Modelos conjuntos para dados longitudinais e de sobrevivência: uma aplicação a dados de qualidade de vida e tempo de sobrevivência de pacientes com câncer.

Aline Campos Reis de Souza
Instituto de Matemática e Estatística - Universidade de São Paulo

Antonio Carlos Pedroso de Lima
Instituto de Matemática e Estatística - Universidade de São Paulo

30 de abril de 2018

Resumo

A modelagem de variáveis contínuas pertencentes a um intervalo do tipo (0,1) é necessária em diversas aplicações, e permite o tratamento de dados como proporções, taxas e frações. O desenvolvimento de modelos de regressão que se adequem à estas características, tem sido o foco principal de muitos trabalhos publicados nas últimas décadas. No contexto da modelagem conjunta para dados longitudinais e de sobrevivência, poucos trabalhos atentam para o uso de distribuições adequadas para a componente longitudinal, sendo comumente empregado o uso de transformações dos dados. Neste trabalho, propomos uma extensão deste tipo de modelagem conjunta, através do uso de modelos de regressão beta com efeitos mistos para explicar as medidas longitudinais associadas. Partimos de uma motivação proveniente de um estudo sobre qualidade de vida e tempo de sobrevivência, realizado em dois hospitais públicos brasileiros especializados no tratamento do câncer: o Instituto do Câncer Dr. Octávio Frias de Oliveira (ICESP) e a Fundação Pio XII (Hospital do Câncer de Barretos). Utilizamos este conjunto de dados para ajustar os parâmetros do modelo proposto, obtendo, entre outros resultados, uma medida de associação estimada entre as duas respostas observadas, considerada atualmente como um fator essencial na indicação de cirurgias ou procedimentos médicos.

1 Introdução

Os estudos longitudinais consistem na observação repetida de uma variável, feita em diferentes instantes de tempo para cada indivíduo da amostra. Nas pesquisas da área médica este tipo de coleta de dados permite observar, além da resposta longitudinal e de covariáveis relevantes ao estudo, o tempo até a ocorrência de um evento de interesse para a investigação, como por exemplo o óbito de um paciente durante o período de um tratamento.

Uma forma de analisar dados com estas características é considerar separadamente a variável longitudinal e o tempo até a ocorrência do evento. Apesar da atrativa facilidade de execução desta abordagem, através do uso de modelos bem consolidados na literatura para cada uma das variáveis, em muitas situações é natural que exista algum tipo de associação entre estas duas respostas, o que conduz ao desenvolvimento de modelos que expliquem de forma apropriada esta relação. Nesta direção, existe na literatura uma grande diversidade de trabalhos envolvendo modelos conjuntos para variáveis longitudinais e tempo até a ocorrência de um evento (nomeadamente o tempo de sobrevivência). Utilizaremos, de maneira abreviada, a nomenclatura *modelos conjuntos* para nos referir a modelos especificados dentro deste contexto.

No presente trabalho, o interesse principal reside na análise do tempo de sobrevivência. Uma abordagem possível neste caso é utilizar os dados longitudinais como covariáveis que variam ao longo do tempo, associando a elas um erro de medida. A extensão do modelo de Cox para variáveis dependentes do tempo, utilizada inicialmente para incorporar a relação entre os dados longitudinais e o tempo de sobrevivência, apresenta fortes restrições que, em algumas aplicações, são difíceis de serem satisfeitas. É o que acontece com as imposições feitas sobre as covariáveis, que devem ser externas e não relacionadas ao mecanismo de falha. A inclusão covariáveis internas no modelo de Cox não é recomendada. Segundo Prentice (1982) esta prática pode gerar viés, uma vez que a observação é obtida com algum grau de incerteza.

