

Modelagem Conjunta de Dados Longitudinais e de Sobrevivência

Juliana Freitas de Mello e Silva

Orientador: Vinícius Diniz Mayrink

Co-orientador: Fábio Nogueira Demarqui

Departamento de Estatística
Universidade Federal de Minas Gerais

27 de outubro de 2017

Conteúdo

- 1 Introdução
- 2 Literatura
- 3 Polinômios de Bernstein
- 4 Proposta de tese
- 5 Referências

Outline

1 Introdução

Conceitos básicos

Alguns conceitos fundamentais em análise de sobrevivência são:

- é utilizada quando se deseja estimar o tempo até a ocorrência de certo evento de interesse (T);
- contém características intrínsecas como falhas/censuras e assimetria;
 - ▶ há três tipos de censura: **à direita**, à esquerda e intervalar;
 - ▶ mecanismo causador da censura: informativo ou **não informativo**;

$$\delta_i = \begin{cases} 1, & \text{se o } i\text{-ésimo indivíduo falhou} \\ 0, & \text{se o } i\text{-ésimo indivíduo é censurado (à direita)} \end{cases}$$

- para cada indivíduo se observa $(t_i, \delta_i, \mathbf{x}_i)$, para $i = 1, 2, \dots, n$.

Funções básicas

Em análise de sobrevivência, considera-se as seguintes funções que são relacionadas entre si:

- função de sobrevivência

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F(t) ;$$

- função risco

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} ;$$

- função de risco acumulado

$$H(t) = \int_0^t h(u) du .$$

Função de verossimilhança

No caso de modelagem paramétrica, censura à direita e não informativa, a função de verossimilhança é dada por:

$$L(\Phi; D) = \prod_{i=1}^n f(t_i|\Phi)^{\delta_i} (S(t_i|\Phi))^{1-\delta_i} = \prod_{i=1}^n h(t_i|\Phi)^{\delta_i} \exp\{-H(t_i|\Phi)\}.$$

Em que:

- Φ é o vetor de parâmetros a ser estimado, que inclui o vetor de coeficientes;
- D representa os dados disponíveis;
- n é o número de indivíduos;
- t_i é o tempo observado do i -ésimo indivíduo;
- δ_i é a indicadora de falha/censura.

Modelo de riscos proporcionais

Na presença de covariáveis, pode-se utilizar o modelo proposto por Cox (1972):

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}\boldsymbol{\beta}),$$

no qual $h_0(t)$ é a função risco de base, \mathbf{x} é o vetor de covariáveis e $\boldsymbol{\beta}$ é o vetor de coeficientes.

No caso de variáveis dependentes do tempo, pode-se usar uma extensão do modelo acima:

$$h(t|\mathbf{x}(t)) = h_0(t) \exp(\mathbf{x}(t)\boldsymbol{\beta}),$$

em que $\mathbf{x}(t)$ é tal que apresenta mudanças ao longo do tempo.

Dados longitudinais e de sobrevivência

- Dados longitudinais são caracterizados, basicamente, por variáveis com medidas repetidas ao longo do tempo;
- De forma geral, o interesse desses estudos é observar o tempo até um determinado evento e uma importante variável com medidas repetidas para cada indivíduo ao longo do tempo;
- De acordo com Ibrahim et al. (2010), a modelagem conjunta de dados longitudinais e de sobrevivência foi inicialmente motivada por estudos sobre a AIDS;
- Essa modelagem foi proposta baseada na idéia de que existe uma correlação significativa entre essas duas variáveis (longitudinal e tempo até evento), que não seria levada em consideração caso a modelagem fosse feita de forma independente.

Motivação: por que acompanhar dados longitudinalmente?

Acompanhar indivíduos ao longo do tempo tem particularidades como:

- informação sobre a evolução de cada indivíduo;
- permite comparações de mesma natureza, eliminando possíveis efeitos de confusão;
- *independência entre* indivíduos e *dependência intra*-indivíduo;
- dados faltantes (em muitos casos);
- erro de medição.

