Likelihood Inference for Discrete Weibull Data with Left Truncation and Right Censoring

Chaobing He

School of Mathematics and Statistics, Anyang Normal University, Anyang 455000, China

Email: chaobing5@163.com

Abstract: The discrete Weibull distribution is a very popular distribution for modeling discrete lifetime data, and it is obtained by discretizing Weibull distribution. Left truncation and right censoring are often observed in lifetime data. Here, the EM algorithm is applied to estimate the model parameters of the discrete Weibull distribution fitted to data containing left truncation and right censoring. The maximization part of the EM algorithm is carried out using the ECM algorithm. The discrete Weibull distribution is also fitted using the Newton-Raphson(NR) method. The asymptotic variance-covariance matrix of the MLEs under the EM framework is obtained through the missing information principle, and asymptotic confidence intervals for the parameters are then constructed.

Keywords: Maximum likelihood estimate; EM algorithm;Lifetime data; Missing information principle; Asymptotic variance-covariance matrix; ECM algorithm; Newton-Raphson method; Asymptotic confidence interval

Mathematics Subject Classification: 62F10

1. Introdution

The discrete Weibull distribution is used to describe mathematically the life of a device, a material, or a structure measured in the numbers of cycles, blows, shocks, or revolutions to failure; see [1–3] for a dicussion in detail. Truncation and censoring occur quite commonly while observing lifetime data; the literature provide detailed accounts in this regard by [4–13]. As In this paper, we are concerned with lifetime data that are left truncated and right censored. Here, we describe in detail the steps of the Expectation Maximization (EM) algorithm for fitting the discrete Weibull distribution to left truncated and right censored data; see [14] for a comprehensive discussion on this topic. For comparative purposes, we also fit the discrete Weibull distribution by the Newton-Raphson (NR) method; the two methods give extremely close results under this setup.

The rest of this paper is organized as follows. In Section 2, we present a description of the form of the data and the corresponding likelihood function. In Section 3, the EM algorithm for the maximum likelihood estimation of the model parameters is described. As the expected complete log-likelihood is a complicated non-linear function of the parameters involved, the maximization part of the EM algorithm is done using the ECM algorithm (see [15]). In this section, we also derive the asymptotic variance-covariance matrix of the maximum likelihood estimates (MLEs) within the EM framework by using the missing information principle of [16]. Then, we use it to construct corresponding asymptotic confidence intervals for the model parameters. The asymptotic confidence intervals, making use of the observed information matrix, are also constructed. Finally, we summarize and conclude the paper in Section 4.

2. Form of data and the likelihood

Let X be the lifetime variable, which follows a discrete Weibull distribution with scale parameter q and shape parameter α. The probability function (pf) of X is given by (see [1])

$$f(k;q,\alpha) = P(X=k) = q^{(k-1)^\alpha} - q^{k^\alpha}, k = 1,2,\cdots, 0<q<1, \alpha>0.$$

Then, the cumulative distribution function (cdf) of X is $F(k;q,\alpha) = P(X \leqslant k) = 1 - q^{k^\alpha}$. We assume that Y and T are discrete random variables. Let Y denote the censoring time variable with pf $g(k)$ and cdf $G(k)$. Let T denote the truncated time variable with pf $h(k)$ and cdf $H(k)$. Suppose that $g(k), G(k), h(k)$ and $H(k)$ do not depend on $\boldsymbol{\theta} = (q,\alpha)', X, Y$ and T are independent. In the random left truncated and right censored (LTRC) model one observes (Z_i, T_i, δ_i) if $Z_i \geqslant T_i$, where $Z_i = \min(X_i, Y_i)$ and $\delta_i = I(X_i \leqslant Y_i), i = 1,2,\cdots,n$. When $Z_i < T_i$, nothing is observed. For convenience, let $\nu_i = I(Z_i \geqslant T_i)$.

