Análise de Diagnóstico no Modelo de Regressão Bivariado com Fração de Cura

Juliana B. Fachini* *Universidade de São Paulo*

Edwin M. M. Ortega † *Universidade de São Paulo*

1 Introdução

Dados de sobrevivência com fração de cura estão se tornando cada vez mais comum em estudos epidemiológicos e ensaios clínicos. O modelo com fração de cura univariado introduzido por Berkson e Gage (1952), tem sido usado para modelar dados de tempo de falha de vários tipos de câncer, incluindo câncer de mama, leucemia, próstata, cabeça e pescoço, dentre outros que apresentam uma significativa proporção de indivíduos curados. Contudo, essa modelagem não apresenta estrutura de riscos competitivos para $S_{pop}(t)$, que representa a função de sobrevivência populacional, se covariáveis são introduzidas. Alternativamente, Yakovlev et al. (1993) propuseram uma nova classe de modelos de mistura com a estrutura de riscos competitivos.

Em análise de sobrevivência, tem-se o interesse em modelar conjuntamente variáveis aleatórias que representam tempos de falha, como o tempo de reincidência do câncer de dois diferentes órgãos, tempo até a primeira e segunda infecção, tempo de reincidência do câncer e da morte, entre outros interesses. Como uma alternativa ao modelo com fração de cura univariado discutido por Yakovlev et al. (1993) e Asselain et al. (1996), Chen, Ibrahim e Sinha (2002) propõem um modelo com fração de cura multivariado, que prova ser bastante útil para modelar dados multivariados com variáveis aleatórias de falha conjunta apresentando fração de cura e cada variável aleatória marginal do tempo de falha poderia também apresentar uma fração de cura.

Por outro lado, quando se ajusta um modelo a um conjunto de dados, é importante estudar a robustez dos resultados obtidos com relação à presença de observações extremas ou observações influentes que podem causar alterações nos resultados das estimativas dos parâmetros do modelo. Para detectar observações influentes nas estimativas dos parâmetros pode ser considerada uma análise de sensibilidade. Neste trabalho é proposto o uso das medidas de infuência local e local total no modelo proposto por Chen, Ibrahim e Sinha (2002), e um conjunto de dados reais é usado para ilustrar a teoria em estudo.

2 Metodologia

2.1 Modelo

Para descrever o modelo bivariado com fração de cura considere: (T_1, T_2) são tempos de falha bivariados. Para um indivíduo arbitrário da população, sejam (N_1, N_2) denotando variáveis latentes (não-observadas) para (T_1, T_2), respectivamente, sendo que, N_k tem distribuição Poisson com média $\theta_k m$, $k = 1, 2$, e (N_1, N_2) são independentes.

A quantidade m é um componente de fragilidade no modelo que induz a correlação entre as variáveis latentes (N_1, N_2), tendo distribuição estável positiva com parâmetro α, em que $0 < \alpha < 1$.

*Address: ESALQ, Universdade de São Paulo, Piracicaba, Brasil. E-mail: jfachini@esalq.usp.br

†Address: ESALQ, Universidade de São Paulo, Piracicaba, Brasil. E-mail: edwin@esalq.usp.br

Pode-se assumir várias distribuições para m, Chen, Ibrahim e Sinha (2002) escolhem a distribuição estável positiva por ser flexível para dados de sobrevivência multivariados.

Seja $R_l = (R_{1l}, R_{2l})$ denotando o tempo aleatório para o l-ésimo fator de risco latente causar o evento de interesse, sendo, R_l, conhecido como o tempo latente para $T_l = (T_{1l}, T_{2l})$. Os vetores aleatórios $R_l, l = 1, 2, \ldots$ são assumidos independentes e identicamente distribuídos, com função distribuição acumulada $F_k(t|\boldsymbol{\lambda}_k) = 1 - S_k(t|\boldsymbol{\lambda}_k), k = 1, 2$, $\boldsymbol{\lambda}_k$ é o vetor de parâmetros desconhecidos da distribuição F_k e F_k é independente de N.

