

ANÁLISE DE SOBREVIVÊNCIA COM ERROS DE CLASSIFICAÇÃO

Magda Carvalho Pires¹, Enrico Antonio Colosimo¹, Arlaine Aparecida da Silva¹

Resumo: *A análise de dados de sobrevivência consiste no estudo do tempo até a ocorrência de um evento de interesse como, por exemplo, a recidiva de uma doença. Quando o tempo exato de ocorrência não é conhecido, mas sabe-se que ele ocorreu no intervalo entre duas avaliações consecutivas do indivíduo, estamos diante de um estudo com censura intervalar. Entretanto, a detecção do evento depende da qualidade dos testes aplicados, pois estes podem estar sujeitos a erros de classificação: um indivíduo pode ser diagnosticado como doente quando na verdade ele está sadio ou vice versa. Nesses casos, ao utilizar métodos tradicionais de Análise de Sobrevivência, estimativas viciadas para os parâmetros da distribuição do tempo de falha são obtidas. Apresentamos, então, um modelo que incorpora a sensibilidade e a especificidade do teste ao modelo de análise de sobrevivência com dados agrupados (caso especial de censura intervalar em que todas as unidades são avaliadas nos mesmos instantes). Estudos de simulação Monte Carlo demonstraram que o método proposto é bastante eficiente, pois suas estimativas apresentam menor vício relativo do que aquelas fornecidas pelo método tradicional.*

Palavras-chave: Análise de Sobrevivência, Censura Intervalar, Testes Diagnósticos, Sensibilidade, Especificidade.

Introdução

Estudos de sobrevivência analisam o tempo até a ocorrência de um evento de interesse (tempo de falha T) podendo ser, por exemplo, o tempo até a morte, o tempo até a cura ou até a recidiva de uma doença. Os dados são caracterizados pela eventual presença de observação parcial da resposta (censura) que ocorre quando o acompanhamento do indivíduo foi interrompido por alguma razão. A censura à direita é o tipo mais comum e indica que o tempo de falha é superior ao tempo observado. Existem também os casos em que não se sabe o tempo exato de ocorrência do evento de interesse, mas sabe-se que ele ocorreu dentro de um intervalo (t_{j-1}, t_j) e, por isso, a censura é denominada censura intervalar.

A distribuição do tempo de falha T pode ser especificada pela sua função de sobrevivência $S(t) = P(T \geq t)$, que é definida como a probabilidade de uma observação não falhar até o instante de tempo t . Para estimarmos essa função, podemos utilizar métodos paramétricos e métodos não paramétricos (Colosimo e Giolo, 2006). Os estimadores paramétricos são construídos a partir de modelos probabilísticos, sendo os mais comuns os modelos Exponencial, Weibull e Lognormal. Depois de especificar a distribuição do tempo de falha é preciso estimar seus parâmetros, o que pode ser realizado pelo método de máxima verossimilhança.

No caso de dados agrupados (caso especial de censura intervalar em que todas as unidades são avaliadas nos mesmos k instantes), a função de verossimilhança é dada por:

$$L(\rho|n, m) = \prod_{j=1}^k [S(t_{j-1}; \rho) - S(t_j; \rho)]^{n_j} [S(t_j; \rho)]^{m_j} \quad (1)$$

¹Departamento de Estatística, ICEX, UFMG, magda@est.ufmg.br, enricoc@est.ufmg.br, arlaineas@yahoo.com.br

em que ρ é o parâmetro (ou vetor de parâmetros) da distribuição de T , $S(t_0) = 1$, n_j é o número de falhas e m_j o número de censuras em t_j .

Nos estudos de sobrevivência, muitas vezes estamos interessados em analisar o tempo até a ocorrência de uma doença, cuja detecção é realizada através de testes diagnósticos (como exames clínicos e laboratoriais) que estão sujeitos a erros de classificação. Os erros de classificação mais comuns consistem em classificar um indivíduo como doente quando na verdade ele está sadio ou classificar um indivíduo como sadio quando na verdade ele está doente. Podemos então quantificar a qualidade dos testes diagnósticos através de duas medidas: a sensibilidade (θ) que é a probabilidade de o teste identificar corretamente a doença entre aqueles que a possuem, e a especificidade (ϕ), que é a probabilidade de o teste excluir corretamente aqueles que não possuem a doença.

