Estimadores, pontual e intervalar, para dados com censuras intervalar

Débora Ohara, Estela Maris Pereira Bereta, Teresa Cristina Martins Dias

Resumo

Dados com censura intervalar ocorrem com frequência em estudos de diversas áreas, em situações em que o evento de interesse é observado com periodicidade. Neste caso, o tempo exato da ocorrência não é conhecido (observado), porém sabe-se que o evento ocorreu dentro de um intervalo (conhecido) de tempo. Este tipo de observação é tratada em análise de sobrevivência usando técnicas apropriadas que considera a presença de censuras intervalar. Considerando este tipo de censura, neste trabalho, apresentamos resultados de estimação pontual e intervalar dos parâmetros de interesse, via algoritmo EM sob a abordagem clássica. Sob a abordagem Bayesiana, obtivemos estimativas pontuais e, para as intervalares, apresentamos uma alternativa ao método proposto por Pradhan e Kundu (2014). Dados reais e simulados foram utilizados para ilustrar a teoria estudada.

Introdução

Tempos até a ocorrência de um evento de interesse são objetos de estudo na área de sobrevivência/confiabilidade. Porém, nem sempre o tempo exato da ocorrência do evento é conhecido e estes são, geralmente, denominados censuras. Neste caso, temos informações parciais a respeito do tempo de ocorrência.

Existem vários tipos de censuras, como por exemplo, à direita, à esquerda e intervalar. A censura é umas das características dos tempos na área de sobrevivência. Outra característica de dados de sobrevivência é o agrupamento, o que ocorre quando as unidades de estudo são avaliadas nos mesmos tempos, sendo este um caso de censura intervalar.

De acordo com o tipo considerado no planejamento do estudo, definimos os métodos de estimação dos parâmetros e/ou para a função dos parâmetros, como por exemplo, a função de sobrevivência/confiabilidade, sob as abordagens clássica, Bayesiana e não-paramétrica, dentre outras. Como exemplos de métodos não-paramétricos, citamos os estimadores de Kaplan e Meier (1958), Turnbull (1976), e de Nelson-Aalen (Nelson (1972) e Aalen (1978)); no caso semi-paramétrico citamos os modelos de Cox

(1972) e na abordagem clássica, os modelos de regressão e os modelos paramétricos, sendo que os mais usados são exponencial, Weibull, log-normal e gama. Porém, estes estimadores não são indicados para o caso de censura intervalar, objeto de estudo neste trabalho. Strapasson (2007) cita alguns autores que propuseram fixar um tempo dentro do intervalo no qual ocorreu o evento e aplicar os métodos usuais de estimação em análise de sobrevivência.

Pradhan e Kundu (2014) apresentam vários métodos de estimação pontual (algoritmo EM, aproximação de Lindley e *importance sampling*) no caso de tempos censurados de forma intervalar, com distribuição exponencial e Weibull. Também, os autores apresentam um algoritmo para a construção de intervalos de confiança, na abordagem Bayesiana.

Neste trabalho, consideramos estudos em que a resposta é o tempo até a ocorrência de um evento de interesse. Por exemplo, pacientes que foram observados durante um período de tempo (especificado) e o tempo até a morte deste paciente foi registrado. Neste caso a morte é o evento de interesse e o tempo é o objeto de estudo. Para uma amostra de n pacientes e os respectivos registros dos tempos observados ou intervalo de censura, apresentamos estimadores pontuais e intervalares, na abordagem clássica, para os parâmetros do modelo Weibull e função de sobrevivência, aplicando este estudo em um conjunto de dados.

Material e metodologia

Suponha um conjunto de dados que representam tempos até a ocorrência do evento de interesse, para n observações. Os dados são representados pelo par $(T_i, \delta_i), i = 1, \ldots, n$, em que T_i representa o tempo (variável aleatória contínua e não-negativa) de ocorrência do evento da i-ésima observação e δ_i é a variável indicadora de censura. Se o tempo até a ocorrência do evento é observado, T_i não pertence ao intervalo $[L_i, R_i]$, em que L_i é o limite inferior e R_i é o limite superior do intervalo de observação. Desta forma, a observação não é censurada e portanto $\delta_i = 1$. Para observações em que o tempo de vida ocorre dentro do intervalo $[L_i, R_i]$, e portanto, o tempo exato é desconhecido, temos censura intervalar e então $\delta_i = 0$.

