O EFEITO DA PRESENÇA DE CENSURAS ALEATÓRIAS NOS INTERVALOS DE CONFIANÇA PARA OS PARÂMETROS DO MODELO LOG-LOGÍSTICO DUPLO

Cleber Giugioli CARRASCO[1]
Francisco LOUZADA-NETO[1]

▪ RESUMO: O modelo log-logístico duplo pode ser utilizado para modelagem de dados de riscos competitivos latentes, particularmente se a função de risco for multimodal. Entretanto, cuidado é necessário para construção de intervalos de confiança para os parâmetros do modelo, particularmente na presença de censuras, visto que os procedimentos usuais podem não ser válidos. Neste artigo propomos procedimentos de simulação que podem ser utilizados para a verificação empírica de que a teoria padrão não deve ser adotada e que se apresentam como métodos alternativos de inferência, possibilitando a obtenção de intervalos de confiança adequados, mesmo na presença de uma quantidade relativamente grande de censuras. Um estudo de Monte Carlo é realizado com o objetivo de comparar os diferentes intervalos de confiança obtidos, por meio da probabilidade de cobertura.

▪ PALAVRAS-CHAVE: Análise de sobrevivência e confiabilidade, estimação intervalar, função de risco, modelo log-logístico duplo, probabilidade de cobertura, técnicas de reamostragem.

1 Introdução

Em análise de sobrevivência é comum encontrarmos estudos onde os indivíduos podem vir a falecer devido a vários tipos de riscos, mas a origem da causa do falecimento é desconhecida, dando origem aos chamados riscos competitivos latentes (Louzada-Neto, 1999). Nesse tipo

1 Departamento de Estatística, Universidade Federal de São Carlos (UFSCar), caixa postal 676, CEP 13565-905, São Carlos, SP, Brasil. E-mail: dfln@power.ufscar.br

de problema, somente o valor mínimo dos tempos de vida relacionados a todos os fatores de riscos é observado.

Em um artigo recente, considerando a presença de dois riscos competitivos, Melo e Louzada-Neto (2000) propuseram a utilização do modelo log-logístico duplo para tratar dados de riscos competitivos latentes, em particular, quando os dados apresentam uma função de risco bimodal. Entretanto, estes autores não consideraram a presença de censuras nos dados, o que é comum em análise de sobrevivência. Vários fatos podem colaborar para a ocorrência de censuras. Entre outros podemos citar: o evento de interesse (falecimento) pode não ocorrer até o final do tempo de estudo, o paciente pode abandonar o experimento antes da observação do evento de interesse, comprometendo a observação do valor do tempo de vida para alguns indivíduos. Dessa forma existe a necessidade da introdução de uma variável indicadora de censura, ou simplesmente censura, que indique se o valor do tempo de vida para um determinado indivíduo foi ou não observado.

Neste trabalho estudamos o efeito da introdução de censuras aleatórias, o que é comum a dados vinculados a experimentos médicos, na estimação intervalar para os parâmetros do modelo log-logístico duplo. Metodologias de estimação intervalar baseadas em processos de reamostragem são propostas. Na Seção 2 apresentamos a formulação do modelo e discutimos brevemente o procedimento de estimação de máxima verossimilhança. Os procedimentos de estimação intervalar são discutidos na Seção 3, incluindo intervalos assintóticos, intervalos percentis bootstrap paramétricos e não paramétricos. Na Seção 4 são apresentados os resultados de um pequeno estudo de simulação onde dados foram gerados com três tamanhos amostrais e três quantidades diferentes de censuras, onde comparamos os diferentes intervalos de confiança através da probabilidade de cobertura. O artigo é finalizado com algumas conclusões na Seção 5.

2 O modelo Poli-Log-Logístico

Suponha que indivíduos são submetidos a m ($\geq$2) causas de morte diferentes. Seja o tempo de vida relacionado a j-ésima causa x_j com uma função densidade de probabilidade $f_j(x)$. Somente o tempo mínimo entre os vários riscos, $ti = min(x_1,...,x_m)$ é observado para cada indivíduo. Assumindo que os x_j são independentes, a função de risco é dada por $h(t) = \sum_{j=1}^{m} h_j(t)$ (Louzada-Neto, 1999).