O uso inadequado da extensão do modelo de Cox para variáveis dependentes do tempo motivou o surgimento de novas metodologias na modelagem conjunta, que visam diminuir o viés causado ao não se levar em conta os erros de medição associados às variáveis longitudinais. Basicamente, um modelo conjunto é construído através de uma componente longitudinal, uma componente para a variável tempo de sobrevivência, e uma estrutura de ligação que relaciona as duas variáveis. A especificação mais usual para explicar o processo longitudinal tem sido feita através de modelos lineares com efeitos mistos. Esta formulação é detalhadamente abordada em Rizopoulos (2012).

Nos últimos anos algumas extensões foram propostas a fim de se obter uma modelagem mais flexível para os perfis longitudinais. Ding & Wang (2008) propõe o uso de B-splines para modelar a trajetória média da componente longitudinal dos modelos conjuntos. Mais recentemente, Huong *et al.* (2017) considera P-splines com base polinomial truncada para parametrizar o processo longitudinal não linear. Vale ressaltar que, apesar da flexibilidade destas propostas, o aumento da dimensão dos efeitos aleatórios causados pelo processo de suavização traz bastante dificuldade de implementação computacional. Além disso, estas extensões se baseiam na normalidade das medidas longitudinais, não considerando seus devidos intervalos de variação.

Neste trabalho desenvolvemos uma extensão para os modelos conjuntos, com medidas longitudinais pertencentes ao intervalo (0,1), considerando para esta componente o modelo de regressão beta com efeitos mistos. Organizamos este artigo da seguinte maneira: na Seção 1 apresentamos a especificação do modelo conjunto, através da formulação dos submodelos longitudinal e de sobrevivência; na Seção 2 encontram-se aspectos relacionados ao processo de estimação dos parâmetros; na Seção 3 consideramos uma aplicação, utilizando dados de qualidade de vida e tempo de sobrevivência e na Seção 4 uma discussão a respeito dos resultados obtidos e seu seguimento é apresentada.

2 Especificação do modelo

Para a especificação do modelo conjunto, consideremos algumas notações. Sejam T_i^* o verdadeiro tempo de ocorrência do evento para um indivíduo i da amostra e T_i o tempo de ocorrência observado. Ao longo do processo de observação alguns indivíduos são censurados. Desta forma, denotando por C_i o tempo de censura associado ao i-ésimo elemento amostral, define-se T_i como sendo $\min\{T_i^*, C_i\}$ e $\delta_i = I(T_i^* \leq C_i)$ como um indicador do evento.

Defina $y_i(t)$ como o valor observado no instante t, referente ao i-ésimo indivíduo. A coleta destas informações resulta em um conjunto de medidas $\mathcal{Y}_i(t) = \{yi(t_{ij}), 0 \leq t_{ij} \leq T_i, j = 1, \cdots, n_i\}$. Associada a cada indivíduo da amostra está a trajetória das variáveis longitudinais, que é um processo especificado como $\mathcal{M}_i(t) = \{m_i(t), t \geq 0\}$, em que $m_i(t)$ corresponde ao verdadeiro valor da medição longitudinal no instante $t \geq 0$. Note que $\mathcal{M}_i(t)$ é um processo latente e portanto não observado. Desta forma para cada indivíduo i da amostra, a história das variáveis longitudinais $\mathcal{Y}_i(t)$ é composta por elementos de $\mathcal{M}_i(t)$ contaminados por um erro de medida.

2.1 Submodelo de sobrevivência

Para medir o efeito da covariável longitudinal no tempo até a ocorrência do evento, é necessário estimar $m_i(t)$, reconstruindo o processo latente $\mathcal{M}_i(t)$ para cada indivíduo. A caracterização da relação entre o processo longitudinal $\mathcal{M}_i(t)$ e o tempo até a ocorrência do evento é feita através de um modelo de riscos relativos com variáveis dependentes do tempo, sendo especificado da seguinte maneira,

$$\begin{aligned} h_i(t) &= \lim_{dt\to 0} \frac{P(t \leq T_i^* < t+dt | T_i^* \geq t, \mathcal{M}_i(t), w_i)}{dt} \\ &= h_0(t)\exp\left(\boldsymbol{\gamma}^\top w_i + \alpha m_i(t)\right), \qquad t \geq 0. \end{aligned} \tag{1}$$

em que $h_0(t)$ denota a função de risco basal, w_i é o vetor de covariáveis do i-ésimo indivíduo, relacionadas ao processo de sobrevivência, $\boldsymbol{\gamma}$ é o vetor de coeficientes de regressão associados à w_i e α é um parâmetro desconhecido que quantifica o impacto do processo longitudinal na observação do evento.

