \prod_{i=n_1+1}^{n_1+n_2} (1 - (1-p_i)^{w_i+1}), \tag{1}$$

where $c = c_1^{n_2} c_2^{n_1}$ is the normalizing constant which does not depend on p_i, where $p_i = 1 - e^{-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}$. From Equation (1), the log-likelihood function of p_i is

$$l(p_i) = \ln(c) + n_1 \ln(p_i) + \sum_{i=1}^{n_1} y_i \ln(1-p_i) + \sum_{i=n_1+1}^{n_1+n_2} u_i \ln(1-p_i) + \sum_{i=n_1+1}^{n_1+n_2} \ln(1 - (1-p_i)^{w_i+1}). \tag{2}$$

Applying the EM algorithm to find the MLE of p, the following conditional expectation is required:

$$\begin{aligned} E_p(Y|U \le Y \le V) &= \sum_{y=U}^{V} y P_p(Y = y|U \le Y \le V) \\ &= \left[U + \frac{(1-p) - (W+1)(1-p)^{W+1} + W(1-p)^{W+2}}{(1-(1-p)^{W+1})(1-(1-p))} \right]. \end{aligned} \tag{3}$$

Equation (3) is used as the E-step in the EM-algorithm and the pseudo-log-likelihood is given as

$$l^*(p_i) \propto n \ln(p_i) + \sum_{i=1}^{n_1} y_i \ln(1-p_i) + \sum_{i=n_1+1}^{n_1+n_2} y_i^* \ln(1-p_i), \tag{4}$$

where

$$y_i^* = E_{p_i}(Y_i|U_i \le Y_i \le V_i) = \left[U_i + \frac{(1-p_i) - (W_i+1)(1-p_i)^{W_i+1} + W_i(1-p_i)^{W_i+2}}{(1-(1-p_i)^{W_i+1})(1-(1-p_i))} \right].$$

Thus, the EM algorithm can be set up as follows. Choose p_0 to be the MLE of the uncensored data. Update the estimates with the following steps.

- *Step 1* Suppose that $p_{(j)}$ is the jth estimate.
- *Step 2* Compute Y_i^* by using Equation (3) with $p = p_{(j)}$.
- *Step 3* Solve Equation (4) for its maximum and set $p_{(j+1)}$ as that maximum.
- *Step 4* Repeat until a convergence criterion is met.

A simulation study is performed to illustrate the usefulness of this approach. Simulations are carried out in R using $N = 100$ replications with a common sample size $n = 250$. Each sample is then censored and the EM algorithm described above is applied to the censored data. The censoring mechanism is as follows; the left end point of the censored interval is Geometric distributed with mean 0.5 and the length of the censored interval is also Geometric distributed but with mean 0.1. Three covariates are used in this simulation. The covariates Z_1 and Z_2 are generated from a Binomial distribution with one trial and probability of success equal to 0.5. The third covariate, Z_3 is generated from a standard Normal distribution. Three cases for the true covariate effects are considered here, similar to Pan [10]. They are $\theta = (1, 1, 1), \theta = (1, 0, 0)$ and $\theta = (0, 0, 1)$ and are chosen since they represent the case where the covariates have an equal effect, where only one Bernoulli covariate has one effect and where only the Normally distributed covariate had an effect. The true values of λ are chosen to be 0.5 and 0.3 as given in Table 1. A

Table 1. Simulation results for the geometric model in the presence of three covariates.

Parameter	True Value	MLE	SD	EMSE	C.I.	Censored Prop
λ	0.5	0.5301	0.0102	0.0072	(0.5101, 0.5501)	0.1562
θ_1	1.0	1.0308	0.0286	0.0087	(0.9747, 1.0869)	
θ_2	1.0	1.0387	0.0211	0.0088	(0.9973, 1.0801)	
θ_3	1.0	1.0597	0.0174	0.0026	(1.0256 1.0938)	
λ	0.5	0.5321	0.0080	0.0127	(0.5164, 0.5478)	0.2058
θ_1	1.0	1.0715	0.0266	0.0051	(1.0194,1.1237)	
θ_2	0.0	−0.0171	0.0289	0.0030	(−0.0737,0.0395)	
θ_3	0.0	−0.0052	0.0261	0.0033	(−0.0564, 0.0460)	
λ	0.5	0.5410	0.0210	0.0159	(0.4998,0.5822)	0.2596
θ_1	0.0	0.0056	0.0464	0.0035	(−0.0853, 0.0965)	
θ_2	0.0	−0.0085	0.0365	0.0021	(−0.0800, 0.0630)	
θ_3	1.0	1.1155	0.0158	0.0133	(1.0845, 1.1465)	
λ	0.3	0.3384	0.0110	0.0015	(0.3168, 0.3600)	0.1972
θ_1	1.0	1.0187	0.0366	0.0030	(0.9470, 1.0904)	
θ_2	1.0	1.1822	0.0279	0.0032	(1.1275, 1.2369)	
θ_3	1.0	1.0515	0.0175	0.0026	(1.0172, 1.0858)	
λ	0.3	0.3529	0.0301	0.0028	(0.2939, 0.4119)	0.2952
θ_1	1.0	1.1160	0.0710	0.0134	(0.9768, 1.2552)	
θ_2	0.0	−0.0283	0.0510	0.0008	(−0.1283, 0.0717)	
θ_3	0.0	−0.0029	0.0333	0.0018	(−0.0682, 0.0624)	
λ	0.3	0.3364	0.0200	0.0013	(0.2972, 0.3756)	0.2982
θ_1	0.0	0.0144	0.0594	0.0042	(−0.01020, 0.1308)	
θ_2	0.0	0.0158	0.0413	0.0022	(−0.0651, 0.0967)	
θ_3	1.0	1.0590	0.0155	0.0035	(1.0286, 1.0894)	