Em relação à correlação intra-indivíduos, Fitzmaurice et al. (2004) afirmam que:

- geralmente as correlações são positivas;
- a correlação diminui com o espaçamento do tempo;
- mesmo com um grande espaçamento, dificilmente ficam próximas de zero;

Exemplos

Como exemplos de dados longitudinais e de sobrevivência, tem-se:

- estudos com pacientes soropositivos
há interesse em analisar o tempo até a ocorrência de um evento, como a progressão para AIDS ou óbito, além da variação do número de células CD4 ao longo do tempo;
- estudos envolvendo pacientes com câncer
nesse contexto, o evento de interesse frequentemente é definido como óbito ou recidiva. Uma variável longitudinal que eventualmente é medida é a qualidade de vida desses pacientes (índice).

Outline

2 Literatura

No início, ao tentar analisar dados longitudinais e de sobrevivência conjuntamente se fez uso, principalmente, de duas abordagens (Ibrahim et al., 2010):

- 1 considerou-se as variáveis longitudinais como variáveis tempo-dependentes e utilizava-se o modelo de Cox adequado para esse caso;
- 2 modelo de dois estágios, proposto por Tsiatis et al. (1995), no qual se modela a variável longitudinal através de uma função trajetória e, em seguida, essa estimativa é imputada no modelo de Cox como variável tempo-dependente.

Considerações sobre os modelos

Sobre a primeira abordagem:

- dados longitudinais geralmente apresentam erro;
- utilizar essa informação diretamente no modelo de Cox pode levar à estimativas viciadas (Ibrahim et al., 2010);
- essa característica é tratada no segundo modelo.

Em relação ao modelo de dois estágios:

- a estimativa da componente longitudinal não diferencia ocorrência de censura ou evento de interesse (Wu et al., 2012);
- a incerteza associada à estimativa da componente longitudinal não é levada em consideração no segundo estágio (Wu et al., 2012).

Modelagem conjunta

Alguns aspectos básicos da modelagem conjunta são:

- fornece estimativas mais precisas e robustas ao utilizar toda informação disponível simultaneamente;
- considera-se principalmente duas componentes, a componente longitudinal e a componente de sobrevivência (Ibrahim et al., 2010);
- a modelagem conjunta trata essas duas componentes como dois submodelos interligados (um para o processo longitudinal “real” e outro para a sobrevivência) (Tsiatis & Davidian, 2004).

Componente longitudinal

Atribui-se um modelo para o processo longitudinal com o intuito de representar a trajetória “real” (não necessariamente observada) dessa componente.

Por exemplo:

$$X_i(u) = \alpha_{0i} + \alpha_{1i}u, \quad i = 1, 2, \dots, n. \quad (1)$$

- $X_i(u)$ representa a variável longitudinal “real” no tempo u ;
- $\alpha_i = (\alpha_{0i}, \alpha_{1i})^T$ é vetor de efeitos aleatórios a nível de indivíduo para o qual comumente se atribui a distribuição Normal.

Componente longitudinal

Uma forma mais geral e bastante utilizada (Tsiatis & Davidian, 2004) é a seguinte:

$$X_i(u) = f(u)^T \boldsymbol{\alpha}_i, \quad i = 1, 2, \dots, n. \quad (2)$$

Nesse caso, $f(u)$ é um vetor de funções do tempo u e $\boldsymbol{\alpha}_i$ é um vetor efeitos aleatórios. Esta forma inclui:

- especificação polinomial;
- *splines*;
- funções não-lineares (em u).

Componente longitudinal

Caso haja intuito de se aproximar ainda mais do processo biológico real, a seguinte forma pode ser adotada:

$$X_i(u) = f(u)^T \boldsymbol{\alpha}_i + U_i(u), \quad i = 1, 2, \dots, n. \quad (3)$$

Observações:

- o termo adicional $U_i(u)$ é um processo estocástico com média zero, (geralmente) independente de $\boldsymbol{\alpha}_i$ e de covariáveis.
- a forma em (3) permite mudanças bruscas no comportamento da componente longitudinal;
- escolhe-se forma da componente longitudinal de acordo com o interesse do estudo, considerando a fidelidade ao processo biológico e o conceito de parcimônia.