Se assumirmos que o tempo de vida relacionado a um particular risco tem uma distribuição log-logística, então o mínimo tempo de vida entre vários riscos terá uma distribuição poli-log-logística, dada por (Melo e Louzada-Neto, 2000)

$$h(t) = \sum_{j=1}^{m} \frac{\beta_j t^{\beta_j - 1}}{\mu_j^{\beta_j} + t^{\beta_j}}, \tag{1}$$

onde μ_j e β_j são os parâmetros a serem estimados. Uma vantagem vinculada à distribuição poli-log-logística em relação a log-logística simples é acomodar várias formas de função de risco, incluindo curvas multimodais. Um caso particular importante de (1) é quando somente dois riscos ($m=2$) são considerados. Neste caso, (1) é reescrito na forma

$$h(t) = \frac{\beta_1 t^{\beta_1 - 1}}{\mu_1^{\beta_1} + t^{\beta_1}} + \frac{\beta_2 t^{\beta_2 - 1}}{\mu_2^{\beta_2} + t^{\beta_2}}. \tag{2}$$

Alguns exemplos numéricos que apresentam funções de risco bi-modais podem ser encontrados em Klein e Moeschberger (1997).

Os estimadores de máxima verossimilhança (EMV) dos parâmetros em (2) podem ser obtidos diretamente através da maximização do logaritmo da função de verossimilhança. Considere uma amostra de variáveis aleatórias independentes $t_1, t_2, \ldots, t_n$, onde $t_i = \min(x_{i1}, \ldots, x_{im})$, e considere também a cada t_i uma variável indicadora de censura aleatória definida por $\delta_i = 1$, se t_i é uma observação do tempo de falha, e $\delta_i = 0$, se t_i é uma observação censurada, portanto, o log da função de verossimilhança é dado por

$$log\ L(\mu, \beta) = \sum_{i=1}^{n} \left\{ \delta_i\ log \left(\sum_{j=1}^{m} \frac{\beta_j t_i^{\beta_j - 1}}{\mu_j^{\beta_j} + t_i^{\beta_j}} \right) - \sum_{j=1}^{m} \frac{\mu_j^{\beta_j}}{\mu_j^{\beta_j} + t_i^{\beta_j}} \right\}. \tag{3}$$

3 Estimação intervalar

Para amostra de tamanho suficientemente grande, inferências para os parâmetros podem ser baseadas nos EMV e em seus erros padrões. Considere $\theta' = (\mu_1, \mu_2, \beta_1, \beta_2)$, $L(\theta)$ a função de verossimilhança e $\hat{\theta}$ o vetor de EMV para θ, utilizando-se a distribuição normal assintótica (Cox & Hinkley, 1974) dos EMV, dada por

$$\hat{\theta} \sim N_{2j}\left(\theta, I^{-1}(\theta)\right), \tag{4}$$

onde $I(\theta)$ é a matriz de informação esperada cujos elementos são dados por $I_{ju}(\theta) = E(-(\partial l / \partial \theta_j)(\partial l / \partial \mu_u)) = -E(\partial^2 l / \partial \theta_j \partial \mu_u)$, j=1,2 e u=1,2. Onde $l(\cdot)$ é o logaritmo da função de verossimilhança.

Entretanto, em análise de sobrevivência podemos ter amostras pequenas ou moderadas, onde a aproximação (4) pode não ser válida. Alternativamente, podemos utilizar correções analíticas (Barndorff-Nielsen, 1983), entretanto, estes procedimentos estão fora dos objetivos do artigo e devem ser considerados em trabalhos futuros. Outra possibilidade é utilizar simulação paramétrica e/ou não-paramétrica (Davison e Hinkley, 1997) para contornar esse problema. Nestes casos, podemos utilizar a técnica *bootstrap* paramétrico e/ou não-paramétrico na construção de intervalos de confiança , através da reamostragem do conjunto de dados original (Davison & Hinkley, 1997). Considere μ_1 o parâmetro de interesse, mas o mesmo procedimento deve ser aplicado aos outros parâmetros do modelo. Para cada amostra calculamos o EMV para μ_1 e temos no final de *R* reamostragens, $\hat{\mu}^*_{1,1} < \ldots < \hat{\mu}^*_{1,R}$ valores dos EMV ordenados. Utilizamos então

$$\hat{\mu}^*_{1,(R+1)\left(\frac{\alpha}{2}\right)} \text{ e } \hat{\mu}^*_{1,(R+1)\left(1-\frac{\alpha}{2}\right)} \quad (5)$$

como os limites inferiores e superiores do intervalo 100(1-α)% de confiança para μ_1, onde α é o nível de significância. Em geral utiliza-se $R = 999$.