The likelihood function for the left truncated and right censored data is given by

$$\begin{aligned} L(q,\alpha) &= \prod_{i=1}^{n}\{[f(z_i;q,\alpha)\bar{G}(z_i-1)h(t_i)]^{\delta_i\nu_i}[g(z_i)\bar{F}(z_i;q,\alpha)h(t_i)]^{(1-\delta_i)\nu_i}u(q,\alpha)^{1-\nu_i}\} \\ &= A\prod_{i=1}^{n}\{[f(z_i;q,\alpha)]^{\delta_i\nu_i}[\bar{F}(z_i;q,\alpha)]^{(1-\delta_i)\nu_i}[u(q,\alpha)]^{1-\sum_{i=1}^{n}\nu_i} \\ &= A\{\prod_{i=1}^{n}[q^{(z_i-1)^\alpha - z_i^\alpha}-1]^{\delta_i\nu_i}\}q^{\sum_{i=1}^{n}\nu_i z_i^\alpha}[u(q,\alpha)]^{n_0}, \end{aligned}$$

where

$$A=\prod_{i=1}^{n}\{[\bar{G}(z_i-1)h(t_i)]^{\delta_i\nu_i}[g(z_i)h(t_i)]^{(1-\delta_i)\nu_i}\},$$

$$u(q,\alpha)=P(Z_i<T_i)=1-\sum_{k=1}^{\infty}h(k)\bar{G}(k-1)\bar{F}(k-1;q,\alpha),$$

$$n_0=n-\sum_{i=1}^{n}\nu_i,$$

and A does not depend on $\theta=(q,\alpha)'$.

The log-likelihood function, after some simplification, becomes

$$\log L=\log A+\sum_{i=1}^{n}\delta_i\nu_i\log[q^{(z_i-1)^\alpha-z_i^\alpha}-1]+\log\sum_{i=1}^{n}\nu_i z_i^\alpha+n_0\log u(q,\alpha).$$

3. Methods of estimation

3.1. The EM algorithm

The EM algorithm is a very powerful and useful tool for analyzing incomplete data; see [14] for an elaborate discussion on the method. The algorithm consists of two steps-the Expectation step (E-step) and the Maximization step (M-step). In the E-step,the conditional expectation of the complete data likelihood is obtained, given the observed incomplete data and the current value of the parameter, real or vector valued. This expected likelihood is essentially a function of the parameter involved, and the current value of the parameter under which the expectation has been calculated. In the M-step,this expected

complete data likelihood is then maximized with respect to the parameter. The E- and M-steps are then iterated till convergence. This algorithm is known to have some desirable and advantageous properties over the direct methods for obtaining the MLEs in the case of incomplete data; see the above-mentioned reference for details.

First, when nothing is observed, i.e., $Z_i < T_i$, we add the data (W_i, δ_i'), where $W_i = \min(X_i, Y_i)$ and $\delta_i' = I(X_i \leqslant Y_i)$. Then we construct the complete data log-likelihood function. With the parameter vector denoted by $\theta = (q, \alpha)'$, had there been no truncation, the complete data likelihood would be

$$\begin{aligned} L(w, \delta_i'; \theta) = & \prod_{i=1}^{n} \{[f(z_i;\theta)\bar{G}(z_i-1)h(t_i)]^{\delta_i\nu_i}[g(z_i)\bar{F}(z_i;\theta)h(t_i)]^{(1-\delta_i)\nu_i} \\ & \times [f(w_i;\theta)\bar{G}(w_i-1)\bar{H}(w_i)]^{\delta_i'(1-\nu_i)}[g(w_i)\bar{F}(w_i;\theta)\bar{H}(w_i)]^{(1-\delta_i')(1-\nu_i)}\} \\ \propto & q^{\sum_{i=1}^{n}\nu_i z_i^{\alpha}+(1-\nu_i)w_i^{\alpha}} \prod_{i=1}^{n} \{[q^{(z_i-1)^{\alpha}-z_i^{\alpha}}-1]^{\nu_i\delta_i}[q^{(w_i-1)^{\alpha}-w_i^{\alpha}}-1]^{(1-\nu_i)\delta_i'}\}. \end{aligned}$$