O tempo de sobrevivência observado é definido como a variável aleatória $T_k = min(R_{kl}, 0 \leq l \leq N_k)$, em que a $P(R_0 = \infty) = 1$ representa a cura e, N_k é independente da sequencia $R_{k1}, R_{k2}, \ldots$, para $k = 1, 2$. De Chen, Ibrahim e Sinha (2002), a função de sobrevivência populacional dado m é dada por:

$$\begin{aligned} S_{pop}(t_1, t_2|m) &= \prod_{k=1}^{2} [P(N_k = 0) + P(R_{k1} > t_k, \ldots, R_{kN} > t_k, N_k \geq 1)] \\ &= \prod_{k=1}^{2} \left[\exp(-m\theta_k) + \left(\sum_{r=1}^{\infty} S_k(t_k|\boldsymbol{\lambda}_k)^r \frac{(m\theta_k)^r}{r!} \exp(-m\theta_k) \right) \right] \\ &= \prod_{k=1}^{2} [\exp(-m\theta_k + m\theta_k S_k(t_k|\boldsymbol{\lambda}_k))] \\ &= \exp[-m(\theta_1 F_1(t_1|\boldsymbol{\lambda}_1) + \theta_2 F_2(t_2|\boldsymbol{\lambda}_2))], \end{aligned}$$

em que $P(N_k = 0) = P(T_k = \infty) = \exp(-\theta_k), k = 1, 2$. É importante notar que N e R_l apenas facilitam a construção do modelo e não há necessidade de alguma interpretação biológica ou física para o modelo ser válido.

Ao considerar que m possui uma distribuição estável positiva com parâmetro α, e utilizando a transformada de Laplace, neste caso a função de sobrevivência populacional é dada por:

$$S_{pop}(t_1, t_2) = \exp\{-[\theta_1 F_1(t_1|\boldsymbol{\lambda}_1) + \theta_2 F_1(t_2|\boldsymbol{\lambda}_2)]^{\alpha}\}, \tag{1}$$

tendo estrutura de riscos proporcionais ao introduzir covariáveis no modelo por meio de (θ_1, θ_2). A variável de fragilidade m induz a correlação de T_1 e T_1 e relaxa a suposição Poisson de N_1 e N_2 ao adicionar a mesma variação extra Poisson por meio das respectivas médias $\theta_1 m$ e $\theta_2 m$.

Da equação (1), pode-se observar que

$$S_{pop}(\infty, \infty) = \exp[-(\theta_1 + \theta_2)],$$

representa a fração de cura conjunta.

As funções de sobrevivência marginais de (1) são

$$S_k(t) = \exp\{-\theta_k^{\alpha}[F_k(t_k|\boldsymbol{\lambda}_k)]^{\alpha}\},$$

com probabilidade de cura $\exp(-\theta_k^{\alpha})$ para T_k, $k = 1, 2$. É importante notar que cada função de sobrevivência marginal tem estrutura de riscos proporcionais ao introduzir covariáveis em θ_k, isto é, $\theta_k = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)$.

Ao considerar uma amostra observada contendo variáveis $(t_{1k}, \delta_{1k}, \boldsymbol{x}_1), \ldots, (t_{nk}, \delta_{nk}, \boldsymbol{x}_n)$, sendo t_{ik} o tempo de vida do k-ésimo evento de interesse do i-ésimo indivíduo, δ_{ik} a respectiva variável indicadora de censura no k-ésimo evento de interesse do i-ésimo indivíduo e $\boldsymbol{x}_i$ o vetor de covariáveis associado ao i-ésimo indivíduo, em que $k = 1, 2$ e $i = 1, 2, \ldots, n$.