Se as reamostras são obtidas através da reposição dos elementos da amostra original temos o *bootstrap* não-paramétrico. Se estas forem geradas através do modelo (2) ajustado com os valores dos parâmetros fixados nos EMV obtidos na amostra original temos o *bootstrap* paramétrico.

4 Resultados

A metodologia descrita na Seção 3 foi aplicada a dados gerados através de (2) considerando-se amostras com 40, 80 e 200 elementos, com os valores dos parâmetros fixos em μ_1= 1650, μ_2= 4500, β_1= 1.9 e β_2= 8, e o número de replicações *boostrap* igual a 999. Para gerar o tempo de sobrevivência *t* para cada indivíduo, consideramos $y_1 = min(X_1, X_2)$ como sendo o tempo de falha e $y_2 = min(X_1, X_2)$ o tempo de censura independente de y_1, onde X_1 e X_2 são variáveis aleatórias independentes com distribuição log-logística. E, tomamos $t = min(y_1, y_2)$ tal que $P(y_1 > y_2) = p$. Três casos foram considerados: amostra completa $(p = 0)$ e com 10% $(p = 0.1)$ e 30% $(p = 0.3)$ de censuras aproximadamente. No total foram gerados nove conjuntos de dados. Apesar de existirem outras formas de gerarmos tempos de sobrevivência censurados, quando assumimos que as censuras ocorrem aleatoriamente, independentemente dos tempos de sobrevivência, a geração dos tempos de sobrevivência *t* conforme descrito acima, em nossa experiência, permite o controle da quantidade de censuras com grande precisão, por isto este procedimento é utilizado neste trabalho.

Os resultados deste estudo de simulação foram condensados na Tabela 1 que apresenta, através de histogramas, as amplitudes dos intervalos de confiança de 90% para os parâmetros de (2). Como esperado a amplitude dos intervalos de confiança diminuem à medida que o tamanho amostral aumenta, enquanto a quantidade de censuras induz o aumento das amplitudes, em particular para amostras pequenas. Este comportamento é também observado para os vícios dos estimadores obtidos pelos vários métodos descritos em (3), que são dados na Tabela 2.

É notório o fato de que, apesar das amplitudes dos intervalos de confiança gerados por todos os métodos descritos em (3) serem similares quando a amostra é completa (0% de censura), quando a porcentagem de censuras aumenta, os intervalos de confiança obtidos via *bootstrap* apresentam-se em geral maiores do que os obtidos via teoria assintótica, indicando que a inflação das variâncias promovida pela presença de censuras nos dados é captada com maior eficiência pelos métodos de reamostragem do que através da teoria assintótica, que tende a subestimar a variabilidade presente nos dados. Este fato também foi observado por Davison & Hinkley (1997) que consideraram uma função de risco Weibull simples.

A Figura 1 apresenta os gráficos "qq-plots" e os "histogramas" das distribuições empíricas dos estimadores dos parâmetros obtidas via *bootstrap* paramétrico, que em geral parecem não ser normalmente distribuídas, indicando que, neste caso, a teoria usual de verossimilhança pode não propiciar resultados suficientemente adequados, corroborando os resultados obtidos nas Tabelas 1 e 2. Também na Figura 1, para verificar a aplicabilidade da teoria assintótica são apresentados os valores do teste de normalidade de Kolmogorov-Smirnov, onde rejeitamos a hipótese de normalidade dos EMV obtidos via técnica *bootstrap* paramétrico ao nível de 0.10 em todos os casos.

Tabela 1 – Histogramas das amplitudes dos intervalos de confiança de 90% para as amostras de tamanho 40, 80 e 200 com ( ) 0%, ( ) 10% e ( ) 30% de censuras.

I.C.				

