Análise bayesiana de um modelo de regressão semi–paramétrico com dados censurados: estimação e diagnóstico.

Alessandra C. Sibim [1], **Vicente G. Cancho**[1]

[1] Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Caixa Postal 668, 13560-970 São Carlos - SP

Resumo: Neste trabalho desenvolvemos um procedimento Bayesiano baseado em métodos de Monte Carlo via cadeias de Markov (MCMC) para um modelo de regressão semi–paramétrico com dados censurados. Além disso, são consideradas técnicas de diagnóstico baseado na divergência de Kullback-Leibler. A metodologia é ilustrada com um conjunto de dados reais.

Palavras-chave: inferência Bayesiana, diagnóstico, divergência de Kullback-Leibler, métodos MCMC, análise de sobrevivência.

1 Introdução

O modelo exponencial por partes é um dos modelos mais populares utilizados em análise de sobrevivência (Ibrahim *et al.*, 2001), a popularidade se deve ao fato deste modelo ser capaz de acomodar funções de taxa de falha com diversas formas, tornando o modelo bastante flexível. Uma das vantagens do modelo é a possibilidade de se trabalhar tanto na versão paramétrica, quanto na versão não-paramétrica. O modelo exponencial por partes não paramétrico pode ser obtido como caso particular do modelo exponencial por partes paramétrico, em que se tem um único tempo de falha associado a cada intervalo.

Uma maneira de avaliar as suposições feitas sobre o modelo e de detectar pontos influentes pode ser efetuada por meio de alguns métodos de diagnóstico. A aplicação destes, constitui uma etapa importante na análise do ajuste do modelo de regressão, verificando a existência de observações extremas com alguma interferência desproporcional nos resultados do ajuste ou existência de afastamento das hipóteses postuladas para o modelo, em especial para a parte aleatória e a parte sistemática. Uma das propostas mais inovadoras nesta área foi apresentada por Cook (1986) que propôs avaliar a influência conjunta das observações sob pequenas perturbações no modelo, ao invés da avaliação pela retirada individual ou conjunta de pontos. Recentemente, Cho *et al.* (2009), propuseram um método bayesiano de análise de influência caso a caso, baseado na divergência Kullback-Leibler (K-L).

Neste trabalho desenvolvemos uma análise bayesiana para um modelo de regressão semi–paramétrico (ou modelo de regressão exponencial por partes) e considerando a proposta Cho *et al.* (2009) um estudo de robustez do modelo é realizado.

2 Modelo de regressão semi–paramétrico

Considere uma partição finita e arbitrária do eixo dos tempos $\{s_1, \ldots, s_j\}$, tal que, $s_0 < s_1 < \ldots < s_j < \infty$, com $s_0 = 0$ e $s_j > t$, para algum t observado, com $t > 0$, e admita que tal partição divida

[1]Contatos: *alesibim@icmc.usp.br, garibay@icmc.usp.br*

o eixo do tempo em J intervalos disjuntos, denotados por $I_1 = (s_0, s_1], I_2 = (s_1, s_2], \ldots, I_j = (s_{j-1}, s_j]$, assume-se que em cada intervalo $I_j = (s_{j-1}, s_j]$, $j = 1, \ldots, J$ a função taxa de falha é constante com $h(t) = \lambda_j \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}$, para $t \in I_j$, $\boldsymbol{x}^\top = (x_1, \ldots, x_p)$ é um vetor de covariáveis e $\boldsymbol{\beta}$ seus respectivos coeficientes de regressão (Ibrahim *et al.*, 2001). Daí podemos demonstrar que a função de sobrevivência do modelo é dada por

$$S(t; \boldsymbol{\lambda}, \boldsymbol{\beta}) = \begin{cases} \exp\{-\lambda_1 t \ \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\}, \text{se } t \in I_1; \\ \exp\left\{ - \left[\sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) + \lambda_j (t - s_{j-1}) \right] \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\} \right\}, \text{se } t \in I_j, j > 1 \end{cases} \quad (1)$$

$\forall j = 1, \ldots, J$. De aqui diante o modelo proposto em (1) será denominado de modelo de regressão semi-paramétrico (MRSP). Uma das dificuldades em se trabalhar com o MRSP está em definir a partição $\{s_1, ..., s_j\}$ a ser utilizada. Em geral, tal partição é escolhida arbitrariamente como discutido por Demarqui *et al.* (2008), Barbosa *et al.* (1996) e Gamerman (1994).