Neste trabalho, adotamos uma modelagem semiparamétrica para o risco (1). Para tanto consideramos que o risco basal $h_0(t)$ é uma função constante por parte, ou seja, $h_0(t)$ é da forma

$$h_0(t) = \sum_{k=1}^{K} \xi_k I(t_{k-1} < \xi_k \leq t_k). \tag{2}$$

2.2 Submodelo de longitudinal

A distribuição beta é muito flexível para modelagem de proporções, taxas e outras varíveis que estejam definidas em intervalos do tipo (0,1). Sua densidade, que pode ter formas bastante diferentes de acordo com a combinação dos valores dos parâmetros que indexam a distribuição, é dada por

$$p(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \qquad 0 < y < 1, \tag{3}$$

em que $p, q > 0$ e $\Gamma(\cdot)$ denota a função gama. Além disso, o valor esperado e a variância de y são dados respectivamente por

$$E(Y) = \frac{p}{p+q}, \qquad \text{e} \qquad V(Y) = \frac{pq}{(p+q)^2 + (p+q+1)}. \tag{4}$$

Com o objetivo de definir um modelo de regressão para variáveis aleatórias com distribuição beta, Ferrari & Cribari-Neto (2004) reparametrizam a densidade (3) como segue. Seja $\mu = E(Y)$ e $\phi = p + q$, temos que $V(Y) = V(\mu)/(1+\phi)$, em que $V(\mu) = \mu(1-\mu)$. Assim, os novos parâmetros obtidos μ e ϕ são, respectivamente, a média da variável aleatória beta e um parâmetro inversamente proporcional à variância de y, ou seja, um parâmetro de precisão. A função densidade reparametrizada em função de μ e ϕ é dada por,

$$p(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1, \tag{5}$$

com $0 < \mu < 1$ e $\phi > 0$.

Sejam $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ variáveis aleatórias independentes, em que $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ é o vetor com as n_i medidas repetidas do i-ésimo indivíduo. Suponha que, condicional ao vetor de efeitos aleatórios $\boldsymbol{b}_i = (b_{i0}, b_{i1})^\top$, as variáveis y_{ij} são independentes para $j = 1, \ldots, n_i$, e têm função de densidade dada por (5), com média μ_{ij} e parâmeto de precisão ϕ. Assuma ainda que o vetor

de efeitos aleatórios segue distribuição normal, com média zero e covariância D. Desta forma, temos a seguinte especificação

$$\begin{aligned} y_{ij}|b_{i0}, b_{i1} &\sim Be(\mu_{ij}, \phi) \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, D). \end{aligned} \tag{6}$$

Algumas suposições podem ser feitas a respeito da estrutura da matriz D, a depender da forma como se associam os elementos do vetor de efeitos aleatórios. Assumiremos que D é da forma $\sigma^2\boldsymbol{I}$, em que $\sigma^2 > 0$ e $\boldsymbol{I}$ denota a matriz identidade. O modelo de regressão é definido através da seguinte relação,

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\boldsymbol{b}_i \tag{7}$$

em que $\boldsymbol{x}_{ij}$ é o vetor de variáveis explicativas associadas aos efeitos fixos, $\boldsymbol{\beta}$ é o vetor de coeficientes de regressão e $\boldsymbol{z}_{ij}$ é o vetor de variáveis explicativas associadas aos efeitos aleatórios. A função $g(\cdot) : (0,1) \to \mathbb{R}$ é a função de ligação, a qual assumimos ser estritamente monótona e duas vezes diferenciável. Algumas funções podem ser atribuídas à $g(\cdot)$, como por exemplo as especificações logito, probito, complemento log-log e log-log, cuja forma funcional se encontra na Tabela 1.