Para realizar os procedimentos inferênciais, a função de verossimilhança descrita em Lawless (2003) para dados bivariados foi considerada. O logaritmo da função de verossimilhança é representado pela equação:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n} \left\{ \delta_{i1}\delta_{i2} f_{pop}(t_{i1}, t_{i2}) + \delta_{i1}(1 - \delta_{i2}) \left[\frac{-\partial S_{pop}(t_{i1}, t_{i2})}{\partial t_{i1}} \right] + (1 - \delta_{i1})\delta_{i2} \left[\frac{-\partial S_{pop}(t_{i1}, t_{i2})}{\partial t_{i2}} \right] + \right. \\ &\left. (1 - \delta_{i1})(1 - \delta_{i2}) S_{pop}(t_{i1}, t_{i2}) \right\} \end{aligned} \tag{2}$$

em que $S_{pop}(t_{i1}, t_{i2})$ é a função de sobrevivência populacional definida na equação (1), $f_{pop}(t_{i1}, t_{i2}) = \frac{\partial^2 S_{pop}(t_{i1}, t_{i2})}{\partial t_{i1} \partial t_{i2}}$ é a função densidade conjunta de (t_{i1}, t_{i2}), $\boldsymbol{\varphi} = (\alpha, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)^T$ é o vetor de parâmetros desconhecidos, sendo que $\boldsymbol{\theta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ para $k = 1, 2$ e $\boldsymbol{\lambda}_k$ é o vetor de parâmetros da distribuição F_k.

Como enunciado anteriormente o parâmetro α, que representa a associação dos tempos de falha, está definido entre $(0, 1)$. Para maximizar a função de verossimilhança definida em (2) sujeito a restrição do parâmetro α, e a demais parâmetros dependendo da distribuição considerada para os tempos de falha. É necessário considerar o método da função barreira adaptada (LANGE, 1999) que é uma combinação do método barreira logaritmo com o algoritmo EM. Sendo assim, o logaritmo da função de verossimilhança sujeita as restrições lineares é representado por

$$l_R(\boldsymbol{\varphi}, \vartheta) = l(\boldsymbol{\varphi}) + \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T \boldsymbol{\varphi} - c_j) \tag{3}$$

em que o parâmetro de ajuste é uma constante positiva, $\vartheta > 0$, $\boldsymbol{u}_j^T \boldsymbol{\varphi} - c_j$ é o conjunto de restrições de inequações lineares, que pode reduzir-se em, $\boldsymbol{v}_j^t \boldsymbol{\varphi} = d_j$ que é o conjunto de restrições de equações lineares para $j = 1, 2, \ldots, q$, $\boldsymbol{\varphi} = (\alpha, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)^T$.

Neste trabalho, o *software* R foi utilizado para obter as estimativas de máxima verossimilhança com restrição nos parâmetros por meio da função *constrOptim*. Maiores detalhes sobre o método da função Barreira adaptada ver, por exemplo, Lange (1999).

3 Influência Local com Restrições

Uma importante etapa na análise estatística de modelos é verificar o quanto as estimativas obtidas a partir do modelo proposto são resistentes a pequenas perturbações nos dados, ou no modelo. Se o modelo ajustado não apresentar uma boa descrição dos dados que foram observados, o mesmo pode conduzir a inferências errôneas.

Devido a isso, destaca-se a importância de realizar um estudo sobre a robustez dos resultados obtidos, considerando vários aspectos que envolvem a formulação do modelo, e as estimativas dos seus parâmetros, ou seja, realizar uma análise de sensibilidade. Tal análise será considerada aqui sob a metodologia de influência local e local total.