Componente longitudinal

Dada uma especificação para a função trajetória $X_i(u)$, os dados longitudinais observados são dados por:

$$\begin{aligned} W_i(t_{ij}) &= X_i(t_{ij}) + e_i(t_{ij}), & i &= 1, 2, \dots, n; \\ & & j &= 1, 2, \dots, m_i. \end{aligned} \quad (4)$$

Em que:

- $W_i(t_{ij})$ é a observação da variável longitudinal do i -ésimo indivíduo no tempo t_{ij} ;
- $e_i \sim Normal(0, \sigma^2)$;
- e_i é independente de α_i (e de $U_i(u)$);
- pode-se assumir estrutura de correlação em e_i .

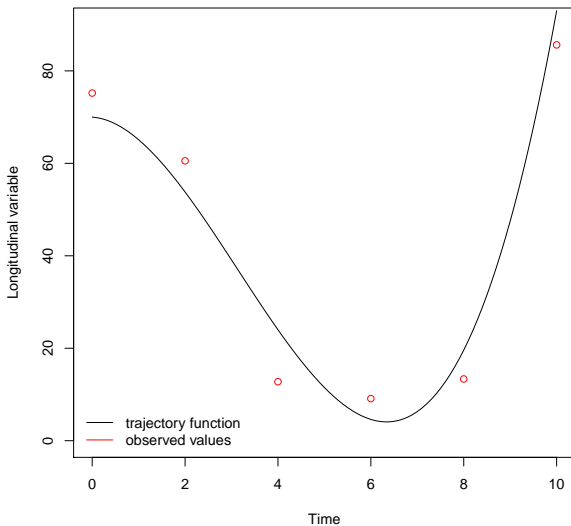


Figura : Representação da função trajetória e dos valores observados.

Componente longitudinal

Segundo Tsiatis & Davidian (2004):

- as formas mais utilizadas para a componente longitudinal são a (2) e a (3);
- se as medidas forem coletadas em grandes espaçamentos no tempo (t_{ij}), é razoável assumir independência entre $e_i(t_{ij})$.
- contudo, se as observações são próximas no tempo de forma que a relação entre elas não pode ser ignorada, deve-se incluir estrutura de correlação.

Componente de sobrevivência

O seguinte modelo é atribuído à função risco:

$$\begin{aligned}h_i(u) &= \lim_{du \rightarrow 0} \frac{\mathbb{P}(u \leq T_i < u + du | T_i \geq u, X_i^H(u), Z_i)}{du} \\&= h_0(u) \exp \{ \gamma X_i(u) + \eta^T Z_i \}, \quad i = 1, 2, \dots, n.\end{aligned}\tag{5}$$

Em que:

- $X_i^H(u) = \{X_i(t), 0 \leq t < u\}$ é o histórico do processo longitudinal até o tempo u ;
- Z_i vetor de covariáveis medidas no tempo inicial;
- $h_0(u)$ é a função risco de base;
- γ mede a relação entre a variável longitudinal ($X_i(u)$) e o tempo até o evento;
- η vetor de coeficientes a serem estimados.

Função de verossimilhança conjunta

Sejam X e Y duas variáveis aleatórias contínuas. Sabe-se que a função densidade conjunta por ser escrita da forma: $f_{(X,Y)}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$.

