Modelos de Fração de Cura e Censura Intervalar

Uma Aplicação para Dados de Anemia

Julio Brettas & Gisela Tunes

Instituto de Matemática e Estatística - Universidade de São Paulo

jbrettas@ime.usp.br -- tunes@ime.usp.br

IME-USP

Resumo

Em um estudo conduzido por pesquisadores da Fundação Pró-Sangue, em São Paulo, tem-se interesse em avaliar o tempo até a ocorrência de anemia em doadores de repetição. No entanto, profissionais da área médica reconhecem a existência de um grupo de doadores não suscetíveis à anemia por doações, os superdoadores. A ocorrência da anemia é definida com base em avaliações periódicas do hematócrito anterior às doações, caracterizando assim a censura intervalar do evento de interesse. São considerados para análise dois modelos semiparamétricos que contemplam tal estrutura de censura e a existência de uma proporção de indivíduos não suscetíveis. O primeiro modelo, proposto por Shen e Liu (2009), trata-se de um modelo de tempo de promoção com estimação baseada em uma variação do algoritmo EM para dados com censura intervalar.Lam et al. (2013) consideram um modelo de fragilidade com distribuição contínua positiva e massa no ponto zero para contemplar indivíduos não suscetíveis, caracterizando também a heterogeneidade das condições de saúde dos pacientes suscetíveis. Os modelos foram aplicados em dados que abrangem doações realizadas entre janeiro de 2003 e dezembro de 2006.

Introdução

Diversas pesquisas na área médica têm o intuito de analisar o tempo até a ocorrência de um evento específico. Porém, há em alguns casos a possibilidade do indivíduo não ser suscetível ao evento de interesse devido à uma possível cura obtida por um tratamento e a proporção de indivíduos assim caracterizada é denominada na literatura como fração de cura. No mesmo contexto, é muito comum ainda encontrar situações em que os tempos de ocorrência de um determinado evento não são diretamente observados, limitando o pesquisador à informação de que o evento encontra-se entre dois instantes de observação, constituindo um caso de censura intervalar.

Objetivo principal do trabalho: aplicação em dados de anemia de modelos semiparamétricos de fração de cura e censura intervalar propostos por Shen e Liu (2009) e Lam et al. (2013). Dados provenientes de estudos conduzidos na Fundação Pró-Sangue foram analisados visando estudar os fatores de influência nos tempos de ocorrência de anemia e na fração de curados em doares de repetição.

Dados de Anemia

Os dados da Fundação Pró-Sangue contém doações realizadas no período de jan/1996 a dez/2006. Foram registrados o dia das doações, a idade (em anos) e o hematócrito (Hct) do doador para cada doação.

Define-se a primeira doação do indivíduo no hemocentro de São Paulo como origem do processo, sendo a ocorrência de anemia tomada como evento de interesse e assume-se também a possibilidade do mesmo evento não ocorrer.

Definem-se como anêmicos os candidatos que apresentarem hematócrito inferior a 39% para homens e a 38% para mulheres. Como as medições de Hct são feitas nos instantes de doação de sangue, tem-se dados com censura intervalar quando a anemia é detectada.

Metodologia

Dois modelos, propostos por Shen e Liu (2009) e Lam et al. (2013), são considerados.

Notação

- $Y_i = \min(T_i, C_i)$, em que T_i é o tempo até o evento de interesse e C_i é o tempo de censura com distribuição independente.
- δ_i é o indicador de censura dado por $\delta_i = I(T_i \leq C_i)$.
- L_i e R_i são os extremos do intervalo de censura que contém T_i
- U_i é a fragilidade com efeito multiplicativo na função de risco.
- K_i é variável latente que define o número de variáveis W a contribuir com a fragilidade.
- $W_h i$ assume distribuição qui-quadrado central com dois graus de liberdade.
- $\boldsymbol{x_i}^{(0)}$ e $\boldsymbol{x_i}^{(1)}$ são vetores de covariáveis assossiadas ao parâmetro de cura e aos parâmetros da regressão de Cox, respectivamente, no modelo de fragilidade.
- $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ são coeficientes de regressão associados à cura e latência (modelo de Cox), respectivamente, no modelo de fragilidade. No modelo de tempo de promoção, adota-se $\boldsymbol{\beta}$ para os coeficientes de cura, conforme o artigo original.
- Z é o vetor de covariáveis associadas à cura no modelo de tempo de promoção.