Consideramos n unidades observacionais sendo que os tempos registrados são denotados por $(T_1, \ldots, T_{n_1})$ e tempos censurados são denotados por $([L_{n_1+1}, R_{n_1+1}], \ldots, [L_{n_1+n_2}, R_{n_1+n_2}])$, para $n = n_1+n_2$. Também, que os tempos seguem uma distribuição Weibull, ou seja, $T \sim \text{Wei}(\alpha, \lambda)$, dada por:

$$f_T(t; \alpha, \lambda) = \alpha\lambda t^{\alpha-1} \exp\{-\lambda t^\alpha\}, \;\; t > 0, \; \alpha > 0 \text{ e } \lambda > 0.$$

Usando as informações de tempos com censura intervalar, a função de verossimilhança é rescrita da seguinte forma:

$$L(\alpha, \lambda | dados) = \ c\alpha^{n_1}\lambda^{n_1} \prod_{i=1}^{n_1} t_i^{\alpha-1} \ e^{-\lambda \sum_{i=1}^{n_1} t_i^{\alpha}} \prod_{i=n_1+1}^{n_1+n_2} (e^{-\lambda t_i^{\alpha}} - e^{-\lambda r_i^{\alpha}}). \qquad (1)$$

Observe que temos um problema de dados incompletos (censuras) e como alternativa aos métodos usuais de estimação, usamos o algoritmo EM. Este é uma ferramenta computacional que calcula o estimador de máxima verossimilhança de forma iterativa. Para aplicar este método, construímos um conjunto de dados completos, formado pelos dados registrados aumentado com os faltantes. Desta forma, obtemos a função verossimilhança associada aos dados completos, chamada de "pseudo-verossimilhança".

O princípio do algoritmo consta de uma sequência de maximizações, através de dois passos. O primeiro é o passo "E" (esperança), em que o valor esperado do logaritmo da "pseudo-verossimilhança" é calculado. O segundo é o passo "M" (maximização), em que a "pseudo-verossimilhança" formada no passo anterior é maximizada (como apresentado em Dempster *et al.* (1977) e Wu (1983)).

O algoritmo, no caso de dados com distribuição Weibull e censura intervalar, pode ser descrito da seguinte forma:

- Passo E: denote as observações censuradas por Z_i, sendo que $Z_i \sim Weibull(\alpha; \lambda)$, $i = n_1 + 1, \ldots, n_1 + n_2$. A função de verossimilhança dos dados completos é dada por:

$$L_c(\alpha, \lambda) = \alpha^n \lambda^n \prod_{i=1}^{n_1} t_i^{\alpha-1} \ e^{-\lambda \sum_{i=1}^{n_1} t_i^{\alpha}} \prod_{i=n_1+1}^{n_1+n_2} z_i^{\alpha-1} \ e^{-\lambda \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}}. \qquad (2)$$

  Para $i = n_1 + 1, \ldots, n_1 + n_2$, os valores de Z_i encontrados são tais que:

  O termo $(e^{-\lambda t_i^{\alpha}} - e^{-\lambda r_i^{\alpha}})$ é substituído por valores de Z, que são encontrados tais que:

$$Z_i = E(T | L_i < T < \ R_i) = \frac{\int_{L_i}^{R_i} \alpha\lambda x^{\alpha} \ e^{-\lambda x^{\alpha}} dx}{e^{-\lambda L_i^{\alpha}} - e^{-\lambda R_i^{\alpha}}}.$$

Desta o log da função de verossimilhança é dado por:

$$\begin{aligned} l_c(\alpha,\lambda) &= n\ln\alpha + n\ln\lambda + (\alpha-1)\left(\sum_{i=1}^{n_1}\ln t_i + \sum_{i=n_1+1}^{n_1+n_2}\ln z_i\right) \\ &\quad -\lambda\left(\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}\right). \end{aligned} \tag{3}$$

Resumindo, o passo "E " é caracterizado pelo cálculo da esperança da distribuição condicional $(T|L_i < T < R_i)$.