3 Análise bayesiana

Denotamos por $\boldsymbol{D} = (n, \mathbf{t}, \boldsymbol{X}, \nu)$ os dados observados, com $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^\top$, $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)^\top$, com $\nu_i = 1$ se o i-ésimo indivíduo for falha e 0 caso contrário. $\boldsymbol{X}$ é uma matriz $n \times p$ de covariáveis. Assumindo $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_j)^\top$, podemos escrever a função de verossimilhança de $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ para n indivíduos por

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D}) \quad = \quad & \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_j \exp(x_i^\top \boldsymbol{\beta}))^{\delta_{ij} \nu_i} \exp\Bigg\{ - \delta_{ij} \Bigg[\lambda_j (t_i - s_{j-1}) \\ & + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1}) \Bigg] \exp(x_i^\top \boldsymbol{\beta}) \Bigg\}. \end{aligned} \quad (2)$$

Assumindo independência entres os parâmetros do MRSP a densidade a *priori* conjunta é dada por $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\lambda})$ em que, $\boldsymbol{\beta_j} \sim N(\mu_{\beta_j}, \sigma^2_{\beta_j})$, $j = 1, \ldots, p$ e $\lambda_k \sim \pi(\lambda_k)$, $k = 1, \ldots, J$. Combinando essa densidade a priori com a verossimilhaça (2) , a densidade a *posteriori* conjunta é dada por $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda} | \boldsymbol{D}) = L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\lambda})$. Como que a densidade a posteriori conjunta não é uma densidade padrão usamos métodos de Monte Carlo via cadeias de Markov (MCMC), tais como o amostrador de Gibbs e Metropolis-Hasting. Podemos mostrar que as densidades a p *posteriori* condicionais para o amostrador de Gibbs são expressas por $\pi(\boldsymbol{\beta} | \boldsymbol{\lambda}, \boldsymbol{D}) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D}) \pi(\boldsymbol{\beta}),$, $\pi(\boldsymbol{\lambda} | \boldsymbol{\beta}, \boldsymbol{D}) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D}) \pi(\boldsymbol{\lambda})$. Como as condicionais não possuem forma fechada usaremos o algoritmo de Metropolis-Hastings dentro do ciclo do algoritmo de Gibbs (Gilks *et al.*, 1996) para gerar amostras de $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$.

3.1 Comparação de modelos

A estatística condicional preditiva ordinária (CPO) é uma ferramenta de avaliação do modelo muito útil e extensamente usada na literatura estatística sob vários contextos (Ibrahim *et al.*, 2001). Seja $\boldsymbol{D}$ os dados completos e $\boldsymbol{D^{(-i)}}$ os dados com a i-ésima observação excluída. Denotamos a densidade a *posteriori* de $\boldsymbol{\gamma}$ dado $\boldsymbol{D^{(-i)}}$ por $\pi(\boldsymbol{\gamma} | \boldsymbol{D^{(-i)}})$, $i = 1, \ldots, n$. Podemos escrever a CPO_i para a i-ésima observação por

$$CPO_i = \int_\Theta g(y_i | \boldsymbol{\gamma}) \pi(\boldsymbol{\gamma} | \boldsymbol{D^{(-i)}}) d\boldsymbol{\gamma} = \left\{ \int_\Theta \frac{\pi(\boldsymbol{\gamma} | \boldsymbol{D})}{g(y_i | \boldsymbol{\gamma})} d\boldsymbol{\gamma} \right\}^{-1}.$$

Para valores altos de CPO_i temos um melhor ajuste do modelo. Uma estimativa de Monte Carlo para CPO_i considerando uma amostra da distribuição a *posteriori* $\pi(\boldsymbol{\gamma} | \boldsymbol{D})$, pode ser obtida se $\boldsymbol{\gamma_1}, \ldots, \boldsymbol{\gamma_Q}$ é

uma amostra de tamanho Q de $\pi(\boldsymbol{\gamma}|\boldsymbol{D})$. A estimativa de Monte Carlo para CPO_i (Chen *et al.*, 2000) é dada por

$$\widehat{CPO_i} = \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{g(y_i|\boldsymbol{\gamma}_q)} \right\}^{-1}. \tag{3}$$

Como em Cancho *et al.* (2010) utilizamos a estatística $B = \sum_{i=1}^{n} \log(\widehat{CPO_i})$ na seleção dos modelos, maiores valores de B indicam o melhor modelo. Um outro critério utilizado neste trabalho é o "Deviance Informartion Criterion "(DIC), menores valores de DIC determinam os melhores modelos.

3.2 Análise de influência caso a caso

Uma maneira comum de avaliar a influência de uma observação no ajuste de um modelo é por meio da deleção de casos (Cook & Weisberg, 1982). Recentemente Cho *et al.* (2009) propuseram uma análise de influência caso a caso baseado na divergência de Kullback–Leibler (K–L).