Modelo de Tempo de Promoção

O modelo apresentado por Shen e Liu (2009) tem seus resultados derivados do modelo de fração de cura proposto por Yakovlev et al. (1993) e tem proposta a seguinte função de sobrevivência

$$S(t|\mathbf{Z}) = \exp(-e^{\alpha+\boldsymbol{\beta}'\boldsymbol{Z}}F(t)), \tag{1}$$

Como em Turnbull (1976), a busca por um estimador de máxima verossimilhança para a função de distribuição F pode restringir-se à classe de equivalência composta de funções escada do tipo:

$$t \mapsto \sum_{j=1}^{m} p_j \mathbf{1}(t \geq r_j),$$

com a restrição $\sum_{j=1}^{m} p_j = 1$.

Shen e Liu (2009) definem um número finito de intervalos disjuntos $\{[s_j, r_j]\}_{j=1}^{m+1}$ construídos como se segue: $s_j \in \{L_i : i = 1, \cdots, n\}$ e $r_j \in \{R_i : i = 1, \cdots, n\}$ de modo que (s_j, r_j) não contenha membro algum de $\{L_i, R_i : i = 1, \cdots, n\}$, e $s_1 \leq r_1 < s_2 \leq r_2 < \cdots s_m \leq r_m < s_{m+1} < r_{m+1} = \infty$. Na construção dos autores, s_{m+1} denota o maior dos tempos de observação.Em seu trabalho, os autores restringem a busca com funções dentro da classe definida por tais intervalos, tomando F contínua pela direita $F(t) = F(t+)$ e assumindo valor constante igual a $F(r_j)$ em $[r_j, s_{j+1}]$ para $j = 1, \cdots, m$, com $F(0) = 0$ e $F(r_m) = F(s_{m+1}) = 1$, com saltos p_j nos instantes imediatamente anteriores a r_j (para $j = 1, \cdots, m$). Deste modo:

$$S_{\boldsymbol{Z_i}}(t) = \exp\left[-e^{\boldsymbol{\theta}'\tilde{\boldsymbol{Z}}_i}\sum_{j=1}^{m} p_j \mathbf{1}(t \geq r_j)\right].$$

em que $p_0 \equiv 0$ por conveniência de notação e $\tilde{\boldsymbol{Z}}_{\boldsymbol{i}} = (1, \boldsymbol{Z_i})$.

Assim, a verossimilhança dos dados depende somente do vetor de parâmetros $\boldsymbol{p}$ e $\boldsymbol{\theta}$. Tomando δ_{ij} como função indicadora assumindo 1 caso $[s_j, r_j]$ pertença a $[L_i, R_i]$, a função de log-verossimilhança pode então ser escrita como

$$l_n(\boldsymbol{\theta}, \boldsymbol{F}) \equiv l_n(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \delta_{ij}(S_{\boldsymbol{Z_i}}(s_j-) - S_{\boldsymbol{Z_i}}(r_j+)) + \delta_{i,m+1} S_{\boldsymbol{Z_i}}(s_{m+1})\right\}. \tag{2}$$

As estimativas para $(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, \hat{F}_n)$ são obtidas a partir da maximização da função de log-verossimilhança acima. A rotina para a maximização consiste em um método iterativo de obtenção da log-verossimilhança esperada seguido pela maximização da log-verossimilhança esperada perfilada para $(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, \hat{F}_n)$, descrevendo assim essencialmente o algoritmo ECM (Meng e Rubin, 1993). Entretanto, as restrições de $\boldsymbol{p}$ e sua possivelmente alta dimensionalidade motivam o uso de técnicas de maximização convexa para a obtenção da estimativa deste vetor, estas encontram-se melhor detalhadas em Shen e Liu (2009) e Boyd e Vandenberghe (2004).