O problema de erros de classificação no contexto de estimação de uma proporção foi primeiramente abordado em [1] e alguns autores, como [3, 7], apresentam propostas para incorporar as probabilidades de erros aos modelos e evitar que estimativas viesadas sejam obtidas. Nos estudos de sobrevivência, [4, 6] incorporam a sensibilidade e a especificidade na função de verossimilhança para casos em que os indivíduos são analisados uma ou duas vezes, respectivamente.

Objetivo

Considere um estudo em que N indivíduos são avaliados periodicamente nos instantes de tempo t_1, t_2, \dots, t_k (não é necessário que $t_1 = 1, t_2 = 2, t_k = k$ nem que os intervalos $(t_{j-1}, t_j]$ sejam de mesmo tamanho). Suponha que o teste para detecção da falha (um evento não catastrófico) seja sujeito a erros de classificação ($\theta < 1$ e $\phi < 1$, parâmetros fixos e conhecidos) e os indivíduos são avaliados até o momento t_j ($j = 1, \dots, k$) em que a falha é detectada, ou seja, indivíduos com indicador de falha positivo são excluídos mesmo que essa falha seja falsa.

Nesse contexto de um estudo de sobrevivência com dados grupados, pretende-se propor um modelo que incorpore as probabilidades de erros de classificação e, assim, seja capaz de fornecer estimativas não viciadas para os parâmetros da distribuição do tempo de falha.

Metodologia

Considere um indivíduo cuja falha foi detectada, por exemplo, em t_3 . Como o teste possui $\theta < 1$ e $\phi < 1$, então a falha (1) pode ter ocorrido entre t_0 e t_1 e não ter sido detectada em t_1 e em t_2 e ter sido detectada em t_3 , (2) pode ter ocorrido entre t_1 e t_2 , não ter sido detectada em t_2 e ter sido detectada em t_3 , (3) pode ter realmente ocorrido em t_3 e (4) pode não ter ocorrido antes de t_3 mas ter sido detectada erroneamente em t_3 . Então, a contribuição desse indivíduo na função de verossimilhança é dada por: $[(S(t_0) - S(t_1))(1 - \theta)^2\theta + (1 - \theta)\theta\phi(S(t_2) - S(t_3)) + (\theta)\phi^2(S(t_2) - S(t_3)) + \phi^2(1 - \phi)S(t_3)]$.

De forma análoga, a contribuição para a verossimilhança de uma unidade com falha detectada em t_j ; $j = 1, \dots, k$ é: $[\phi^{(j-1)}(1 - \phi)S(t_j) + \theta \sum_{l=1}^j \phi^{(l-1)}(1 - \theta)^{(j-l)}(S(t_{l-1}) - S(t_l))]$, em que $S(t) = P(T \geq t)$ e $S(t_0) = 1$.

Considere agora um indivíduo censurado em t_3 . Nesse caso, a falha (1) pode ter ocorrido antes do tempo t_1 e não ter sido detectada em t_1, t_2 e t_3 , (2) pode ter ocorrido entre t_1 e t_2 e não ter sido detectada em t_2 e t_3 , (3) pode ter ocorrido entre t_2 e t_3 e não ter sido detectada em t_3 ou (4) pode realmente não ter falhado antes de t_3 . Assim, a contribuição deste indivíduo é dada por: $[(S(t_0) - S(t_1))(1 - \theta)^3 + (S(t_1) - S(t_2))\phi(1 - \theta)^2 + (S(t_2) - S(t_3))\phi^2(1 - \theta) + S(t_3)\phi^3]$.

Seguindo o mesmo raciocínio, a contribuição para a verossimilhança de uma unidade que foi censurada em t_j ; $j = 1, \dots, k$ é: $[\phi^j S(t_j) + (1 - \theta) \sum_{l=1}^j (S(t_{l-1}) - S(t_l))\phi^{(l-1)}(1 - \theta)^{(j-l)}]$.

