Modelo semiparamétrico de fração de cura com aplicação a um estudo de prognóstico de pacientes com insuficiência cardíaca

Bhering, F.L; Cesar, R.S.; Tunes-da-Silva, G.
Departamento de Estatística - IME-USP

1 Introdução

Os modelos usuais de sobrevivência, apesar de terem sido desenvolvidos de forma a incorporar observações censuradas, são baseados na suposição de que os pacientes censurados (i.e., cujos eventos de interesse não foram observados) irão em algum momento apresentar o evento de interesse (óbito, por exemplo). No entanto, existem situações em que tal suposição não é verdadeira.

Um exemplo é um estudo com pacientes com insuficiência cardíaca que tiveram entrada no ambulatório de Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo, no período de janeiro de 1991 à fevereiro de 2010. Observe que esse estudo apresentou um período longo de seguimento (18 anos). Os pesquisadores envolvidos acreditam que muitos pacientes apresentam insuficiência cardíaca podem conviver com ela por muitos anos e eventualmente falecem por outras causas não relacionadas com o quadro de insuficiência cardíaca. De fato, neste estudo, as curvas de Kaplan-Meier construídas se estabilizam em um patamar acima do valor zero, indicando a existência de pacientes que não falecem devido a insuficiência cardíaca.

Nas situações em que existe uma proporção de pacientes que não apresentam o evento de interesse (denominados pacientes *curados*, apesar de, em alguns casos, tais pacientes não serem clinicamente considerados curados), modelos de sobrevivência apropriados para incorporar a possibilidade de pacientes curados são necessários. Modelos paramétricos com fração de cura foram propostos por vários autores como Farewell (1986), Yamaguchi (1992)

e Peng (1998). Neste trabalho, é considerado um modelo de fração de cura semiparamétrico, baseada em Peng (2003). Alguns outros modelos semiparamétricos também foram propostos, como Kuk e Chen (1992) e Taylor (1995), porém o modelo de Peng (2003) possui a vantagem de ser de fácil implementação computacional.

2 Modelo semiparamétrico com fração de cura

Defina a variável T como o tempo até o evento de interesse (óbito do paciente, por exemplo) e a variável U como indicador de falha não-cura deles, ou seja, $U = 1$ se o paciente que apresentou o evento e $U = 0$ se o paciente é curado. O modelo pode ser escrito da seguinte forma:

$$S(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S_u(t|\mathbf{x}) + 1 - \pi(\mathbf{z}), \tag{1}$$

em que $\mathbf{x} = (x_1, x_2, \ldots, x_m)^{'}$ e $\mathbf{z} = (z_1, z_2, ..., z_q)^{'}$ são vetores de covariáveis, $S(t|\mathbf{x}, \mathbf{z})$ é a função de sobrevivência para toda população das as covariáveis, $S_u(t|\mathbf{x}) = P(T > t|U = 1, \mathbf{x})$ é a função de sobrevivência dos pacientes não-curados dado um vetor de covariáveis $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_m})^{'}$ e $\pi(\mathbf{z})$=$P(U = 1|\mathbf{z})$ é a probabilidade de o paciente sendo não-curado dado o vetor de covariáveis $\mathbf{z}$. Observe que, neste modelo, são definidos dois vetores de covariáveis, um associado à probabilidade de cura e o outro associado à função de sobrevivência dos não curados. Esses vetores $\mathbf{x}$ e $\mathbf{z}$ podem ter variáveis em comum ou não.

No modelo de fração de cura considerado, assume-se que a probabilidade de não curados está associada ao vetor de covariáveis $\mathbf{z}$ pela função logística:

$$\text{logit}(\pi) = \eta, \tag{2}$$

em que $\eta = (1, \mathbf{z}^{'})\boldsymbol{\gamma}$ é o preditor linear e $\boldsymbol{\gamma}$ é o vetor de parâmetros desconhecidos. Para modelar a o tempo de sobrevivência dos pacientes não curados, um modelo semiparamétrico de riscos proporcionais é utilizado. Assim, a função de risco (ou função de taxa de falha) no tempo t de um paciente não-curado com vetor de covariáveis $\mathbf{x}$ pode ser escrita na forma