Correspondingly, the complete data log-likelihood function is given by

$$\begin{aligned} \log L(w, \delta_i'; \theta) = & \log q \sum_{i=1}^{n} \nu_i z_i^{\alpha} + \sum_{i=1}^{n} \delta_i\nu_i \log[q^{(z_i-1)^{\alpha}-z_i^{\alpha}}-1] + \log q \sum_{i=1}^{n}(1-\nu_i)w_i^{\alpha} \\ & + \sum_{i=1}^{n} \delta_i'(1-\nu_i)\log[q^{(w_i-1)^{\alpha}-w_i^{\alpha}}-1]. \end{aligned}$$

The E-Step: In the E-step, we calculate the conditional expectation of the complete data log-likelihood, i.e., we calculate

$$Q(\theta, \theta^{(m)}) = E_{\theta^{(m)}}[\log L(w, \delta_i'; \theta)|z, \delta, \nu].$$

Clearly, the expectations of interest are

$$E_{\theta^{(m)}}(W_i^{\alpha}|W_i < T_i), \quad E_{\theta^{(m)}}\{\delta_i' \log[q^{(W_i-1)^{\alpha}-W_i^{\alpha}}]|W_i < T_i\}.$$

To derive these conditional expectations, we first consider the conditional

probability function of (W_i, δ_i'), given $W_i < T_i$, given by

$$P(W_i = k, \delta_i' = 1) = P(X_i = k, Y_i \geqslant k | W_i < T_i) = \frac{[q^{(k-1)^\alpha} - q^{k^\alpha}]\bar{G}(k-1)\bar{H}(k)}{u(q,\alpha)}$$
$$\triangleq \psi_1(k,\theta),$$
$$P(W_i = k, \delta_i' = 0) = P(Y_i = k, X_i > k | W_i < T_i) = \frac{g(k)q^{k^\alpha}\bar{H}(k)}{u(q,\alpha)} \triangleq \psi_2(k,\theta).$$

Hence

$$P(W_i = k, \delta_i' = l) = l\psi_1(k,\theta) + (1-l)\psi_2(k,\theta) \triangleq \varphi(l,k,\theta), l = 0,1; k = 1,2,\cdots.$$

So the conditional probability function of W_i, given $W_i < T_i$, given by

$$P(W_i = k) = \psi_1(k,\theta) + \psi_2(k,\theta) \triangleq \psi(k,\theta).$$

Based on the above conditional density functions, the required expectations can be derived to be

$$E_{\theta^{(m)}}\{\delta_i' \log[q^{(W_i-1)^\alpha - W_i^\alpha} - 1] | W_i < T_i\} = \sum_{k=1}^{\infty} \log[q^{(k-1)^\alpha - k^\alpha} - 1]\psi_1(k,\theta^{(m)})$$
$$\triangleq e_1(\theta,\theta^{(m)}),$$

$$E_{\theta^{(m)}}[W_i^\alpha | W_i < T_i] = \sum_{k=1}^{\infty} k^\alpha \psi(k,\theta^{(m)}) \triangleq e_1(\theta,\theta^{(m)}).$$

We obtain

$$Q(\theta,\theta^{(m)}) = \log q \sum_{i=1}^{n} \nu_i z_i^\alpha + \sum_{i=1}^{n} \delta_i \nu_i \log[q^{(z_i-1)^\alpha - z_i^\alpha} - 1]$$
$$+ n_0[e_1(\theta,\theta^{(m)}) \log q + e_2(\theta,\theta^{(m)})].$$

The quantity $Q(\theta,\theta^{(m)})$ needs to be maximized with respect to θ. Evidently,the maximization poses a challenge as the function involved is a complicated non-linear function of q and α,and the process used for this purpose is described

next.

