Estimativas da Função de Probabilidade de Cura do Modelo de Mistura Padrão com Censura Informativa

Josenildo de Souza Chaves,[1]

Josemar Rodrigues[2]

Resumo

Neste trabalho, consideramos que uma fração p_0 da população são de indivíduos curados ou imunes e a fração restante $1-p_0$ são de não curados. Assumimos a distribuição exponencial para o tempo de sobrevivência e a uniforme-exponencial para o tempo de censura. O impacto causado nas estimativas dos parâmetros do modelo de mistura padrão pela censura informativa uniforme-exponencial é analisado em um estudo com dados de leucemia. Estimativas de máxima verossimilhança da função de probabilidade de cura e intervalos de confiança assintóticos dos parâmetros de interesse são calculadas com o uso da matriz de informação de Fisher e com o uso da matriz de informação observada. Estimativas não-paramétricas da função de probabilidade de cura são também calculadas.

Palavras-chave: análise de sobrevivência, modelo com taxa de cura, censura informativa, distribuição uniforme-exponencial, Matriz de informação de Fisher.

1 Introdução

Os modelos com fração de cura foram desenvolvidos originalmente para analisar dados de sobrevivência de indivíduos portadores de câncer por Boag (1949) e Berkson & Gage (1952), que propuseram um modelo de mistura em que há uma proporção de indivíduos curados ou imunes na população. Iremos nos referir ao modelo de Berkson & Gage (1952) como modelo de mistura padrão. Chen *et al.* (1999) e Tsodikov *et al.* (2003) sugerem modelos alternativos a esse, baseados no número de células suscetíveis a desenvolver metástase. Rodrigues *et al.* (2009) propõem uma teoria unificada da qual o modelo de mistura padrão, que será definido na Subseção 1.2, e o modelo de Chen *et al.* (1999) são casos particulares.

1.1 Modelo Uniforme-Exponencial

Neste trabalho, utilizamos o modelo uniforme-exponencial que pode ser encontrado em Patterson & Smith (1985) e Ghitany (1993) para o tempo de censura X. Suponha que há interesse em modelar uma situação que pode ser representada por uma função de risco híbrida das distribuições uniforme e exponencial. Definimos, como uma função explícita do tempo, uma distribuição uniforme-exponencial para uma v.a. X, denotada por $X \sim UExp(\lambda_c, T_0)$, como segue.

[1]Departamento de Matemática, Universidade Federal do Maranhão, Av. dos Portugueses, s/n, Campus do Bacanga, 65.085-580, São Luís-MA, Brasil. e-mail: chavesjs@yahoo.com.br

[2]Departamento de Estatística, Universidade Federal de São Carlos, Caixa Postal 676, 13565-905, São Carlos-SP, Brasil. e-mail: vjosemar@power.ufscar.br

Definição 1.1 Uma variável aleatória X tem distribuição uniforme-exponencial (UExp) com parâmetros λ_c e T_0, o tempo de duração do estudo (conhecido), se sua função de distribuição de probabilidade, G, for dada por

$$G(x;\lambda_c) = 1 - \left(1 - \frac{x}{T_0}\right) e^{-\lambda_c x}, \tag{1}$$

em que $0 < x < T_0$ e $\lambda_c > 0$.

1.2 Modelo de Mistura Padrão

O modelo de mistura padrão completamente parametrizado é dado pela seguinte definição.

Definição 1.2 Seja T uma variável aleatória representando o tempo de ocorrência do evento de interesse e seja M uma variável aleatória Bernoulli representando o número de causas do evento de interesse, em que $M = 0$ se o indivíduo não está em risco (indivíduo curado ou imune) e $M = 1$ caso contrário. Seja $P(M = 0) = p_0 \in (0, 1)$ a proporção de indivíduos curados que nunca falharão. O modelo de mistura padrão é dado por

$$S_p(t;\boldsymbol{\theta}, p_0) = p_0 + (1 - p_0)S(t;\boldsymbol{\theta}|M = 1), \tag{2}$$

sendo que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ e p_0 pertencem ao espaço paramétrico $\boldsymbol{\Theta} = \mathbb{R}^k \times (0, 1)$ e $S(t;\boldsymbol{\theta}|M = 1)$ é uma função de sobrevivência própria associada aos indivíduos em risco da população.

2 Censura Informativa

Para caracterizar a censura informativa que iremos trabalhar, introduzimos, motivada por Lawless (1982), Kalbfleisch & Prentice (2002) e Siannis *et al.* (2005), a seguinte definição.

Definição 2.1 Sob o modelo (2) o mecanismo de censura é informativo se existir uma relação funcional entre os parâmetros da distribuição de probabilidade da v.a. T e os parâmetros da distribuição de probabilidade da v.a. X, o tempo de censura. Caso contrário, a censura é não-informativa.

