Universidade Estadual de Maringá
Programa de Pós-graduação em Bioestatística

FELIPE EMANOEL BARLETTA MENDES

Joint model para dados longitudinais e multi-estado: Uma aplicação para câncer de próstata

Dissertação de Mestrado

Maringá - Paraná
2018

FELIPE EMANOEL BARLETTA MENDES

Joint model para dados longitudinais e multi-estado: Uma aplicação para câncer de próstata

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística.
Orientadora: Profª. Drª. Isolde Previdelli
Coorientador: Prof. Dr. Paulo Canas Rodrigues

Maringá - Paraná
2018

FELIPE EMANOEL BARLETTA MENDES

Joint model para dados longitudinais e multi-estado: Uma aplicação para câncer de próstata

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística.
Orientadora: Profª. Drª. Isolde Previdelli
Coorientador: Prof. Dr. Paulo Canas Rodrigues

Trabalho aprovado.
Maringá - Paraná, 21 de Março de 2018:

Profª. Drª. Isolde T. S. Previdelli
Orientador

Paulo Canas Rodrigues
Coorientador

Josmar Mazucheli
Membro 3

Giovani Loiola Silva
Membro 4

Maringá - Paraná
2018

Dedico esta dissertação a minha família e todas as famílias que de alguma forma foram vítimas do câncer de próstata.

Agradecimentos

Agradeço a professora Isolde Previdelli, por todo apoio, orientação, compreensão e dedicação inspiradoras. Tais virtudes sempre colocadas com muito amor, e sem este amor tenho certeza que não conseguiria finalizar este mestrado. Amor que sinto em meu âmago que dever ser muito semelhante ao amor de uma mãe. Como não conheci minha mãe biológica, me arrisco a inferir, sem medo de cometer um erro padrão com grande variância, que recebi uma amostra significativa deste sentimento tão nobre.

Também agradeço imensamente a coorientação do professor Paulo Canas, que mesmo de longe, apoiou e deu significativa contribuição e também foi um dos agentes principais para que eu pudesse ter a experiência de passar alguns meses na *University of Tampere* na Finlândia, disponibilizando moradia e local de estudos naquela Universidade e ainda se tornando um grande amigo e conquistando minha total admiração pela pessoa que é.

Ao meu pai, pelo exemplo de vida e pelo total apoio do meu período estabelecido em Maringá-PR e à Finlândia. Sempre foi o meu maior incentivador e fomentador pela cultura e educação. Toda a minha formação se deve muito a sua existência em minha vida.

À minha namorada Thalita, por toda a inspiração, amor e atenção dedicados e toda paciência. Sei que teve momentos em que apenas falava e pensava em minhas atividades do mestrado, porém com esta fase finalizada, irei retribuir com o mesmo amor depositado por ela em mim. Os momentos de companheirismo, incentivo e lazer proporcionados por sua presença foram fundamentais.

Aos meus colegas Omar Pereira, Beatriz e Tiago Suguiura, pela constante ajuda e por terem acreditado e confiado em mim desde o início e por terem criado um ambiente de estudo muito descontraído e saudável por meio de suas companhias.

Um especial agradecimento ao meu padrinho Miguel, e a todos os nossos amigos em comum, que assim como ele e eu, temos o mesmo propósito primordial.

Quero também minha demonstrar minha gratidão a todos os membros da banca Walmes Zeviani, Josmar Mazucheli, Giovani Silva e todos os professores que participaram de minha formação no programa de Bioestatística da UEM. Não posso esquecer de meus professores da graduação, tais como o professor Paulo Justiniano, professora Silvia Shimakura, que além de

terem um papel fundamental na minha formação, sempre me incentivaram para a continuação de meus estudos e cederam espaço no Laboratório de Estatística e Geoinformação (LEG) da UFPR para minha preparação, tanto antes da prova de seleção para o mestrado quanto em meu regresso da Finlândia para finalizar a redação da dissertação. E devo citar também o professor Elias Krainski, que foi meu colega de graduação e nesta última fase deu importante ajuda enquanto estive no LEG.

E por fim, agradecer o professor Anssi Auvinen, da *Faculty of Social Sciences*, *University of Tampere* que cedeu os dados para realizar as análises desta dissertação. E agradecer também o professor Loic Ferrer da *Université de Bordeaux* que deu preciosas dicas para o ajuste do modelo.