Suponha que $K(P, P_{-i})$ denota a divergência K-L entre P e P_{-i}, onde P denota a distribuição a *posteriori* de γ para os dados completos e P_{-i} é a distribuição a *posteriori* de γ sem o i-ésimo caso. Especificamente,

$$K(P, P_{-i}) = \int p(\gamma|\mathbf{D}) \log \left\{ \frac{\mathbf{p}(\gamma|\mathbf{D})}{\mathbf{p}(\gamma|\mathbf{D_{-i}})} \right\} d\gamma, \tag{4}$$

$K(P, P_{-i})$ mede o efeito de omitir o i-ésimo caso dos dados completos na distribuição a *posteriori* de γ. Note que $K(P, P_{-i}) \neq K(P_{-i}, P)$ em geral. Após alguma álgebra, podemos demonstrar uma expressão simplificada para $K(P, P_{-i})$ é dada por,

$$K(P, P_{-i}) = \log E_\gamma \left[\frac{L(\gamma|\mathbf{D_{-i}})}{L(\gamma|\mathbf{D})} \Big| D \right] + E_\gamma \left[\log \left\{ \frac{L(\gamma|\mathbf{D})}{L(\gamma|\mathbf{D_{-i}})} \right\} \Big| D \right], \tag{5}$$

onde $E[\cdot|D]$ representa a média a *posteriori* de γ. A equação (5) nos permite computar $K(P, P_{-i})$, para $i = 1, \ldots, n$, usando estimativas de Monte Carlo.

Segundo McCulloch (1989) e Cho *et al.* (2009) a calibração de $K(P, P_{-i})$, pode ser feita resolvendo em p_i a equação

$$K(P, P_{-i}) = K(B(0,5), B(p_i)) = -\frac{\log\{4p_i(1-p_i)\}}{2}, \tag{6}$$

onde $B(p)$ denota uma distribuição de Bernoulli com probabilidade de sucesso p. Isto implica que descrever resultados usando $p(B|D_{-i})$ ao invés de $p(B|D)$ é equivalente com a descrição de um evento não observado com probabilidade p_i quando a probabilidade correta é 0,5. Após o cálculo de $K(P, P_{-i})$ em (5) a solução em p_i da equação (6) é dada por $p_i = 0,5[1 + \sqrt{1 - \exp\{-2K(P, P_{-i})\}}]$. Isto implica que $0,5 \leq p_i \leq 1$. Além disso, se $p_i \gg 0,5$ implica que o i-ésimo caso é influente.

4 Aplicação e Discussão

Os dados considerados neste trabalho referem-se ao estudo descrito em Klein & Moeschberger (2003), os quais relacionam os tempos de reincidência de infecção renal em 38 pacientes. Consideramos a presença de cinco covariáveis: x_1, idade do paciente (em anos); x_2, gênero do paciente (0-masculino, 1-feminino); x_3, x_4 e x_5 indicam o tipo da doença apresentada pelo paciente, sendo GN, AN e PKD respectivamente. Em nosso estudo consideramos somente o primeiro tempo observado.

Pelo método gráfico baseado no tempo total em teste (TTT) transformado, descrito por Barlow & Campo (1975) identificamos a forma da função de risco com o objetivo de avaliar se o modelo é adequado

aos tempos de vida. Em nosso caso, temos indício de que os dados apresentam a função de risco não monótona, como observado na Figura 1, assim ajustamos MRSP como descrito na seção 2, considerando $J = 3, 4, 5$ e 6 com as cinco covariáveis descritas acima.

Para nossa análise Bayesiana consideramos densidades a *prioris* independentes para os parâmetros do MRSP, com $\beta_j \sim N(0, 100)$, j=1,... 5, e $\lambda_k^* = \exp(\lambda) \sim N(0, 100)$, k=1,... J. Considerando essas densidades a priori geramos amostras de Gibbs através do algoritmo de Gibbs com Metropolis-Hasting da seguinte forma: Geramos duas cadeias paralelas cada uma com 35.000 iterações com um *burn in* de 5.000 e saltos de tamanho 10, resultando uma amostra de Gibbs de tamanho 6.000. Para monitorar a convergência do amostrador de Gibbs utilizamos a aproximação desenvolvida por Gelman e Rubin(1992). A Tabela 1 apresenta os valores para os critérios de seleção B e DIC para os modelos ajustados. Ambos critérios indicam um MRSP com $J = 4$ intervalos disjuntos na partição do eixo dos tempos.

Tabela 1: Estimativas para os modelos.

Critério	Modelo			
	$J = 3$	$J = 4$	$J = 5$	$J = 6$
DIC	320,3	316,3	322,4	318,4
B	-162,5	-160,3	-164,5	-162,1

Na tabela 2 são apresentados os resumos da distribuição a *posteriori* do MRSP para o melhor modelo, onde observamos que um 95% de confiança que somente as covariáveis sexo e PKD são significativas.

Tabela 2: Estimativas a *posteriori* para os tempos.