2.1 Função de Verossimilhança e Matriz de Informação de Fisher

O conjunto de dados $\mathcal{D}$ que temos interesse em analisar consiste de n observações (Y_i, δ_i) das variáveis aleatórias (Y, δ). Seja T o tempo de vida dos indivíduos em risco e seja X o tempo de censura variáveis aleatórias independentes com funções de distribuição F e G, respectivamente. A variável aleatória Y_i é definida por

$$Y_i = \min\{T_i, X_i\},\ i = 1, \ldots, n, \tag{3}$$

sendo que $T_1, \ldots, T_n$ são variáveis aleatórias i.i.d. com função de distribuição F, independentes de $X_1, \ldots, X_n$, as quais são i.i.d. com função de distribuição G. A variável indicadora δ_i é tal que $\delta_i = 0$, se a i-ésima observação é censurada e $\delta_i = 1$, se a i-ésima observação está associada a uma falha. Note que δ_i, $i = 1, \ldots, n$, pode ser representada do seguinte do seguinte modo

$$\delta_i = \begin{cases} 1, & \text{se } T_i \leq X_i, \\ 0, & \text{se } T_i > X_i. \end{cases} \tag{4}$$

No modelo (2) existem dois mecanismos de censura atuando simultaneamente, um dado pela variável aleatória δ_i definida em (4) e outro pela variável aleatória

$$M_i = \begin{cases} 1, \text{ se o indivíduo } i \text{ está em risco,} \\ 0, \text{ se o indivíduo } i \text{ não está em risco (curado),} \end{cases} \tag{5}$$

em que, de acordo com a Definição 1.2, $M_i \sim Bernoulli(1 - p_0)$ i.i.d., $i = 1, \ldots, n$.

2.2 Relação Funcional Entre os Parâmetros das Distribuições de T e X

Uma configuração de dependência entre os parâmetros das distribuições de T, X, M e δ_i, $i = 1, \ldots, n$ segue do fato de que a probabilidade de censura do i-ésimo indivíduo, p_c, é dada por

$$\begin{aligned} p_c &= P\left(\delta_i = 0\right) \\ &= p_0 \overbrace{P\left(\delta_i = 0 | M_i = 0\right)}^{1} + (1 - p_0) \overbrace{P\left(\delta_i = 0 | M_i = 1\right)}^{\kappa^*}, \quad i = 1, \ldots, n. \end{aligned} \tag{6}$$

Logo,

$$p_c = p_0 + (1 - p_0)\kappa^*. \tag{7}$$

Nas equações (6) e (7), κ^* representa a probabilidade do i-ésimo indivíduo ser censurado dado que está em risco. Esta probabilidade determina uma relação funcional entre os parâmetros das distribuições de T e X caracterizada pelo seguinte teorema.

Teorema 2.1 Considere o modelo de mistura padrão dado em (2) e seja X uma v.a. representando o tempo de censura com função de distribuição $G(x; \boldsymbol{\psi})$ independente da v.a. T, o tempo de sobrevivência dos indivíduos em risco. Então, a probabilidade κ^* do i-ésimo indivíduo ser censurado dado que está em risco, é dada por

$$\kappa^* = \int_0^\infty S\left(x; \boldsymbol{\theta} | M_i = 1\right) g\left(x; \boldsymbol{\psi} | M_i = 1\right) dx, \;\; i = 1, 2, \ldots, n, \tag{8}$$

em que $g(\cdot) = G'(\cdot)$.

Prova. Temos que

$$\begin{aligned}
\kappa^* &= P(\delta_i = 0|M_i = 1) = P(T > X|M_i = 1) \\
&= \int\int_{\{(t,x):\, t>x, x>0\}} f(t, x; \boldsymbol{\theta}, \boldsymbol{\psi}|M_i = 1)\, dtdx, \\
&= \int\int_{\{(t,x):\, t>x, x>0\}} f(t; \boldsymbol{\theta}|M_i = 1)\, g(x; \boldsymbol{\psi}|M_i = 1)\, dtdx \\
&= \int_0^\infty \left\{ \int_x^\infty f(t; \boldsymbol{\theta}|M_i = 1)\, dt \right\} g(x; \boldsymbol{\psi}|M_i = 1)\, dx.
\end{aligned}$$

Portanto, κ^* é dada pela equação (8). ■

Suponha que a censura seja informativa, isto é, a função de distribuição G da variável aleatória X_i envolve $\boldsymbol{\theta}$ e p_0. A função de verossimilhança com censura informativa, motivada por Lagakos (1979), Lawless (1982) e Maller & Zhou (1996), é dada por

$$L^*(\boldsymbol{\theta}, p_0, \boldsymbol{\psi}; \mathcal{D}) = \prod_{i=1}^{n} f_p^{\delta_i}(y_i; \boldsymbol{\theta}, p_0) S_p^{1-\delta_i}(y_i; \boldsymbol{\theta}, p_0) \prod_{i=1}^{n} 1 - G(y_i; \boldsymbol{\psi})^{\delta_i} g^{1-\delta_i}(y_i; \boldsymbol{\psi}). \tag{9}$$

sendo que $g(\cdot) = G'(\cdot)$ e $\boldsymbol{\psi}$ é um vetor de parâmetros associado à distribuição de censura $G(\cdot)$.