$$h_u(t|\mathbf{x}) = h_{u0}(t)\exp(\zeta),$$

em que $h_{u0}(t)$ é uma função de taxa de falha basal arbitrária e não-especificada, $\zeta = \mathbf{x}^{'}\boldsymbol{\beta}$ e $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos. Dessa forma, a equação (1) pode ser reescrita como

$$S(t|\mathbf{x}, \mathbf{z}) = \pi(\mathbf{z})S_{u0}(t)^{\exp(\zeta)} + 1 - \pi(\mathbf{z}), \tag{3}$$

em que $S_{u0}(t)$ corresponde a função de sobrevivência basal. Os parâmetros desconhecidos que precisam ser estimados são $\boldsymbol{\gamma}$ e $\boldsymbol{\beta}$, além da função de sobrevivência basal $S_{u0}(t)$. Existem propostas na literatura algumas formas de estimar os parâmetros, tais como simulação de Monte Carlo (Kuk e Chen, 1992) e método EM (Taylor, 1995). Nesse trabalho, será utilizado um método também baseado no algoritmo EM, proposto por Peng(2003), que apresenta a vantagem de ser de fácil implementação computacional. Nesta metodologia, o passo M no algoritmo EM consiste no ajuste de um modelo de riscos proporcionais e um modelo logístico, respectivamente, com alguns coeficientes fixos. Com isso, a implementação torna-se simples, pois rotinas já prontas em *softwares* estatísticos podem ser utilizadas.

Suponha que se tenha dados da forma $(T_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$ com $i = 1, 2, \ldots, n$, em que T_i denota o tempo de sobrevivência observado para o i-ésimo paciente, δ_i o indicador de falha, sendo 1 se T_i for o tempo de falha (óbito) e 0 caso contrário, e $\mathbf{x}_i$ e $\mathbf{z}_i$ são os valores observados para os dois vetores de covariáveis. No modelo sugerido por Peng e Dear (2000), um vetor $\mathbf{u} = (u_1, u_2, ..., u_n)^{'}$ é definido, em que u_i é o valor de U para o i-ésimo paciente. Observe que o vetor $\mathbf{u}$ não é completamente observável, pois caso $\delta_i = 1$, então $u_i = 1$, mas se $\delta_i = 0$, então u_i não é observável e pode ser tanto 1 como 0. Dado $\mathbf{u}$, a função de verossimilhança é

$$\prod_{i=1}^{n} \pi(\mathbf{z}_i)^{u_i} \{1 - \pi(\mathbf{z}_i)\}^{1-u_i} h_u(t_i|\mathbf{x}_i)^{\delta_i} S_u(t_i|\mathbf{x}_i)^{u_i}.$$

O algoritmo EM (Expectation-Maximization) começa com valores inicias $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ e $S_{u0}^{(0)}(t)$. Denote por $\boldsymbol{\gamma}^{(r)}$, $\boldsymbol{\beta}^{(r)}$ e $S_{u0}^{(r)}(t)$ as estimativas após a r-ésima iteração. O passo E do algoritmo consiste no cálculo da esperança condicional de função de log-verossimilhança, que consiste na soma das seguintes funções:

$$L_1(\boldsymbol{\gamma}) = \log \prod_{i=1}^{n} \pi(\mathbf{z}_i)^{p_i^{(r)}} \{1 - \pi(\mathbf{z}_i)\}^{1-p_i^{(r)}}, \tag{4}$$

$$L_2(\boldsymbol{\beta}, S_{u0}(t)) = \log \prod_{i=1}^{n} [h_{u0}(t_i) \exp(\zeta_i)]^{\delta_i} S_{u0}(t_i)^{p_i^{(r)} \exp(\zeta_i)}, \tag{5}$$

em que $p_i^{(r)} = E\{u_i|\boldsymbol{\gamma}^{(r)}, \beta^{(\mathbf{r})}, \mathbf{S_{u0}^{(r)}(t)}\}$=$P\{u_i = 1|\boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}, S_{u0}^{(r)}(t)\}$, e é dado por

$$\delta_i + (1 - \delta_i) \frac{\pi^{(r)}(\mathbf{z}_i) S_{u0}^{(r)}(t_i)^{\exp(\zeta_i(r)}}{1 - \pi^{(r)}(\mathbf{z}_i) + \pi^{(r)}(\mathbf{z}_i) S_{u0}^{(r)}(t_i)^{\exp(\zeta_i^{(r)})}}, \tag{6}$$