number of different starting points were used in the EM-algorithm in order to capture the global maximum.

The 'MLE' reported is the average value of the $N = 100$ estimates obtained and the estimated mean-squared error, EMSE is calculated using the equation

$$\text{EMSE}(\hat{p}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{p} - p)^2$$

where $\hat{p}$ is the estimate of p and a total of N simulation were computed. The standard deviation (SD) of the estimates is evaluated in the simulation and their confidence intervals (C.I.) could be evaluated. The 'Censored proportion' line in the table gives the mean proportion of censoring in the $N = 100$ simulated samples.

In the $N = 100$ simulations, the samples were found to be between 14% and 29% censored. The MLEs of $\lambda, \theta_1, \theta_2, \theta_3$ were computed using the EM algorithm described above. See Table 1 for the results from these simulations. The MLEs are fairly close to the actual value and the EMSE are small. This approach yields very useful, accurate and reliable results. Note that we initialized the EM-algorithm from a number of different starting points and it shows that the likelihood does have a unique maximum.

2.3 *The case of no covariates*

The case where there are no covariates which is considered in DP[3] comes out as a special case of what we already have, by taking $Z = 0$. In this case, the likelihood function of the observed data is written as in Equation (1) but with $p = 1 - \mathrm{e}^{-\lambda}$. The log-likelihood function of p, the conditional expectation for the E-step of the EM-algorithm and the pseudo-log-likelihood are shown in Equations (2)–(4), respectively. Hence, the EM algorithm is set up as follows. Choose $p_{(0)}$ to be the MLE of the uncensored data, that is, $p_{(0)} = n_1/(n_1 + \sum_{i=1}^{n_1} y_i)$. Update the estimates with the following steps.

- *Step 1* Suppose that $p_{(j)}$ is the jth estimate
- *Step 2* Compute Y_i^* by using Equation (3) with $p = p_{(j)}$
- *Step 3* Set $p_{(j+1)} = n/(n + \sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_1+n_2} y_i^*)$
- *Step 4* Repeat until convergence is met.

Simulations are run in order to test the validity of the program. We considered different sample sizes namely $n = 50$, 100, 250 and 500. For each sample size n, $N = 100$ samples were simulated. Each sample was then censored and the EM algorithm described above was applied to the censored data. See Table 2 for the results from these simulations. The 'p est' reported is the average value of the $N = 100$ estimates obtained. The geometric model converges numerically to the true value in all cases, which is consistent with the result found in DP [3] for the geometric lifetimes. The estimates are converging to the true values as the sample size n increases but it appears to converge rather slowly.

2.4 *Asymptotic distribution of the MLE*

It can be checked that the conditions for the validity of the properties of the MLEs, hold. For completeness, we give below the derivatives of the log-likelihood function from Equation (2),

Table 2. Simulation results for Geometric (0.3) lifetimes.

n	(p_l, p_z)	(0.5,0.9)	(0.2,0.9)	(0.3,0.8)
50	p est	0.3159	0.3109	0.3097
	EMSE of p	0.0017	0.0019	0.0018
	SD	0.0288	0.0276	0.0309
	Censored proportion	0.1526	0.1538	0.0912
	C.I.	(0.2595, 0.3723)	(0.2568, 0.3650)	(0.2491, 0.3703)
100	p est	0.3092	0.3095	0.3086
	EMSE of p	0.0009	0.0008	0.0008
	SD	0.0236	0.0229	0.0227
	Censored proportion	0.1438	0.1628	0.1003
	C.I.	(0.2629, 0.3555)	(0.2646, 0.3544)	(0.2641, 0.3531)
250	p est	0.3056	0.3077	0.3050
	EMSE of p	0.0005	0.0006	0.0005
	SD	0.0178	0.0176	0.0170
	Censored proportion	0.1471	0.1592	0.0968
	C.I.	(0.2707, 0.3405)	(0.2732, 0.3422)	(0.2717, 0.3383)
500	p est	0.3054	0.3065	0.3035
	EMSE of p	0.0002	0.0002	0.0002
	SD	0.0136	0.0135	0.0131
	Censored proportion	0.1468	0.1605	0.0946
	C.I.	(0.2787, 0.3321)	(0.2770, 0.3330)	(0.2778, 0.3292)

where $p = 1 - \mathrm{e}^{-\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}}$:

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{n_1} \frac{\mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}}{\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} - 1} - \sum_{i=1}^{n_1} y_i \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} - \sum_{i=n_1+1}^{n_1+n_2} u_i \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} + \sum_{i=n_1+1}^{n_1+n_2} \frac{\mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (w_i + 1)}{\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (w_i + 1)} - 1}$$

and for $j = 1, 2, 3$,

$$\frac{\partial l}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{n_1} \frac{\boldsymbol{Z}_j \lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}}{\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} - 1} - \sum_{i=1}^{n_1} y_i \boldsymbol{Z}_j \lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} - \sum_{i=n_1+1}^{n_1+n_2} u_i \boldsymbol{Z}_j \lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} + \sum_{i=n_1+1}^{n_1+n_2} \frac{\boldsymbol{Z}_j \lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (w_i + 1)}{\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (w_i + 1)} - 1}.$$

The second derivatives are given by

$$\begin{aligned}
\frac{\partial^2 l}{\partial \lambda^2} &= -\sum_{i=1}^{n_1} \frac{\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} + 2(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i)}}{(\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} - 1)^2} - \sum_{i=n_1+1}^{n_1+n_2} \left(\frac{(w_i + 1)^2 \mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (w_i + 1) + 2(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i)}}{(\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (w_i + 1)} - 1)^2} \right) \\
\frac{\partial^2 l}{\partial \lambda \partial \theta_j} &= -\sum_{i=1}^{n_1} \frac{Z_j \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (\lambda \mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} - \mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} + 1)}{(\mathrm{e}^{\lambda \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} - 1)^2} - \sum_{i=1}^{n_1} y_i Z_j \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} - \sum_{i=n_1+1}^{n_1+n_2} u_i Z_j \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} \\
&\quad - \sum_{i=n_1+1}^{n_1+n_2} \frac{(w_i + 1) Z_j \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} (\lambda (w_i + 1) \mathrm{e}^{\lambda (w_i + 1) \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i} - \mathrm{e}^{\lambda (w_i + 1) \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} + 1)}{(\mathrm{e}^{\lambda (w_i + 1) \mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{Z}_i}} - 1)^2}
\end{aligned} \tag{5}$$

$$\frac{\partial^2 l}{\partial \theta_j^2} = -\sum_{i=1}^{n_1} \frac{Z_j^2 \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i} (\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i + \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - e^{\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} + 1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - 1)^2} - \sum_{i=1}^{n_1} y_i Z_j^2 \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i} - \sum_{i+n_1+1}^{n_1+n_2} u_i Z_j^2 \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}$$
$$- \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_j^2 \lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i} (\lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i + \lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} + 1)}{(e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - 1)^2} \tag{6}$$

$$\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{n_1} \frac{Z_j Z_k \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i} (\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i + \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - e^{\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} + 1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - 1)^2} - \sum_{i=i}^{n_1} y_i Z_j Z_k \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}$$
$$- \sum_{i=n_1+1}^{n_1+n_2} u_i Z_j Z_k \lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}$$
$$- \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_j Z_k \lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i} (\lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i + \lambda (w_i+1) e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - e^{\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} + 1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \mathbf{Z}_i}} - 1)^2}. \tag{7}$$

By substituting the MLE found by using the algorithm above into the information matrix, we obtain the 'observed information' matrix, namely the Hessian matrix of the log-likelihood function (see Efron and Hinkley [4]) as follows:

$$\hat{I}_{4\times 4} = \begin{bmatrix} \frac{\partial^2 l}{\partial \lambda^2} & \frac{\partial^2 l}{\partial \lambda \partial \theta_1} & \frac{\partial^2 l}{\partial \lambda \partial \theta_2} & \frac{\partial^2 l}{\partial \lambda \partial \theta_3} \\ \frac{\partial^2 l}{\partial \theta_1 \partial \lambda} & \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_3} \\ \frac{\partial^2 l}{\partial \theta_2 \partial \lambda} & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_2^2} & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_3} \\ \frac{\partial^2 l}{\partial \theta_3 \partial \lambda} & \frac{\partial^2 l}{\partial \theta_3 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_3 \partial \theta_2} & \frac{\partial^2 l}{\partial \theta_3^2} \end{bmatrix},$$

where $\lambda = \hat{\lambda}$ and $\theta_i = \hat{\theta}_i$. Hence, $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ is asymptotically Normal with mean zero and covariance $I(\boldsymbol{\theta})^{-1}$. This large sample approximation can be used to construct the required C.I., as we do in Section 2.2 and the ensuing illustration.