Tabela 2 – Vícios estimados em relação aos EMV para amostras de tamanho 40, 80 e 200 com 0%, 10% e 30% de censuras

Parâmetros	Bootstrap Paramétrico			Bootstrap Não-Paramétrico		
μ_1	0%	10%	30%	0%	10%	30%
40	0.0315	0.1134	0.6523	0.0093	0.0559	0.0649
80	0.0051	0.1002	0.3343	0.0069	0.0125	0.0329
200	0.0094	-0.0745	-0.2473	0.0059	0.0060	0.0131
β_1	0%	10%	30%	0%	10%	30%
40	-0.0025	-0.0352	-0.0491	-0.0031	-0.0328	-0.0401
80	0.0029	-0.0059	0.0756	0.0076	0.0146	0.0221
200	-0.0014	0.0157	0.0216	0.0082	0.0067	0.0665
μ_2	0%	10%	30%	0%	10%	30%
40	-0.0129	0.0166	0.0912	-0.0200	-0.0304	-0.0404
80	-0.0054	0.0233	0.1072	-0.0063	0.0064	-0.0074
200	-0.0048	-0.0370	-0.0787	0.0005	0.0015	-0.0280
β_2	0%	10%	30%	0%	10%	30%
40	0.1972	0.2103	0.2550	-0.0624	0.0910	0.2677
80	0.2099	0.2195	0.3264	0.1544	0.2280	0.2641
200	0.0278	0.0354	0.0564	0.0689	0.0932	0.1383

Portanto, o método de reamostragem possibilitou a verificação empírica que a aplicabilidade da teoria assintótica usual neste caso não é válida, uma vez que a normalidade dos EMV dado em (4) não é satisfeita.

Também foi feito um estudo de Monte Carlo para comparar os métodos de construção de intervalos de confiança através da probabilidade de cobertura (PC). Amostras foram geradas com 40, 80 e 200 elementos, com os valores dos parâmetros fixados em μ_1= 1650, μ_2= 4500, β_1= 1.9 e β_2= 8, e $R = 399$. Novamente três casos foram analisados: amostra completa, com 10% e 30% de censuras aproximadamente.

A PC é determinada repetindo o procedimento de construção do intervalo de confiança *B* vezes nas quais verificamos em cada uma se o verdadeiro valor do parâmetro está contido neste intervalo de confiança. Assim, a PC pode ser calculada através de $\sum_{b=1}^{B} \varphi(vp \in IC)$, onde $\varphi(\cdot)$ é uma função indicadora, vp é o verdadeiro valor do parâmetro e IC é o intervalo de confiança. Neste trabalho consideramos $B = 500$.