$logit[\pi^{(r)}(\mathbf{z}_i)] = \eta_i^{(r)} = (1, \mathbf{z}_i^{'})\boldsymbol{\gamma}^{(r)}$, e $\zeta_i^{(r)} = \mathbf{x}_i \boldsymbol{\beta}^{(r)}$. O passo M na $(r+1)$-ésima iteração maximiza (4) e (5) separadamente para obter $\boldsymbol{\gamma}^{(r+1)}$, $\boldsymbol{\beta}^{(r+1)}$

e $S_{u0}^{r+1)}(t)$. O algoritmo é iterado até convergir. Note que (4) é a função de log-verossimilhança de um modelo de regressão logística com n valores $p_i^{(r)}$'s vindo de uma distribuição binomial com probabilidade de resposta $\pi(z_i)$=exp($\boldsymbol{\gamma}'\mathbf{z}_i$)/[1+exp($\boldsymbol{\gamma}'\mathbf{z}_i$)]. Assim, pode ser maximizado utilizando uma programa qualquer de regressão logística para obter $\boldsymbol{\gamma}^{(r+1)}$. Do mesmo modo, a função (5) é a função de log-verossimilhança do modelos de riscos proporcionais de Cox com a inclusão de uma covariável adicional $\log p_i^{(r)}$ com coeficiente fixo 1. Assim, facilmente também programável. Os erros padrões para os parâmetros não são diretamente conseguidos por meio desse procedimento EM, no entanto Peng e Dear(2000) sugeriram um método aproximado baseado em Louis (1982) para o algoritmo EM. Outra alternativa para obtenção dos erros padrão é a utilização do jackknife ou da técnica Bootstrap.

3 Resultados e discussão

Atualmente, como desenvolvimento e aprimoramento de drogas para tratamento de doenças antes incuráveis, uma proporção de pacientes, felizmente, são considerados curados. Torna-se, dessa forma, extremamente importante a utilização de modelos estatísticos apropriados para a análise correta dos dados. Neste trabalho, foi considerado um modelo semiparamétrico de fração de cura. Neste modelo, a estimação de parâmetros de interesse é feita por meio do algoritmo EM, porém de uma forma que torna o problema de simples implementação computacional. A metodologia é ilustrada com a análise de um conjunto de dados do InCor (Instituto do Coração) com pacientes portadores de insuficiência cardíaca.

Referências

[1] FAREWELL, V. T. *Mixture models in survival analysis: are they worth the risk?*. Canadian Journal of Statistics, V. 14(3), 1986.

[2] KUK, A. Y. C.; CHEN, C.*A mixture model combining logistic regression with proportional hazards regression.* Biometrika, V. 79, p. 531-541, 1992.

[3] LOUIS, T.A. *Finding the observed information matrix when using the EM algorithm.* Journal Roy. Statist. Soc. Ser. B, V. 44(2), p. 226-233, 1982.

[4] PENG, Y. *Fitting semiparametric cure models.* Computational Statistics & Data Analysis, V.41, p. 481-490, 2003.

[5] PENG, Y.; DEAR, K.B.G.; DENHAM, J.W. *A generalized F mixture model for cure rate estimation*. Statist. Med, V. 17, p. 813-830, 1998.

[6] PENG, Y.; DEAR, K. B. G. *A nonparametric mixture model for cure rate estimation*. Biometrics, V. 56, p. 237-243, 2000.

[7] TAYLOR, J. M. G. *Semiparametric estimation in failure time mixture models*. Biometrics, V. 51, p. 899-907, 1995.

[8] YAMAGUCHI, K. *Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of permanent employment in Japan*. Journal of American Statistical Association, V. 87(418), p. 284-292, 1992.