The M-Step: In the maximization step, the quantity $Q(\theta, \theta^{(m)})$ is maximized with respect to θ over the parameter space Θ to obtain the improved estimate of the paramater as

$$\theta^{(m+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\theta, \theta^{(m)}).$$

The E-step and the M-step are then continued iteratively until convergence to obtain the MLE of the parameter θ. From the complicated form of the function $Q(\theta, \theta^{(m)})$, it is obvious that there are no explicit MLEs for the parameters, and one has to depend on a numerical maximization procedure. Here, we make use of the ECM algorithm (see [15]). In this algorithm, the function $Q(\theta, \theta^{(m)})$ by k-step maximizing conditional expectation, where k is the length of θ, to get the updated estimate $\theta^{(m+1)}$. This algorithm is a special case of the generalized EM algorithm (see [17]), and is closely related to the original EM algorithm. The properties of this algorithm are quite close to that of the EM algorithm. However, the ECM algorithm may converge slower than the EM algorithm.

In the M-step, the ECM algorithm is carried out as follows:

After getting $Q(q, \alpha|q^{(m)}, \alpha^{(m)})$, on the $(m+1)$th iteration of this ECM algorithm, calulate $q^{(m+1)}$ as the value of q that maximizes $Q(q, \alpha^{(m)}|q^{(m)}, \alpha^{(m)})$. Then calulate $\alpha^{(m+1)}$ as the value of α that maximizes $Q(q^{(m+1)}, \alpha|q^{(m+1)}, \alpha^{(m)})$. The updated estimate $\theta^{(m+1)} = (q^{(m+1)}, \alpha^{(m+1)})'$ are got. The M-step is then continued iteratively until convergence. It can be seen readily that the expression of $Q(q, \alpha|q^{(m)}, \alpha^{(m)})$ contains some summations that require special techniques for their evaluation. In our study, we calculated the summations by the sum function of the R software. It has been observed in our extensive empirical study that the numerical MLEs converge to the true parameter values quite accurately.

3.2. Newton-Raphson method

The NR method is a direct approach for obtaining the MLEs by maximizing the likelihood function. It involves calculation of the first and second derivatives of the observed log-likelihood with respect to the parameters. Herein, we use the NR method for comparative purpose. The NR method, although works well in general, fails to converge in some cases under this setup. In our study, we employed the NR method by a default function of the R software, called the maxNR function. We observed in our empirical study that the EM and the NR methods yield close results in most cases.

3.3. Asymptotic variances and covariance of the MLEs

Unlike the NR method, in the EM algorithm,the asymptotic variances and covariance of the MLEs are not directly obtained as a byproduct of the algorithm. Within the EM framework,the missing information principle (see [16]) can be applied to obtain the observed information matrix as

$$\text{Observed information} = \text{Complete information} - \text{Missing information}.$$

Then, inverting the observed information matrix, one can obtain the asymptotic variance-covariance matrix of the MLEs.

Let $I(\theta)$, $I_1(\theta)$ and $I_2(\theta)$ denote the complete information matrix, observed information matrix and the missing information matrix, respectively. The complete information matrix is given by

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log L(W,\delta';\theta)\right]. \tag{1}$$

The expected missing information can be easily obtained as

$$I_2(\theta) = -n_0 E\left[\frac{\partial^2}{\partial\theta^2}\log\varphi(l,k,\theta)\right]. \tag{2}$$

Thus, by the missing information principle, the observed information matrix can be obtained

$$I_1(\theta) = I(\theta) - I_2(\theta). \tag{3}$$

The asymptotic variance-covariance matrix of the MLE of θ can be obtained finally by inverting the observed information matrix $I_1(\theta)$ in Eq. (3) and evaluating at $\theta = \hat{\theta}$.