Parâmetro	Média	Desvio Padrão	2,50%	97,50%
β_1(idade)	0,0038	0,0160	-0,0272	0,0355
β_2(sexo)	-1,0702	0,4796	-1,9635	-0,0888
β_3(GN)	0,3007	0,5903	-0,8650	1,4446
β_4(AN)	0,7823	0,6182	-0,3926	2,0229
$\beta_5(PKD)$	-2,0382	0,9989	-4,1766	-0,2405
λ_1^*	-4,5998	0,7977	-6,2344	-3,1163
λ_2^*	-3,5512	0,8227	-5,2232	-2,0339
λ_3^*	-5,0047	0,8488	-6,7316	-3,4225
λ_4^*	-3,9977	0,7500	-5,5658	-2,6048

Com as amostras de Gibbs foram estimadas as medidas de divergência de K–L para cada uma das observações esses resultados são graficados na Figura 2, onde as observações 10, 15, 21 e 33 apresentam maiores valores quando comparados com as demais observações, para verificar se essas observações são influentes estimamos a calibração da medida de divergência K–L, essas estimativas são apresentados na Tabela 3 conjuntamente com as respectivas estimativas da divergência K–L, indicado que as observações 10, 15 e 21 são influentes.

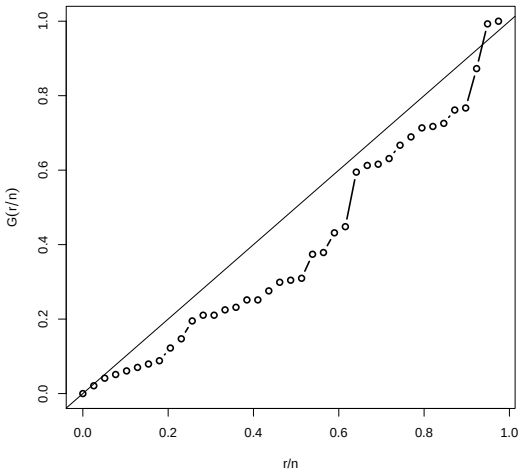

Figura 1: Gráfico TTT plot para dados de infecção renal.

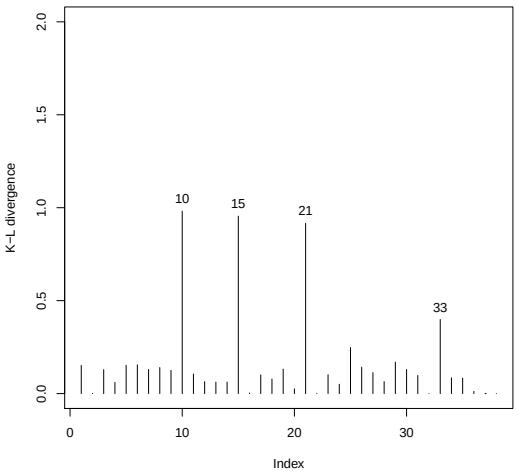

Figura 2: Gráfico de índices de $K(P, P_{-i})$ para infecção renal.

Tabela 3: Identificação dos casos influentes para dados de infecção renal.

Identificação do caso			Influência caso a caso	
Paciente	Tempo	Idade	$K(P, P_{-i})$	Calibração
10	154	51,5	0,981227	0,963543
15	536	17	0,954761	0,961478
21	562	46,5	0,916202	0,958250
33	152	57	0,398699	0,870641

Agradecimentos

Os autores agradecem a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro.

Referências

Barbosa, P. E., Colosimo, E. A. & Louzada-Neto, F. (1996). Accelerated life tests analyzed by a piecewise exponential distribution via generalized linear models. *IEEE Transactions on Reliability*, **45**, 619–623.

Barlow, R. & Campo, R. (1975). Total Time on Test Processes and Applications to Failure Data Analysis.

Cancho, V., Ortega, E. & Paula, G. (2010). On estimation and influence diagnostics for log-Birnbaum-Saunders Student-t regression models: Full Bayesian analysis. *Journal of Statistical Planning and Inference*.

Chen, M. H., Shao, Q. M. & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. Springer-Verlag, New York.

Cho, H., Ibrahim, J. G., Sinha, D. & Shu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, **65**, 116–124.

Cook, R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, **48**, 133–169.

Cook, R. & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall New York.

Demarqui, F., Loschi, R. & Colosimo, E. (2008). Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis*, **14**, 333–356.

Gamerman, D. (1994). Bayes estimation of the piece-wise exponential distribution. *IEEE Transactions on Reliability*, **43**, 128–131.

Gilks, W., Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Ibrahim, J., Chen, M. & Sinha, D. (2001). *Bayesian survival analysis*. Springer Verlag.

Klein, J. & Moeschberger, M. (2003). *Survival analysis: techniques for censored and truncated data*. Springer Verlag.