3 Função de Probabilidade de Cura

Dado que um indivíduo tenha sido acompanhado até o tempo $t > 0$ e o evento falha não tenha sido observado, desejamos determinar a probabilidade de que este indivíduo esteja curado.

Baseada em Maller & Zhou (1996), definimos a função de probabilidade de cura do seguinte modo:

Definição 3.1 Dado que um indivíduo tenha sobrevivido até o tempo $t > 0$, a probabilidade de que este indivíduo esteja curado é determinada pela função de probabilidade de cura

$$p(t) = P(M = 0|T > t). \tag{10}$$

A função $p(t)$, é obtida pelo seguinte teorema.

Teorema 3.1 Dado que um indivíduo tenha sobrevivido até o tempo $t > 0$, a função de probabilidade de cura, é dada por

$$p(t) = \frac{p_0}{S_p(t)}. \tag{11}$$

Prova. A prova é imediata pela aplicação da fórmula de Bayes em (10), observando-se que $P(T > t|M = 0) = 1, \forall t$ e $P(M = 0) = p_0$. ■

4 Aplicações com Dados Reais e Conclusões

Analisamos dados de pacientes com leucemia descritos em (Kersey *et al.*, 1987). Determinamos estimativas de máxima verossimilhança (EMV) e intervalos de confiança dos parâmetros de interesse. Estimativas de Kaplan-Meier (vide por exemplo, Kaplan & Meier, 1958; Colosimo & Giolo, 2006) da função de sobrevivência e da função de probabilidade de cura são também calculadas.

Assumimos no modelo de mistura padrão (2) e função de verossimilhança (9) que os tempos de vida dos indivíduos em risco seguem uma distribuição exponencial com parâmetro λ e que os tempos de censura seguem (1). Os intervalos de confiança assintóticos são construídos com censura não-informativa com matriz de informação observada e com censura informativa com matriz de informação de Fisher. Para construir intervalos de confiança de κ^* utilizamos o método delta que pode ser encontrado em Migon & Gamerman (1999).

As estimativas de Kaplan-Meier (EKM) da função de probabilidade de cura $p(t)$ são obtidas, de acordo com Maller & Zhou (1996), do seguinte modo:

$$\hat{p}(t) = \frac{\hat{p}_n}{\hat{S}_n(t)}, \tag{12}$$

em que, $\hat{p}_n$ é dado pelo valor mínimo do EKM de $S_p(t)$ e $\hat{S}_n(t)$ é dado pelo EKM de $S_p(t)$.

Verificamos que as EMV de λ e da fração de cura (não sofrer recorrência da leucemia após o transplante alogênico) p_0 calculadas com o uso da função de verossimilhança com censura não-informativa são aproximadamente iguais às EMV calculadas com o uso da função de verossimilhança com censura informativa, não dependendo de T_0. Entretanto, as alterações no erro padrão das EMV de λ e p_0, podem provocar diferenças importantes nos comprimentos dos intervalos de confiança assintóticos. Notamos que o valor de κ^* pode ser considerado pequeno, por isso, os resultados obtidos com censura não-informativa e matriz de informação observada não podem ser descartados. Situações como esta, em que é possível considerar que o mecanismo de censura seja não-informativo, já estão configuradas em estudos de simulação (Chaves, 2010). Entretanto, é necessário esclarecer de acordo com (8) que a relação de dependência entre os parâmetros da distribuição de sobrevivência e distribuição de censura estabelecida por κ^* não permite admitir que o mecanismo de censura informativo seja ignorado.

Referências

Berkson, J. & Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**, 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society B*, **11**, 15–53.

Chaves, J. S. (2010). *Modelo de mistura padrão com censura informativa uniforme-exponencial*. Tese de doutorado, DEs – Universidade Federal de São Carlos, São Carlos, Brasil.

Chen, M.-H., Ibrahim, J. G. & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**, 909–919.

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Edgard Blücher, São Paulo.

Ghitany, M. E. (1993). On the information matrix of exponential mixture models with long-term survivors. *Biometrical Journal*, **35**, 15–27.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, second edition.

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

Kersey, J. H., Weisdorf, D., Nesbit, M. E., Lebien, T. W., Woods, W. G., Mcglave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B., Hurd, D. & Ramsay, N. K. C. (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, **317**, 461–467.

Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, **35**, 139–156.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Maller, R. A. & Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Migon, H. S. & Gamerman, D. (1999). *Statistical Inference: An Integrated Approach*. Arnold, London.

Patterson, B. H. & Smith, P. J. (1985). Asymptotic and finite sample behavior of the time on test estimator under random censorship when lifetimes are not exponential. *Communications in Statistics – Theory and Methods*, **14**, 1643–1658.

Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009). On the unification of the long-term survival models. *Statistics and Probability Letters*, **79**, 753–759.

Siannis, F., Copas, J. & Lu, G. (2005). Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, **6**, 77–91.

Tsodikov, A. D., Ibrahim, J. G. & Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078.