Define

$$\varphi_1(z, \delta; \theta) = \log q \sum_{i=1}^{n} \nu_i z_i^{\alpha} + \sum_{i=1}^{n} \delta_i \nu_i \log[q^{(z_i-1)^{\alpha} - z_i^{\alpha}} - 1]$$

and

$$\varphi_2(w_i, \delta_i'; \theta) = w_i^{\alpha} \log q + \delta_i' \log[q^{(w_i-1)^{\alpha} - w_i^{\alpha}} - 1].$$

The elements of the complete information matrix $I(\theta)$ are given by

$$E\left[\frac{\partial^2}{\partial q^2} \log L(W, \delta'; \theta)\right] = -\left[\frac{\partial^2}{\partial q^2}\varphi_1(z, \delta; \theta) + n_0 \sum_{k=1}^{\infty}\sum_{l=0}^{1} \psi(k, l; \theta) \frac{\partial^2}{\partial q^2}\varphi_2(k, l; \theta)\right],$$

$$-E\left[\frac{\partial^2}{\partial \alpha^2} \log L(W, \delta'; \theta)\right] = -\left[\frac{\partial^2}{\partial \alpha^2}\varphi_1(z, \delta; \theta) + n_0 \sum_{k=1}^{\infty}\sum_{l=0}^{1} \psi(k, l; \theta) \frac{\partial^2}{\partial \alpha^2}\varphi_2(k, l; \theta)\right],$$

$$-E\left[\frac{\partial^2}{\partial q \partial \alpha} \log L(W, \delta'; \theta)\right] = -\left[\frac{\partial^2}{\partial q \partial \alpha}\varphi_1(z, \delta; \theta) + n_0 \sum_{k=1}^{\infty}\sum_{l=0}^{1} \psi(k, l; \theta) \frac{\partial^2}{\partial q \partial \alpha}\varphi_2(k, l; \theta)\right].$$

The elements of the missing information matrix $I_2(\theta)$ are given by

$$-n_0 E\left[\frac{\partial^2}{\partial q^2} \log \psi(W, \delta'; \theta)\right] = -n_0 \sum_{k=1}^{\infty}\sum_{l=0}^{1} \psi(k, l; \theta) \frac{\partial^2}{\partial q^2} \log \psi(k, l; \theta),$$

$$-n_0 E\left[\frac{\partial^2}{\partial \alpha^2} \log \psi(W, \delta'; \theta)\right] = -n_0 \sum_{k=1}^{\infty}\sum_{l=0}^{1} \psi(k, l; \theta) \frac{\partial^2}{\partial \alpha^2} \log \psi(k, l; \theta),$$

$$-n_0 E\left[\frac{\partial^2}{\partial q \partial \alpha} \log \psi(W, \delta'; \theta)\right] = -n_0 \sum_{k=1}^{\infty}\sum_{l=0}^{1} \psi(k, l; \theta) \frac{\partial^2}{\partial q \partial \alpha} \log \psi(k, l; \theta).$$

The observed Fisher information matrix can be obtained from Eq. (3). Finally, by inverting $I_1(\theta)$, the asymptotic variance-covariance matrix of the MLEs can be obtained.

3.4. Confidence intervals

After obtaining the MLEs and their asymptotic variances, the asymptotic confidence intervals for q and α can be constructed using the asymptotic normality of the MLEs. Here, we study the asymptotic confidence intervals for q and α corresponding to both the EM algorithm and the NR method. Evidently, the confidence intervals corresponding to the EM algorithm and the NR method will be different, due to the different estimates of the parameters and the standard errors obtained from these methods. Then, these confidence intervals are compared in terms of their coverage probabilities through a Monte Carlo simulation study.