Uma medida de análise de sensibilidade é a metodologia de influência local. Gu e Fung (2001) mostraram que se o afastamento da verossimilhança restrita $LD(\boldsymbol{w}) = 2[l_R(\hat{\boldsymbol{\varphi}}, \vartheta) - l_R(\hat{\boldsymbol{\varphi}}_{\boldsymbol{w}}, \vartheta)]$ é usado, em que $\hat{\boldsymbol{\varphi}}_{\boldsymbol{w}}$ denota o estimador de máxima verossimilhança sujeito a restrições nos parâmetros sob o modelo perturbado, a curvatura normal, seguindo a idéia de Cook (1986), para $\hat{\boldsymbol{\varphi}}$ na direção $\boldsymbol{d}$, $\| \boldsymbol{d} \| = 1$, é dada por $C_{\boldsymbol{d}} = 2\|\boldsymbol{d}^T \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\varphi}, \vartheta)^{-1} \boldsymbol{\Delta} \boldsymbol{d}\|$, em que $\boldsymbol{\Delta} = \partial^2 l_R(\boldsymbol{\varphi}, \vartheta | \boldsymbol{w}) / \partial \boldsymbol{\varphi} \partial \boldsymbol{w}^T$ é uma $b \times n$ matriz que depende do esquema de perturbação usado, em que b é o número de parâmetros e $\ddot{\boldsymbol{L}}_R(\boldsymbol{\varphi}, \vartheta) = -\partial^2 l_R(\boldsymbol{\varphi}, \vartheta) / \partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^T$ é a matriz de informação do modelo postulado, ambas matrizes avaliadas em $\boldsymbol{\varphi} = \hat{\boldsymbol{\varphi}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$, em que $\boldsymbol{w}_0$ é o vetor de não perturbação (veja Cook, 1986). Uma importante informação, é saber a direção que produz a maior influência local na estimativa dos parâmetros. Essa direção é dada por $\boldsymbol{d}_{max}$, sendo este o autovetor normalizado correspondente ao maior autovalor $C_{\boldsymbol{d}_{max}}$ da matirz $F = \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\varphi}, \vartheta)^{-1} \boldsymbol{\Delta}$. O gráfico do autovetor $\boldsymbol{d}_{max}$ contra a ordem das observações pode identificar as observações mais influentes para o esquema de perturbação considerado.

Lesaffre e Verbeke (1998) sugerem considerar também a direção do i-ésimo indivíduo que corresponderia a $\boldsymbol{d}_i = (0, \ldots, 1, \ldots, 0)^T$, tal que o i-ésimo elemento é um. Sendo assim, a curvatura normal chamada influência local total do i-ésimo indivíduo, é representada por:

$$C_i = 2|\Delta_i^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\varphi}, \vartheta)^{-1} \Delta_i|.$$

O gráfico de C_i contra a ordem das observações pode ser usado como diagnóstico em influência local. Os possíveis pontos, tal que $C_i \geq 2\bar{C}$ em que $\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$, merecem uma atenção especial.

Ao considerar as medidas de influência local e local total podem ser realizados os esquemas de perturbação de casos, em que é suposto uma perturbação no logaritmo da verossimilhança. Perturbação da variável resposta por meio da adição de um fator de perturbação, no caso de respostas bivariadas pode-se perturbar cada resposta individualmente, ou conjuntamente. E a perturbação da covariável, em que é adicionado um fator de perturbação em uma particular variável explicativa contínua.

4 Aplicação

Os dados utilizados são referentes aos tempos de sobrevivência de retinopatia diabética que começou em 1971. Pacientes com retinopatia diabética em ambos os olhos e com acuidade visual menor ou igual a 20/100 para ambos os olhos, fizeram parte do estudo. Um olho foi selecionado aleatoriamente para o tratamento e o outro foi observado sem tratamento. No total, 1742 pacientes foram acompanhados durante 7 anos, e no final 197 pacientes fizeram parte do subconjunto em estudo definido por algum critério de estudo de retinopatia diabética. Assume-se t_{i1} e t_{i2} são os tempos de falha para tratamento e controle respectivamente, com respectiva variável indicadora de censura δ_{i1} e δ_{i2}. A covariável considerada foi $x_{i1} = 1$, se o paciente é um diabético adulto e $x_{i1} = 0$, se o paciente é um diabético jovem.

Na Figura 1 é apresentado as estimativas de sobrevivência de Kaplan-Meier e o ajuste marginal da distribuição Weibull para ambos os tempos, verificando que é adequado supor distribuição Weibull para os tempos de sobrevivência. Também foi feito um gráfico que verificou a suposição de riscos proporcionais, mas esse não será apresentado aqui.