Modelo de Fragilidade

Lam et al. (2013) adaptam a ideia de (Aalen,1992) de que a fração de cura populacional pode ser modelada através de um efeito aleatório com massa no ponto zero. Efeitos de covariáveis são contemplados, influenciando a distribuição da fragilidade e os tempos de latência por conta da regressão de Cox. Tem-se então:

$$S(t|\boldsymbol{x_i}^{(0)}, \boldsymbol{x_i}^{(1)}) = \exp\left[-\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x_i}^{(0)})}{2}\left\{1 - \frac{1}{1 + 2\Lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x_i}^{(1)})}\right\}\right],$$

$$\lambda(t|u_i, \boldsymbol{x_i}^{(1)}) = u_i\lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x_i}^{(1)}),$$

Tal que:

$$K_i|\boldsymbol{x_i}^{(0)} \sim Poisson\left(\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x_i}^{(0)})}{2}\right), \qquad U_i = \begin{cases} 0 & \text{se } k_i = 0; \\ W_{1i} + W_{2i} + \cdots + W_{k_i i} & \text{se } k_i > 0 \end{cases}$$

A estimação de $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ é obtida diretamente através da maximização da seguinte verossimilhança parcial L_C com dados completos dada por

$$L_C(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{D}) = \prod_{i=1}^{n} \frac{\exp\left\{\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x_i}^{(0)})}{2}\right\}\left\{\frac{\exp(\boldsymbol{\theta}'\boldsymbol{x_i}^{(0)})}{2}\right\}^{k_i}}{k_i!} \prod_{k_i>0} \frac{u_i^{k_i-1}\exp(\frac{-u_i}{2})}{2^{k_i}\Gamma(k_i)} \prod_{i=1}^{n}\left\{\frac{u_i\exp(\boldsymbol{\beta}'\boldsymbol{x_i}^{(1)})}{\sum_{m\in R(y_i)} u_m\exp(\boldsymbol{\beta}'\boldsymbol{x_m}^{(1)})}\right\}^{\delta_i}, \tag{3}$$

em que $R(y_i)$ é o conjunto de indivíduos em risco até o instante y_i. A função de risco acumulada basal $\Lambda_0(t)$ pode ser estimada com uma adaptação do estimador de Nelson-Aalen apresentada no artigo dos autores.

Entretanto, as quantidades K_i e U_i não são observáveis. Os autores utilizam então o método da imputação múltipla para estimar $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$. A ideia consiste em imputar as variáveis latentes utilizando posterioris das mesmas. Geram-se M réplicas do banco de dados com dados imputados e tem-se a estimativa definida como a média dos estimadores de máxima verossimilhança condicional, constituindo então uma iteração do processo.

Além disso, as expressões apresentadas acima dependem de tempos observados ou censurados à direita, sendo necessária a geração dos mesmos a partir da distribuição condicional dados os intervalos de censura. O algoritmo é apresentado em maiores detalhes em Lam et al. (2013) e seus fundamentos são encontrados em Tanner e Wong (1987a).