3. A practical example

In this section, we apply the proposed techniques to a time-to-pregnancy data in [1]. The study classified women as either smokers or non-smokers where a 'current smoker' is defined as a woman who smokes an average of one or more cigarettes per day during the first cycle in which she was trying to get pregnant. A total of 586 women were asked the number of cycles required to conceive. A couple is considered clinically infertile after 12 unsuccessful cycles, so medical interventions tend to begin after the 12th unsuccessful cycle. The number of menstrual cycles to pregnancy is our discrete survival time and we take the variable *smoke* (Z_1) as the covariate with regression coefficient θ_1.

For the complete data set, it is observed that the MLEs of λ and θ_1 are 0.3150 and -0.3136, respectively. In order to create a set of middle-censored data, we randomly choose several actual failure data and replace them by random censoring intervals. The data were censored by a random interval whose left end was a geometric random variable with mean 5 and the width was geometric with mean 10. It is found that 26.79% of data were censored resulting in 429 uncensored

observations and 157 censored observations. Applying the model given in Section 2.2, it is found that the estimates of the regression coefficients are $\hat{\lambda} = 0.3045$ and $\hat{\theta}_1 = -0.3189$. The 95% C.I. based on the asymptotic distribution of λ and θ_1 are $(0.2754, 0.3336)$ and $(-0.5446, -0.0932)$, respectively.

In order to assess how much of a change it makes in the estimates or C.I. when one uses the discretized geometric distribution in lieu of the original exponential distribution, we fit this model with the exponential distribution instead of the geometric distribution. The estimates of the regression coefficients are $\hat{\lambda} = 0.3303$ and $\hat{\theta}_1 = -0.3343$. The 95% C.I. based on the asymptotic distribution of λ and θ_1 are (0.3181,0.3425) and $(-0.5404, -0.1282)$, respectively. The data were censored exactly like the geometric case resulting in 27.13% censored observations, specifically 427 uncensored observations and 159 censored observations. These comparisons show that the estimates are very close as are the C.I.

4. Conclusion

We considered inference for discrete lifetimes, when the data are middle-censored and extend it to the case when covariates are present. We validate and confirm the estimation and inference procedures discussed, from extensive simulation studies, which show that the MLEs of the regression coefficients are very close to the true values in all the cases. The model is applied to a real data set on Stanford heart transplant survival and is shown to give very meaningful and useful results.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] D.D. Baird and A.J. Wilcox, *Cigarette smoking associated with delayed conception*, J, Am. Med. Assoc. 253 (1985), pp. 2979–2983.
[2] D.R. Cox and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, London, 1984.
[3] N. Davarzani and A. Parsian, *Statistical inference for discrete middle-censored data*, J. Statist. Plan. Inference 141 (2011), pp. 1455–1462.
[4] B. Efron and D.V. Hinkley, *Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information*, Biometrika 65 (1978), pp. 457–482.
[5] J. Huang, *Asymptotic properties of nonparametric estimation based on partly interval-censored data*, Statist. Sin. 9 (1999), pp. 501–519.
[6] S.K. Iyer, S.R. Jammalamadaka, and D. Kundu, *Analysis of middle censored data with exponential lifetime distributions*, J. Statist. Plan. Inference 138 (2008), pp. 3550–3560.
[7] S.R. Jammalamadaka and V. Mangalam, *Nonparametric estimation for middle-censored data*, J. Nonparametr. Stat. 15 (2003), pp. 253–265.
[8] S.R. Jammalamadaka and S.K. Iyer, *Approximate self consistency for middle censored data*, J. Statist. Plan. Inference 124 (2004), pp. 75–86.
[9] P.H. Leiderman, D. Babu, J. Kagia, H.C. Kraemer, and G.F. Leiderman, *African infant precocity and some social influences during the first year*, Nature 242 (1973), pp. 247–249.
[10] W. Pan, *Extending the iterative convex minorant algorithm to the Cox model for interval-censored data*, J. Comput. Graph. Stat. 8 (1999), pp. 109–120.
[11] Q. Yu, G.Y.C. Wong, and L. Li, *Asymptotic properties of self-consistent estimators with mixed interval-censored data*, Ann. Inst. Statist. Math. 53 (2001), pp. 469–486.