A partir disso, a função de verossimilhança no caso de censura à direita e não informativa é dada por:

$$L(\Phi; D) = \prod_{i=1}^n \int \left\{ h(u_i)^{\delta_i} \exp \{-H(u_i)\} \left(\prod_{j=1}^{m_i} p(W_i(t_{ij}) | (\alpha_i, \sigma^2)) \right) p(\alpha_i | \cdot) \right\} d\alpha_i$$

(supondo normalidade dos erros)

$$\begin{aligned} &= \prod_{i=1}^n \int \left[h_0(u_i) \exp \{ \gamma X_i(u_i) + \eta^T Z_i \} \right]^{\delta_i} \\ &\quad \exp \left\{ - \int_0^{u_i} h_0 \exp \{ \gamma X_i(w) + \eta^T Z_i \} dw \right\} \\ &\quad \frac{1}{(2\pi\sigma^2)^{m_i/2}} \exp \left\{ - \sum_{j=1}^{m_i} \frac{\{W_i(t_{ij}) - X_i(t_{ij})\}^2}{2\sigma^2} \right\} p(\alpha_i | \cdot) d\alpha_i \end{aligned}$$

Extensões na literatura

Há, na literatura, vários outros tipos de modelagem e extensões, como:

- modelagem paramétrica para a componente de sobrevivência;
- modelos que relaxam algumas das suposições (normalidade de α_i);
- estruturas de dependência;
- inclusão de características como fração de curados.

A seguir, alguns exemplos.

Exemplo I

Wulfsohn & Tsiatis (1997):

- tiveram por objetivo principal verificar a relação da contagem de células CD4 com a sobrevivência de pacientes HIV positivos;
- utilizaram abordagem frequentista;
- a função risco de base foi modelada utilizando o método de máxima verossimilhança não paramétrica e os demais parâmetros por máxima verossimilhança paramétrica;
- a estimação foi realizada via EM.

Wulfsohn & Tsiatis (1997)

Neste trabalho, os autores fizeram a modelagem na estrutura:

- componente longitudinal

$$\begin{aligned}X_i(u) &= \alpha_{0i} + \alpha_{1i}u \\W_i(t_{ij}) &= \textcolor{red}{X_i(t_{ij})} + e_i(t_{ij}) = \textcolor{red}{\alpha_{0i} + \alpha_{1i}t_{ij}} + e_i(t_{ij}),\end{aligned}$$

sendo $e_{ij} \sim \text{Normal}(0, \sigma_e^2)$, com $\text{Cov}(e_{ij}, e_{ij'}) = 0$ (para $j \neq j'$) e $\boldsymbol{\alpha}_i \sim \text{Normal}_2(\boldsymbol{\alpha}, \Sigma)$ (e_i independente de $\boldsymbol{\alpha}_i$);

- componente de sobrevivência

$$\begin{aligned}h(u|\boldsymbol{\alpha}_i, \mathbf{W}_i, \mathbf{t}_i) &= h(u|\boldsymbol{\alpha}_i) \\&= h_0(u) \exp\{\gamma X_i(u)\} = h_0(u) \exp\{\gamma(\alpha_{0i} + \alpha_{1i}u)\}.\end{aligned}$$

Exemplo II

Brown & Ibrahim (2003):

- estudaram o efeito de vacinas contra câncer no tempo até a recidiva;
- o objetivo principal desse trabalho era flexibilizar a modelagem da componente longitudinal, para isso fizeram uso do processo Dirichlet;
- a forma da função trajetória foi especificada para permitir um crescimento no início, seguido de queda;
- a função risco de base foi modelada com a distribuição Exponencial por Partes;
- esse trabalho utilizou Inferência Bayesiana para estimação dos parâmetros.

Brown & Ibrahim (2003)

A estrutura utilizada neste trabalho foi:

- componente longitudinal

$$\begin{aligned}X_i(u) &= \alpha_{0i} + \alpha_{1i}u + \alpha_{2i}u^2 \\W_i(t_{ij}) &= \mathbf{X}_i(t_{ij}) + e_i(t_{ij}) = \alpha_{0i} + \alpha_{1i}t_{ij} + \alpha_{2i}t_{ij}^2 + e_i(t_{ij}),\end{aligned}$$

$e_{ij} \sim Normal(0, \sigma_e^2)$, com $Cov(e_{ij}, e_{ij'}) = 0$ (para $j \neq j'$)