- Passo M: use os valores obtidos no passo "E" para maximizar o log da função de verossimilhança (equação 3), com respeito a α e λ.

  De forma iterativa obtemos os estimadores:

$$\lambda^{(k+1)}(\alpha) = \frac{n}{\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha,\lambda^{(k)})},$$

  e

$$\begin{aligned} \alpha^{(k+1)}(\lambda) &= n\left(\frac{\sum_{i=1}^{n_1} t_i^{\alpha}\ln t + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha^{(k)},\lambda^{(k)})\ln z_i(\alpha^{(k)},\lambda^{(k)})}{\sum_{i=1}^{n_1} t_i^{\alpha} + \sum_{i=n_1+1}^{n_1+n_2} z_i^{\alpha}(\alpha^{(k)},\lambda^{(k)})}\right)^{-1} \\ &\quad -n\left(\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2}\ln z_i(\alpha^{(k)},\lambda^{(k)}\right). \end{aligned}$$

Este é um processo iterativo e portanto deve ser realizado até se atingir a convergência. Aqui, adotamos o seguinte critério de parada:

$$\left\|\lambda^{(k+1)} - \lambda^{(k)}\right\| < \epsilon$$

e

$$\left\|\alpha^{(k+1)} - \alpha^{(k)}\right\| < \epsilon$$

em que ϵ é um valor pré-fixado maior que zero.

Resultados

Para ilustrar a teoria, simulamos os tempos e intervalos de censura com os mesmo valores de parâmetros de Pradhan e Kundu (2014), mostrados abaixo.

0,882	1,1739	0,4123	0,4565	1,9935	1,0662	1,3516	0,313
1,3364	1,6493	0,3	0,8187	0,0253	0,6841	0,2672	1,1791
0,346	0,8371	0,9184	0,8331	0,5123	0,1045	0,2159	0,0992

e os intervalos são: $[0,7286; 2,7756][0,4465; 1,7119][0,0204; 2,7927]$ $[0,6566; 1,9712]$ $[1,5674; 2,4757][0,1700; 2,3342]$.

No processo iterativo aplicando o algoritmo EM obtemos as estimativas pontuais $\hat{\alpha} = 1,9227$ e $\hat{\lambda} = 1,1037$ e os intervalos com 95% de confiança para α e λ, respectivamente, são $(1,8851; 1,9603)$ e $(1,0671; 1,1402)$. Os autores encontraram $\hat{\alpha} = 1,4945$ e $\hat{\lambda} = 1,1864$.

Proposta

As estimativas pontuais e intervalares no caso clássico foram apresentadas e implementadas no *software* R, para o caso da distribuição Weibull. Sob o enfoque Bayesiano, as estimativas pontuais foram obtidas. O foco deste trabalho é obter estimativas intervalares, utilizando um método alternativo ao método apresentado por Pradhan e Kundu (2014), que estão sendo estudadas e implementadas. No trabalho final, apresentaremos os resultados das estimações pontuais e intervalares, no enfoque Bayesiano.

Referências

Aalen, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

Cox, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, B*, v. 34(2), p. 187–220, 1972.

Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (methodological)*, pages 1–38, 1977.

Kaplan, E. L.; Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53(282), p. 457–481, 1958.

Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, v. 14(4), p. 945–966, 1972.

Pradhan, B.; Kundu, D. Analysis of interval-censored data with Weibull lifetime distribution. *Sankhya B*, v. 76(1), p. 120–139, 2014.

Strapasson, E. *Comparação de modelos com censura intervalar em análise de sobrevivência*. Tese (Doutorado), Universidade de São Paulo, 2007.

Turnbull, B. W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

Wu, C. J. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.