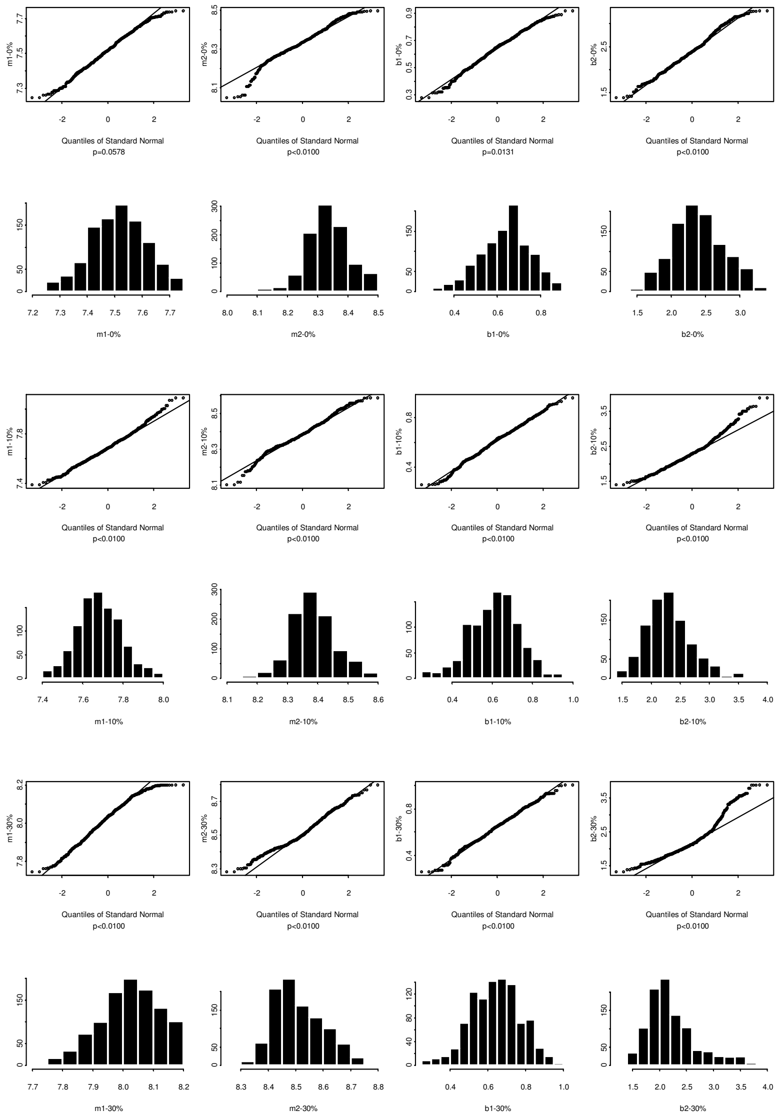

FIGURA 1 – QQ-Plots e Histogramas das distribuições empíricas dos estimadores dos parâmetros obtidas via bootstrap paramétrico para a amostra de tamanho 80, completa, com 10% e 30% de censuras respectivamente.

A Tabela 3 apresenta os resultados das PC para todos os intervalos de confiança apresentados, notamos que as PC obtidas pelos procedimentos via técnica de reamostragem estão mais próxima da probabilidade de cobertura nominal que foi fixada em 0.90, do que os intervalos de confiança via teoria assintótica, particularmente na presença de uma quantidade relativamente grande de censuras e, para amostras pequenas e moderadas. Entre os intervalos de confiança via reamostragem, os intervalos percentis *bootstrap* não-paramétrico possuem PC mais próxima das PC nominal em comparação com os intervalos obtidos via *bootstrap* paramétrico.

Conclusões

Quando temos dados com a presença de dois riscos competitivos latentes, em particular quando os dados apresentam funções de riscos bimodais, o modelo log-logístico duplo pode ser utilizado para o ajuste. Entretanto, cuidado é necessário para construção de intervalos de confiança para os parâmetros, visto que os procedimentos usuais podem não ser válidos. Neste artigo, estudamos o efeito da presença de censuras aleatórias na construção de intervalos de confiança para os parâmetros do modelo (2), independentemente do tipo de intervalo de confiança. Os procedimentos de simulação adotados permitem a verificação empírica de que a teoria padrão não deve ser adotada e apresentam-se como métodos alternativos de inferência, possibilitando a obtenção de intervalos de confiança adequados, mesmo na presença de uma quantidade relativamente grande de censuras. Além disso, o procedimento de construção de intervalo de confiança via técnica *bootstrap* não-paramétrico apresenta PC mais próxima da PC nominal do que os outros dois procedimentos, particularmente para amostras pequenas e moderadas e, na presença de dados censurados.

Agradecimentos

Os autores agradecem os comentários e sugestões do corpo editorial da revista bem como dos revisores e o apoio da CAPES e CNPq, que financiaram parcialmente este trabalho.

CARRASCO, C. G.; LOUZADA-NETO, F. The presence of randomizes censors in confidence intervals for parameters of the poly-log-logistic model. *Rev. Mat. Estat.*, (São Paulo), v.21, n.1, p.85-95, 2003.

- *ABSTRACT: The poly-log-logistic model can be used for modeling latent competing risk data when the hazard function is multimodal. We study such model and its properties in the presence of censored data. Interval estimation for the parameter of the model can run into problems if asymptotic theory is used, particularly if the sample size is small or moderate. We overcome this problem by constructing bootstrap confidence intervals. We run a Monte Carlo study to compare, the proposed confidence interval procedures through to coverage probability.*

- *KEYWORDS: Bootstrap techniques, coverage probability, hazard function, maximum likelihood estimation, bi-log-logistic model, reliability and survival analysis.*

Referências

BARNDORFF-NIELSEN, O. On a formula for the distribution of maximum likelihood estimador. *Biometrika*, v.70, p.343-365, 1983.

COX, D.R.; HINKLEY, D.V. *Theoretical statistics*. London: Chapman and Hall, 1974. 511p.

DAVISON, A.C.; HINKLEY, D.V. *Bootstrap methods and their application*. Cambridge: Cambridge University Press, 1997. 582p.

KLEIN, J.P.; MOESCHBERGER, L.M. *Survival analysis: techniques for censored and truncated data*. New York: Springer, 1997. 501p.

LOUZADA-NETO, F. Poly-hazard regression models for lifetime data. *Biometrics*, v.55, p.1121-1125, 1999.

MELO, A. A. S.; LOUZADA-NETO, F. Estimação intervalar para os parâmetros do modelo log-logístico múltiplo. *Rev. Mat. Estat.*, v.18, p.125-133, 2000.

Recebido em 25.07.2001.
Aprovado após revisão em 17.10.2002.