$\boldsymbol{\alpha}_i = (\alpha_{0i}, \alpha_{1i}, \alpha_{2i})$ segue, *a priori*, um processo Dirichilet. Isso permite que $(\alpha_{0i}, \alpha_{1i}, \alpha_{2i})$ tenham diferentes comportamentos;

- componente de sobrevivência

$$\begin{aligned}h(u|\mathbf{W}_i) &= h_0(u) \exp \{ \gamma X_i(u) + \eta^T Z_i \} \\&= h_0(u) \exp \{ \gamma (\alpha_{0i} + \alpha_{1i}u + \alpha_{2i}u^2) + \eta^T Z_i \}.\end{aligned}$$

Outline

3 Polinômios de Bernstein

Polinômios de Bernstein

- Polinômios de Bernstein (PB) são bastante utilizados para estimar funções de densidade;
- Segundo Chang et al. (2005) e Osman & Ghosh (2012), o PB fornece boas aproximações polinomiais além de possuírem propriedades interessantes;
- Alguns trabalhos em análise de sobrevivência (com censura à direita) que utilizaram o PB foram:
 - ▶ Chang et al. (2005), fizeram uso do PB com grau aleatório para modelar a função risco acumulado (Bayesiano);
 - ▶ **Osman & Ghosh (2012)** trataram, sobretudo, curvas de sobrevivência que se cruzam (clássico);
 - ▶ Chen et al. (2014) utilizaram o PB no contexto de tempos de falha acelerados, incluindo variáveis dependentes do tempo (Bayesiano).

Polinômios de Bernstein

Considere uma função contínua $H(\cdot)$ no intervalo $(0, \tau]$. O Polinômio de Bernstein de grau m que aproxima essa função é dado por:

$$B(t; m, H) = \sum_{k=0}^m H\left(\frac{k}{m}\tau\right) \binom{m}{k} \left(\frac{t}{\tau}\right)^k \left(1 - \frac{t}{\tau}\right)^{m-k} \quad (6)$$

- o teorema de Weierstrass garante que $B(\cdot; m, H) \rightarrow H(\cdot)$ uniformemente no intervalo $(0, \tau]$ quando $m \rightarrow \infty$;

Polinômios de Bernstein

Por sua vez, a derivada da função $H(\cdot)$, denotada por $h(\cdot)$, pode ser aproximada por:

$$\begin{aligned} b(t; m, H) &= \frac{\partial B(t; m, H)}{\partial t} \\ &= \sum_{k=1}^m \left\{ H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right) \right\} \frac{f_{\beta}(t/\tau; k, m-k+1)}{\tau} \end{aligned} \quad (7)$$

sendo $f_{\beta}(\cdot; k, m-k+1)$ a função de densidade de uma Beta com parâmetros $(k, m-k+1)$;

- da mesma forma, tem-se que $b(\cdot; m, H) \longrightarrow h(\cdot)$ uniformemente em $(0, \tau]$ quando $m \longrightarrow \infty$;

Osman & Ghosh (2012)

A modelagem proposta por Osman & Ghosh (2012) foi feita da seguinte forma:

- assuma que existe $\tau < \infty$ tal que $\tau = \inf\{t : S(t) = 0\}$;
- o modelo para a função de risco é dado por:

$$h_m(t, \gamma) = \sum_{k=1}^m \gamma_k \mathbf{g}_{m,k}(t) = \boldsymbol{\gamma}^T \mathbf{g}_m(t), \quad 0 \leq t < \infty \quad (8)$$

em que

- ▶ $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_m)^T$ é tal que $\gamma_k \geq 0$ para todo k ;
- ▶ $\mathbf{g}_m(t) = (g_{m,0}(t), g_{m,1}(t), \dots, g_{m,m}(t))^T$ é chamada função de base e satisfaz $g_{m,k}(\cdot) \geq 0$ e $\int_0^\infty g_{m,k}(u) du < \infty$ para todo $k \leq m$.