Portanto, a função de verossimilhança é dada por:

$$L(\rho|n, m, \phi, \theta) = \prod_{j=1}^k [\phi^{(j-1)}(1-\phi)S(t_j) + \theta \sum_{l=1}^j \phi^{(l-1)}(1-\theta)^{(j-l)}(S(t_{l-1}) - S(t_l))]^{n_j} \times \times [\phi^j S(t_j) + (1-\theta) \sum_{l=1}^j (S(t_{l-1}) - S(t_l))\phi^{(l-1)}(1-\theta)^{(j-l)}]^{m_j}. \quad (2)$$

O modelo proposto (2) reduz-se a (1) quando $\phi = 1$ e $\theta = 1$, ou seja, quando não há erros de classificação. Além disso, esse modelo é uma generalização do proposto em [5], que consideram $t_1 = 1, t_2 = 2, t_k = k$. Métodos numéricos podem ser utilizados para maximizar (2) e obter os estimadores dos parâmetros.

Resultados e Discussões

Para avaliar o desempenho do método proposto, utilizamos simulação Monte Carlo em rotina implementada no *software R*: foram geradas 50 observações com tempo até a falha de distribuição exponencial com parâmetro igual a 1. Os tempos de avaliação de cada observação foram $t_1 = 0.5, t_2 = 1.0, t_3 = 1.5, \dots, t_8 = 4.0$. A partir dessa amostra original, foram simuladas 1000 amostras para cada cenário em que ϕ varia entre 0,80 e 0,90 e θ varia entre 0,85 e 0,95. Os resultados obtidos são sumarizados na Tabela 1, em que $\hat{\alpha}_*$ e $\hat{\alpha}_{surv}$ são estimativas obtidas pelo método proposto e tradicional, respectivamente; e $\text{Var}(\hat{\alpha})$ é a variância assintótica calculada a partir de (2).

As estimativas do método proposto apresentaram menor vício relativo em comparação às estimativas do método tradicional de análise de sobrevivência. Nesse sentido, a eficiência do método cresce na medida em que θ e ϕ decrescem. Resultados semelhantes foram obtidos para diferentes tamanhos amostrais e distribuições.

Tabela 1: Estimação do parâmetro da distribuição do tempo de falha via simulação

θ	ϕ	Média($\hat{\alpha}_*$)	Var($\hat{\alpha}_*$)	Var($\hat{\alpha}$)	Média($\hat{\alpha}_{surv}$)
0,95	0,95	1,028	0,028	0,028	1,084
0,95	0,90	1,025	0,031	0,034	1,183
0,95	0,85	1,022	0,040	0,041	1,285
0,95	0,80	1,037	0,050	0,050	1,411
0,90	0,95	1,018	0,032	0,028	1,041
0,90	0,90	1,032	0,038	0,036	1,149
0,90	0,85	1,030	0,043	0,043	1,251
0,90	0,80	1,033	0,053	0,053	1,361
0,85	0,95	1,027	0,031	0,030	1,015
0,85	0,90	1,034	0,042	0,037	1,111
0,85	0,85	1,034	0,050	0,046	1,209
0,85	0,80	1,058	0,059	0,060	1,330

Conclusões

O método proposto nesse trabalho apresentou desempenho satisfatório ao estimar os parâmetros da distribuição do tempo de falha na situação em que os indivíduos são avaliados periodicamente por testes sujeitos a erros de classificação conhecidos.

Referências

- [1] BROSS, I., Misclassification in 2 X 2 Tables. *Biometrics* **10**, 478-486, 1954.
- [2] COLOSIMO, E. A.; GIOLO, S. R., *Análise de Sobrevivência Aplicada*. São Paulo: Edgard Blucher, 2006.
- [3] JOHNSON, N. L.; KOTZ, S.; WU, X., *Inspection Errors for Attributes in Quality Control*. London: Chapman & Hall, 1991.
- [4] MCKEOWN, K.; JEWELL, N. P., Misclassification of Current Status Data. *Lifetime Data Analysis*, **16**, 215-230, 2010.
- [5] MEIER, A. S.; RICHARDSON, B. A.; HUGHES, J. P., Discrete Proportional Hazards Models for Mismeasured Outcomes. *Biometrics*, **59**, 947-954, 2003.
- [6] PAGGIARO, A.; TORELLI, N., The Effect of Classification Errors in Survival Data Analysis, *Statistical Methods & Applications*, **13**, 213-225, 2004.
- [7] QUININO, R. C.; PIRES, M. C.; SUYAMA, E., Estimation of the Conformance Fraction in a Presence of Misclassification Errors: a Bayesian Analysis in an Absence of Expert's Knowledge. *Brazilian Journal of Operations & Production Management*, **7(1)**, 181-193, 2010.

Agradecimentos

Os autores agradecem o apoio recebido da PRPq-UFMG.