Resultados

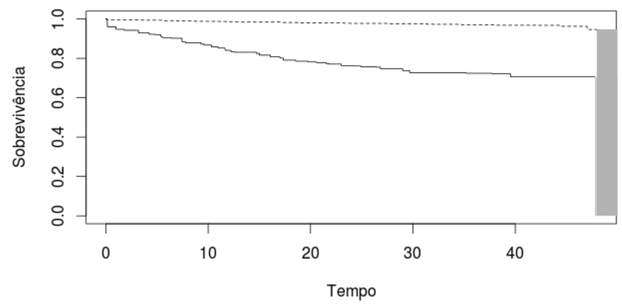

Figura 1: Estimativas da Função de Sobrevivência com homens e mulheres representados pelas linhas pontilhada e contínua, respectivamente

Sexo	β_0 **(Intercepto)**	β_1 **(Idade)**	β_2 **(Htc)**
H	11.070 (0.843)	0.021 (0.004)	-0.339 (0.019)
M	8.891 (0.454)	-0.019 (0.002)	-0.227 (0.011)

Tabela 1: Estimativas e Desvios Padrões para Modelo de Tempo de Promoção

Os modelos foram aplicados em dados considerando doações realizadas a partir de janeiro de 2003 contendo somente doadores de repetição (doações semestrais). Das tabelas, observa-se que os efeitos estimados das covariáveis tem efeitos nos mesmos sentidos para os modelos dos diferentes autores; O avanço da idade acarreta o aumento da probabilidade de incidência de anemia para os homens, entretanto, apresenta efeito contrário na probabilidade de incidência feminina, condizendo com o contexto biológico do período da menopausa; Embora a incidência aumente, a idade tem efeito negativo no modelo de Lam et al. (2013) sobre os tempos latentes, implicando que o risco diminui com o aumento da idade, o que pode ser interpretado como um aumento da probabilidade de ser suscetível à anemia, com os indivíduos suscetíveis tendo uma maior sobrevida. Para ambos modelos propostos, a diminuição do hematócrito implica no aumento da incidência de anemia em todos os cenários.

Sexo	θ_0 **(Intercepto)**	θ_1 **(Idade)**	θ_2 **(Htc)**	β_1 **(Idade)**	β_2 **(Htc)**
H	9.592 (1.329)	0.028 (0.007)	-0.315 (0.030)	-0.142 (0.029)	-0.761 (0.057)
M	7.843 (0.630)	-0.010 (0.003)	-0.203 (0.015)	-0.025 (0.007)	-0.052 (0.022)

Tabela 2: Estimativas e Erros Padrões para Modelo de Fragilidade

Para ilustrar os efeitos sobre a proporção de cura, fixa-se o gênero masculino, idade de 30 anos e hematócrito de 40%, temos 62.6% e 75.5% de proporção de curados para os modelos de tempo de promoção e fragilidade, respectivamente.

Conclusões

- Para ambos os modelos, o efeito da idade sobre a incidência nos homens tem natureza contrária ao das mulheres.
- Os resultados alimentam a evidência da existência de uma fração imune à anemia por doações.
- A especificação incorreta do modelo de fração de cura tem como consequência frações de cura estimadas discrepantes. Um estudo exaustivo de simulações com o intuito de avaliar a robustez dos modelos está sendo realizado e será apresentado futuramente.
- Uma complicação destes dados é a presença de censuras informativas: a distribuição do tempo altera-se após uma doação, demandando o estudo de alternativas como o uso de covariáveis dependentes do tempo, por exemplo.

Referências

Boyd, S. e Vandenberghe, L. (2004). *Convex Optimization.*, Cambridge University Press, Cambridge, UK.

Lam, K., Wong, K. e Zhou, F. (2013). A semiparametric cure model for interval-censored data, *Biometrical Journal* **55(5)**: 771–788.

Meng, X. e Rubin, D. (1993). Maximum likelihood estimation via the ecm algorithm: a general framework., *Biometrika* **80(2)**: 267–278.

Shen, Y. e Liu, H. (2009). A semiparametric regression cure model for interval-censored data, *Journal of the American Statistical Association* **104(487)**: 1168–1178.

Tanner, M. e Wong, W. (1987a). The calculation of posterior distributions by data augmentation., *Journal of the American Statistical Association* **82**: 528–540.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society. Series B (Methodological)* **38(3)**: 290–295.