- para a função de risco acumulado, tem-se:

$$H_m(t, \gamma) = \sum_{k=1}^m \gamma_k G_{m,k}(t) = \gamma^T \mathbf{G}_m(t), \quad 0 \leq t < \infty \quad (9)$$

em que

- ▶ $\mathbf{G}_m(t) = (G_{m,0}, G_{m,1}, \dots, G_{m,m})^T$ também é função de base, com

$$G_{m,k}(t) = \int_0^t g_{m,k}(u) du;$$

- a proposição desses autores é utilizar o PB com grau m nessas funções de base;

Isto é:

$$\begin{aligned} h_m(t, \gamma) &= \sum_{k=1}^m \gamma_k \mathbf{g}_{m,k}(t) \\ &\approx \sum_{k=1}^m \left\{ H\left(\frac{k}{m}\tau\right) - H\left(\frac{k-1}{m}\tau\right) \right\} \frac{f_\beta(t/\tau; k, m-k+1)}{\tau}. \end{aligned}$$

A partir disso é possível especificar de forma completa a função de verossimilhança.

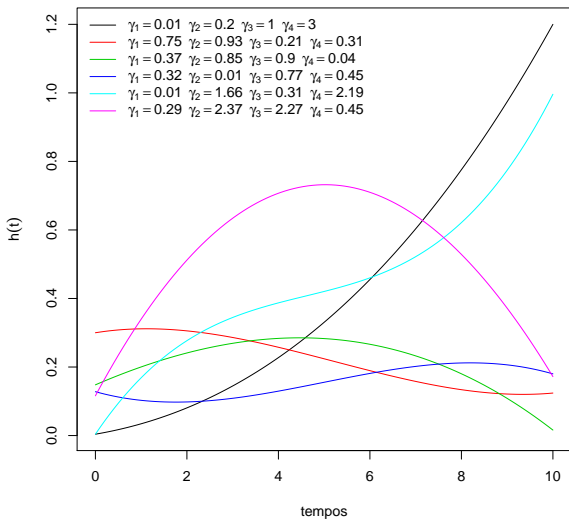


Figura : Função de risco aproximada a partir de diferentes combinações de γ , com grau $m = 4$.

Considerações sobre o PB

Algumas considerações importantes sobre o PB:

- são bastante flexíveis;
- um dos desafios é a escolha de m ;
- geralmente não é necessário um grau (m) muito alto;

Outline

4 Proposta de tese

Objetivos para a tese

O objetivo principal da tese é modelar dados longitudinais e de sobrevivência conjuntamente via Polinômios de Bernstein, utilizando abordagem Bayesiana.

Itens a serem possivelmente explorados:

- dados faltantes;
- erro de medição nas observações longitudinais;
- PB na função trajetória;

Outline

5 Referências

Referências



Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004) *Applied Longitudinal Analysis*. John Wiley & Sons.



Ibrahim, J. G., Chu, H. and Chen, L. M. (2010) Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, **28**, 2796-2801.






Osman, M. and Ghosh, S. K. (2012) Nonparametric regression models for right-censored data using Bernstein polynomials. *Computational Statistics & Data Analysis*, **56**, 559 – 573.






Tsiatis, A. A. and Davidian, M. (2004) Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, **14**, 809–834.

Referências

-  Brown, E. R. and Ibrahim, J. G. (2003) A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221–228.
-  Chang, I.-S., Hsiung, C. A., Yuh-Jennwu and Yang, C.-C. (2005) Bayesian survival analysis using Bernstein polynomials. *Scandinavian Journal of Statistics*, **32**, 447–466.
-  Chen, Y., Hanson, T. and Zhang, J. (2014) Accelerated hazards model based on parametric families generalized with Bernstein polynomials. *Biometrics*, **70**, 192–201.

Referências

-  Huang, Y., Dagne, G. and Wu, L. (2011) Bayesian inference on joint models of HIV dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine*, **30**, 2930–2946.
-  Wu, L., Liu, W., Yi, G. Y. and Huang, Y. (2012) Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012.
-  Wulfsohn, M. S. and Tsiatis, A. A. (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Obrigada! :)