Por meio da Figura 1 observa-se a existência de uma significativa fração de indivíduos curados para ambos tempos, pois o limite da estimativa de sobrevivência de t_1 tende a $0,67$, $\lim_{t_1\to\infty} \hat{S}(t_1) = 0,67$, e de t_2 tende a $0,39$, $\lim_{t_1\to\infty} \hat{S}(t_2) = 0,39$. Por esse motivo foi considerado o modelo bivariado com fração de cura descrito em metodologia e suposto distribuição marginal Weibull para t_1 e t_2, consequentemente, com função de distribuição acumulada definida por:

$$F_k(t_k|\boldsymbol{\lambda}_k) = 1 - \exp[-(t_k/\xi_k)^{\gamma_k}]$$

para $k = 1, 2$, logo a função de sobrevivência populacional é dada por:

$$S_{pop}(t_1, t_2) = \exp\Big\{ - \Big[\exp(\beta_{01}\boldsymbol{x}_0 + \beta_{11}\boldsymbol{x}_1)\{1 - \exp[-(t_1/\xi_1)^{\gamma_1}]\} + \exp(\beta_{02}\boldsymbol{x}_0 + \beta_{12}\boldsymbol{x}_1)\{1 - \exp[-(t_2/\xi_2)^{\gamma_2}]\}\Big]^{\alpha}\Big\}.$$

Tabela 1: Estimativa de máxima verossimilhança para o modelo de bivariado com fração de cura na estrutura de riscos competitivos

Parâmetro	Estimativa	Erro-Padrão	valor de p
β_{01}	-0,8310	0,2968	0.0051
β_{11}	-0,5461	0,3453	0.1138
β_{02}	0,3299	0,7639	0.6658
β_{12}	0,4421	0,2473	0.0739
ξ_1	36,7098	12,5378	-
γ_1	1,2806	0,1830	-
ξ_2	85,7296	80,1823	-
γ_2	1,1242	0,1428	-
α	0,8043	0,0585	-

Da Tabela 1 a fração de cura conjunta médio estimado foi $\hat{S}_{pop}(\infty, \infty) = 0,169$, a fração de cura marginal médio estimado para t_1 foi $\hat{S}_1(\infty) = 0,649$, e a fração de cura marginal médio

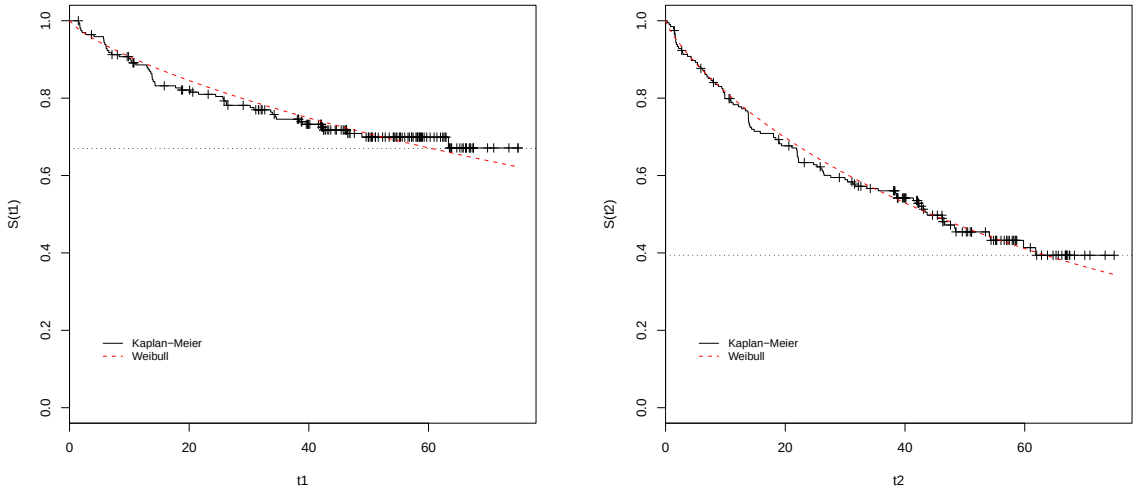

Figura 1: Curvas de sobrevivência estimadas marginalmente por Kaplan-Meier para os dados de retinopatia