Tabela 1: Funções de ligações e suas formas funcionais.

Ligação	Forma Funcional $g(\mu)$	μ
Logito	$\log(\mu/(1-\mu))$	$\exp(\eta)/(1+\exp(\eta)$
Probito	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
C. log-log	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\eta))$
Log-log	$-\log(-\log(\mu))$	$\exp(-\exp(-\eta))$

2.3 Função de verossimilhança

Para definir a função de verossimilhança conjunta do modelo algumas suposições devem ser consideradas. Assume-se que o vetor de efeitos aleatórios $\boldsymbol{b}_i$ é subjacente aos processos longitudinal e de sobrevivência, e que, condicional a $\boldsymbol{b}_i$ estes processos são independentes. Assim, sob estas hipóteses, denotando o vetor de parâmetros por $\boldsymbol{\theta} = (\theta_t, \theta_{y,b})^\top$, temos que a distribuição marginal de $\boldsymbol{y}_i$ é dada por

$$p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}) = p(T_i, \delta_i|\boldsymbol{b}_i, \theta_t)p(\boldsymbol{y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}_{y,b}) \tag{8}$$

E portanto, temos que a distribuição marginal de $(T_i, \delta_i, \boldsymbol{y}_i)$ é da forma

$$p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{\theta}) = \int p(T_i, \delta_i|\boldsymbol{b}_i, \theta_t) \prod_{j=1}^{n_i} p(\boldsymbol{y}_i|\boldsymbol{\beta}, \phi)p(\boldsymbol{b}_i|D)d\boldsymbol{b}_i. \tag{9}$$

A função log-verossimilhança de $\boldsymbol{\theta}$ é da forma

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \int p(T_i, \delta_i|\boldsymbol{b}_i, \theta_t) \prod_{j=1}^{n_i} p(\boldsymbol{y}_i|\boldsymbol{\beta}, \phi) p(\boldsymbol{b}_i|D) d\boldsymbol{b}_i \\
&= \sum_{i=1}^{n} \log \int [h_0(t) \exp(\boldsymbol{\gamma}^\top w_i + \alpha m_i(t))]^{\delta_i} \exp\left(\int_0^{T_i} -h_0(s) \exp(\boldsymbol{\gamma}^\top w_i + \alpha m_i(s)) ds \right) \\
&\times \prod_{j=1}^{n_i} \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y^{\mu_{ij}\phi - 1} (1-y)^{(1-\mu_{ij})\phi - 1} \\
&\times 2\pi^{-q_b/2} \det(D)^{-1/2} \exp\left(\frac{\boldsymbol{b}_i \top D^{-1} \boldsymbol{b}_i}{2} \right) d\boldsymbol{b}_i \qquad (10)
\end{aligned}
$$

em que q_b é a dimensão do vetor de efeitos aleatório $\boldsymbol{b}$ e μ_{ij} é obtido pela relação (7). A avaliação da contribuição de cada indivíduo i da amostra na função de verossimilhança depende da resolução das seguintes integrais com respeito aos efeitos aleatórios:

$$
\begin{aligned}
\mathcal{I}_i &= \int [h_0(t) \exp(\boldsymbol{\gamma}^\top w_i + \alpha m_i(t))]^{\delta_i} \exp\left(\int_0^{T_i} -h_0(s) \exp(\boldsymbol{\gamma}^\top w_i + \alpha m_i(s)) ds \right) \\
&\times \prod_{j=1}^{n_i} \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y^{\mu_{ij}\phi - 1} (1-y)^{(1-\mu_{ij})\phi - 1} \\
&\times 2\pi^{-q_b/2} \det(D)^{-1/2} \exp\left(\frac{\boldsymbol{b}_i \top D^{-1} \boldsymbol{b}_i}{2} \right) d\boldsymbol{b}_i \qquad (11)
\end{aligned}
$$

Por causa de dificuldades enfrentadas no desenvolvimento da integração, ou pela não existência de solução analítica, é necessária a implementação de técnicas de análise numérica nestes casos.