One can also construct parametric bootstrap confidence intervals for q and α in the following way. First of all, based on a given data of size n, the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = (q, \alpha)'$ is obtained. Then, using $\hat{\boldsymbol{\theta}}$ as the true value of the parameter, a sample of size n in the same sampling framework with left truncation and right censoring is produced. This process is repeated for 1000 Monte Carlo simulation runs, and the MLEs are obtained for each of these samples. Then, based on these 1000 estimates, the bootstrap bias and variance for the estimates of q and α are obtained. In the final step, a $100(1-\beta)$ parametric bootstrap confidence interval for q is obtained as

$$\text{LCL: } \hat{q} - b_q - z_{\beta/2}\sqrt{\sigma_q}, \quad \text{UCL:} \hat{q} - b_q + z_{\beta/2}\sqrt{\sigma_q},$$

where b_q and σ_q are the bootstrap bias and variance for the estimate of q, respectively, and z_β is the upper β-percentage point of the standard normal distribution. The confidence interval for α can be constructed in a similar manner. The parametric bootstrap confidence intervals can then be compared

to the other asymptotic confidence intervals in terms of coverage probabilities.

4. Conclusions

In this paper, we consider the likelihood inference for discrete Weibull data with left truncation and right censoring. The EM algorithm is applied to estimate the model parameters, and the maximization part of the EM algorithm is carried out using the ECM algorithm. The asymptotic variance-covariance matrix of the MLEs under the EM framework is obtained through the missing information principle, and asymptotic confidence intervals for the parameters are constructed.

Reference

[1] T. Nakagawa, S. Osaki, The discrete weibull distribution, Reliability, IEEE Transactions on 24 (5) (1975) 300–301.

[2] M. A. Khan, A. Khalique, A. Abouammoh, On estimating parameters in a discrete weibull distribution, Reliability, IEEE Transactions on 38 (3) (1989) 348–350.

[3] K. Kulasekera, Approximate mle's of the parameters of a discrete weibull distribution with type i censored data, Microelectronics Reliability 34 (7) (1994) 1185–1188.

[4] A. C. Cohen, Truncated and censored samples, Theory and Applications.

[5] N. Balakrishnan, A. C. Cohen, Order statistics and inference: Estimation methods (1991).

[6] W. Q. Meeker, L. A. Escobar, Statistical methods for reliability data, Vol. 78, Wiley New York, 1998.

[7] Y.-T. Hwang, C.-c. Wang, A goodness of fit test for left-truncated and right-censored data, Statistics & Probability Letters 78 (15) (2008) 2420–2425.

[8] M. Jácome, M. Iglesias-Pérez, Presmoothed estimation with left-truncated and right-censored data, Communications in Statistics—Theory and Methods 37 (18) (2008) 2964–2983.

[9] Y. Hong, W. Q. Meeker, J. D. McCalley, Prediction of remaining life of power transformers based on left truncated and right censored lifetime data, The Annals of Applied Statistics 3 (2) (2009) 857–879.

[10] P.-s. Shen, Hazards regression for length-biased and right-censored data, Statistics & Probability Letters 79 (4) (2009) 457–465.

[11] P.-s. Shen, Estimation in the cox proportional hazards model with doubly censored and truncated data, Journal of Statistical Computation and Simulation 81 (11) (2011) 1717–1725.

[12] N. Balakrishnan, D. Mitra, Likelihood inference for lognormal data with left truncation and right censoring with an illustration, Journal of Statistical Planning and Inference 141 (11) (2011) 3536–3553.

[13] N. Balakrishnan, D. Mitra, Left truncated and right censored weibull data and likelihood inference with an illustration, Computational Statistics & Data Analysis 56 (12) (2012) 4011–4025.

[14] G. J. McLachlan, T. Krishnan, The EM algorithm and extensions, Vol. 382, John Wiley & Sons, 2008.

[15] X.-L. Meng, D. B. Rubin, Maximum likelihood estimation via the ecm algorithm: A general framework, Biometrika 80 (2) (1993) 267–278.

[16] T. A. Louis, Finding the observed information matrix when using the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1982) 226–233.

[17] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38.