estimado para t_2 foi $\hat{S}_1(\infty) = 0.223$. Ao comparar as estimativas da proporção de indivíduos curados marginais obtidas por meio da modelagem com a probabilidade calculada empiricamente pelas curvas de Kaplan-Meier apresentadas na Figura 1, observa-se que os valores são próximos.

A covariável em estudo é significativa para o tempo de falha controle para a fração de cura. A estimativa do parâmetro α indica pouca associação entre t_1 e t_2, pois quanto mais próximo de 1 estiver a estimativa de α menos correlacionado estão os tempos, e quanto mais próximo de 0 estiver a estimativa de α mais correlacionado estão os tempos de falha.

Ao realizar a análise de influência local e local total, foi considerado o esquema de perturbação de casos, perturbação da variável t_1, perturbação da variável t_2, e perturbação da variável t_1 e t_2 conjuntamente. Em que as observações $\#5, \#22, \#49, \#102, \#111$ e $\#172$ são as que mais se destacam das demais, considerando os quatro esquemas em análise. Essa conclusão foi obtida por meio de gráficos que não serão apresentados neste texto.

Essas observações apresentam algumas características em comum, isto é, as observações $\#5, \#22$ e $\#172$ são censuradas em um dos tempos e falharam no outro, tendo uma grande diferença entre o t_1 e t_2. Já as observações $\#49, \#102$ e $\#111$ falharam em ambos os tempos em estudo.

5 Conclusões

Neste trabalho, discutiu-se a aplicação do modelo bivariado com fração de cura na estrutura de riscos competitivos em um conjunto de dados, obtendo resultados coerentes com a teoria do modelo em estudo. Uma análise de influência local e local total foi realizada, encontrando possíveis pontos influentes, sendo necessário uma investigação mais rigorosa desses pontos em conjunto com o pesquisador dos dados e a implementação de outras metodologias, bem como, influência global e impacto das observações influentes para concluir se elas causam alterações nas estimativas dos parâmetros e devem ser retiradas da análise.

6 Referências

ASSELAIN, B.; FOURQUET, A.; HOANG, T.; TSODIKOV, D.; YAKOVLEV, A. Y. A parametric regression model of tumor recurrence: An application to the analysis of clinical data on breast cancer. **Statistics and Probability Letters**, v. 29, p. 271-278, 1996.

BERKSON, J.; GAGE, R.P. Survival curve for cancer patients follwing treatment. **Journal of the American Statistical Association**, Alexandria, v. 47, p. 501-515, 1952.

CHEN, M. H.; IBRAHIM, J; SINHA,D. Bayesian inference for multivariate survival data with a cure fraction. **Journal of Multivariate Analysis**, v. 80, p. 101-126, 2002.

COOK, R.D. Assement of local influence (with discussion). **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 48, n. 2, p. 133-169, 1986.

GU, H; FUNG, W.K. Local influence for the restricted likelihood with applications **Sankhya: The Indian Journal of Statistical**, Indian, v. 63, pt. 2, p. 250-259, 2001.

LANGE, K. **Numerical analysis for statisticians**. 1nd ed. New York: Springer, 1999. 356 p.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, Washington, v. 54, n. 2, p. 570-582, 1998.

YAKOVLEV, A.; ASSELAIN, B.; BARDOU, V., FOURQUET, A. HOANG, T. ROCHEFEDIERE; TSODIKOV,A. A Stochastic models of tumor latency and their biosta-tistical applications. **Biometrie et analyse de Donnes Spatio-Temporelles**, Paris, v. 12, p. 66-82, 1993.