3 Aplicação: Dados ICESP

Com o objetivo de avaliar o tempo de sobrevivência e a qualidade de vida de pacientes com neoplasia maligna que deram entrada em unidade de terapia intensiva, um estudo de coorte prospectivo, apresentado em Normilio-Silva *et al.* (2016), foi realizado em dois hospitais públicos brasileiros cuja especialidade é o tratamento do câncer: o Instituto do Câncer Dr. Octávio Frias de Oliveira (ICESP) e a Fundação Pio XII (Hospital do Câncer de Barretos), ambos situados estado de São Paulo. Um total de 803 indivíduos maiores de 18 anos de idade, com malignidades comprovada e admitidos em UTIs dos hospitais participantes, foram incluídos no estudo. As variáveis resposta coletadas foram o índice de utilidade da qualidade de vida, e o tempo de sobrevivência dos pacientes.

A primeira variável foi obtida através da aplicação do questionário EQ-5D-3L, cujo resultado foi convertido em um índice de utilidade, que assume valores no intervalo (0,1). A medição da qualidade de vida foi feita antes da entrada na UTI, e acompanhada ao longo dos seguintes dias após a internação: 15, 90, 180, 365 e 540. O tempo de sobrevivência foi definido como o tempo até a ocorrência do óbito, tendo sua origem a partir da data da internação na UTI. O tempo máximo de estudo é 720 dias. Com relação às covariáveis, foram coletadas informações sócio-demográficas dos pacientes , informações de aspecto clínico ,a coexistência de doenças ,e uma covariável que especifica o centro de coleta dos dados (ICESP ou Hospital do Câncer de Barretos).

Na Figura 1 apresentamos os perfis do índice de utilidade da qualidade de vida considerando a amostra completa, juntamente com o seu perfil médio. Notam-se padrões de comportamento variados de acordo com cada paciente, e uma tendência média de crescimento ao longo do tempo.

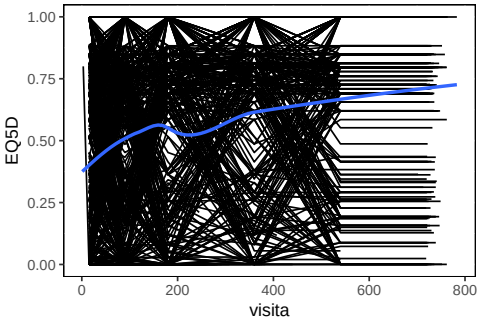

Figura 1: Perfis do índice de utilidade da qualidade de vida e perfil médio, considerando a amostra completa.

Esta trajetória da qualidade de vida nos dias subsequentes à alta da UTI é esperada, uma vez que pacientes debilitados são submetidos a cuidados intensivos para a melhora de sua saúde. Além disso, as ocorrências de óbitos ao longo do tempo faz com que permaneçam na amostra indivíduos que apresentam uma melhor reação aos tratamentos, e que portanto tendem a ter uma sobrevida e um índice de utilidade maior. Na Figura 2 apresentamos a curva de sobrevivência obtida via estimador de Kaplan-Meier.

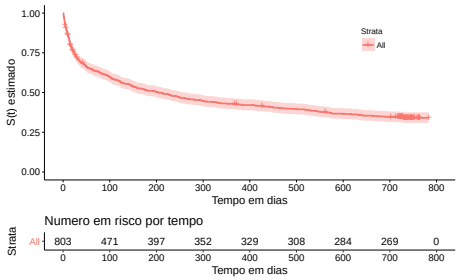

Figura 2: Curva de sobrevivência obtida através do estimador de Kaplan-Meier, considerando a amostra completa.

Para captar a relação entre as duas respostas de interesse, utilzamos a análise conjunta, através da seguinte formulação. Denotemos por $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^\top$ as n_i medidas do índice de utilidade da qualidade de vida, associadas ao i-ésimo indivíduo da amostra. Sejam T_i o tempo de ocorrência de óbito e δ_i o indicador de censura, para o i-ésimo paciente observado. Consideramos como submodelo longitudinal a especificação dada em (6), em que a estrutura de regressão (7) é definida por

$$g(\mu) = \beta_0 + \beta_1 t + b_0, \tag{12}$$

e a função de ligação utilizada é a logito. O submodelo de sobrevivência é especificado de acordo com a formulação (1) da maneira que segue,

$$h_i(t) = h_0(t) \exp\left(\alpha m_i(t)\right), \qquad t \geq 0, \tag{13}$$

com função de risco basal constante por partes. A avaliação do logaritmo da função verossimilhança depende da resolução da integral (11), cuja solução numérica é feita utilizando a quadratura de Gauss-Hermite. O ajuste do modelo foi feito através de métodos de maxização da verossimilhança. As estimativas dos parâmetros foram obtidas utilizando a função *optim*, disponível no pacote *stats* do R, e estão dispostas na Tabela 2. O modelo conjunto encontra uma forte associação entre o índice de utilidade de qualidade de vida e o risco de morte: uma diminuição de 0.1 no índice de utilidade de qualidade de vida corresponde a um aumento de $\exp(\alpha) = 1,28$ vezes no risco de morte.

Tabela 2: Estimativas dos parâmetros do modelo conjunto, utilizando os dados do ICESP.

Parâmetros	Estimativas
β_0	0.4407
β_1	0.0002
$\log \phi$	-1.1711
$\log D$	-1.2599
α	-2.4795
$\log \xi_1$	-2.9603
$\log \xi_2$	-3.3639
$\log \xi_3$	-4.5189
$\log \xi_4$	-5.0206
$\log \xi_5$	-5.5994
$\log \xi_6$	-5.2213

4 Discussão

Atualmente, na área da saúde, há uma necessidade cada vez maior de se estabelecer não apenas as estimativas de prognóstico ou sobrevida, mas principalmente a relação entre tais quantidades. O uso da modelagem conjunta neste caso permite captar o grau de associação das respostas de interesse, considerando a inclusão de características intrínsecas de cada indivíduo. Este aspecto do modelo contribui de maneira notável da tomada de decisões médicas, auxiliando na indicação individualizada de tratamentos.

A especificação mais usual para explicar o processo longitudinal tem sido feita através de modelos lineares com efeitos mistos, considerando medidas longitudinais com distribuição normal. Neste trabalho propomos o uso de modelos de regressão beta com efeitos mistos para modelar variáveis longitudinais pertencentes ao intervalo (0,1).

O seu desenvolvimento até a realização deste Simpósio, terá seguimento no cálculo da matriz de variância-covariância assintótica através do método *Bootstrap*, na inserção de um coeficiente angular aleatório no modelo de regressão beta misto, e na comparação dos resultados obtidos por meio de análises conjuntas, ou separadas, utilizadas neste tipo de dados.

Referências

Ding, J. & Wang, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**(2), 546–556.

Ferrari, S. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Huong, P. T. T., Nur, D. & Branford, A. (2017). Penalized spline joint models for longitudinal and time-to-event data. *Communications in Statistics-Theory and Methods*, **46**(20), 10294–10314.

Normilio-Silva, K., de Figueiredo, A. C., Pedroso-de Lima, A. C., Tunes-da Silva, G., da Silva, A. N., Levites, A. D. D., de Simone, A. T., Safra, P. L., Zancani, R., Tonini, P. C. *et al.* (2016). Long-term survival, quality of life, and quality-adjusted survival in critically ill patients with cancer. *Critical care medicine*, **44**(7), 1327–1337.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**(2), 331–342.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.