Resumo

Joint models para dados longitudinais e de eventos no tempo, são uma ferramenta poderosa que leva em conta estes dois tipos de dados simultaneamente em um único modelo, permitindo inferir sobre a dependência e associação entre o biomarcador longitudinal (por exemplo, *prostate-specific antigen*, PSA) e eventos no tempo para uma melhor avaliação do efeito de um tratamento. Esses modelos são úteis para estudos no campo da saúde que visam a compreensão da doença (por exemplo, câncer de próstata), considerando o seu desenvolvimento ao longo do tempo e a quantidade de tempo até o paciente atingir o estado absorvente (por exemplo, morte). Os modelos conjuntos mais utilizados, que resultam de uma combinação de um modelo longitudinal e análise de sobrevivência, não permitem monitorar a ligação entre o biomarcador longitudinal e as transições entre os múltiplos estados da doença até atingir o estado absorvente. Para entender melhor esta ligação entre o biomarcador longitudinal e as transições entre os múltiplos estados, nesta dissertação usou-se um modelo conjunto que combina o modelo longitudinal e o modelo Markov multi-estado. Uma aplicação é apresentada onde um conjunto de dados do câncer de próstata é considerado. Os parâmetros do modelo são estimados pelo método da máxima verossimilhança, o que é feito em dois estágios: (i) no primeiro estágio, os efeitos fixos e aleatórios são estimados com base no biomarcador longitudinal PSA; e (ii) no segundo estágio essas estimativas são usadas, para vincular o modelo longitudinal com o modelo multi-estado de Markov, permitindo a mensuração do impacto para o risco de morte, considerando as covariáveis demográficas, em cada transição entre os estados da doença ao longo do tempo. Desta forma, o modelo é capaz de avaliar a trajetória do biomarcador, definir os riscos das transições entre estados de saúde e quantificar o impacto da dinâmica do PSA em cada intensidade de transição.

Palavras-chave:*Joint model*, Modelo longitudinal, Modelo multi-estado de Markov, Câncer de próstata.

Abstract

Joint models for longitudinal and time-to-event data are a powerful tool that take into account these two of data types simultaneously into a single model, allowing to infer about the dependence and association between the longitudinal biomarker (e.g. prostate-specific antigen, PSA) and time-to-event for a better assessment of the effect of a treatment. These models are useful for studies in the field of health that aim at understanding the disease (e.g. prostate cancer), considering its development over time and the amount of time until the patient reaches the absorbent state (e.g. death). The most used joint models, that result from a combination of a longitudinal model and survival analysis, do not allow to monitor the link between the longitudinal biomarker and the transitions between the multiple states of the disease until it reaches the absorbent state. In order to better understand this link between the longitudinal biomarker and the transitions between the multiple states, in this dissertation we use a joint model that combines the longitudinal model and the multi-state Markov model. An application is presented where a data set from prostate cancer is considered. The parameters of the model are estimated by maximum likelihood, which is performed in two stages: (i) in the first stage the fixed and random effects are estimated based on the longitudinal biomarker PSA; and (ii) in the second stage these estimates are used to link the longitudinal model with the multi-state Markov model, allowing the measurement of the impact for the risk of death, considering demographic covariables, in each transition between the states of the disease along time. In this way, the model is able to assess the biomarker's trajectory, define the risks of transitions between health states, and to quantify the impact of the PSA dynamics on each transition.

Key words: Joint model, Longitudinal model, Multi-state Markov model, Prostate cancer.

Lista de figuras

Lista de tabelas

Sumário

Capítulo 1

Introdução

Há muitas opções para modelar fenômenos caracterizados por transições entre estágios ou estados de evolução ao longo do tempo (SAINT-PIERRE et al., 2003). Em estudos longitudinais e análise de sobrevivência na área de saúde, há muitos modelos para monitoramento de progressão de doenças como HIV e câncer, por exemplo, mas nem todos consideram os estágios da doença com a perspectiva de estudar a ligação entre a transição dos estados até uma recaída ou estado de morte e ainda levar em conta o tempo até a ocorrência do evento de interesse, como pode ser visto em Tsiatis e Davidian (2004), Putter et al. (2007) e Putter (2016) .

Na prática, múltiplos tipos de estados de uma doença podem ocorrer sucessivamente antes que o paciente atinja o último estado (FRYDMAN; SZAREK, 2009; JACKSON et al., 2011), e para monitorar as transições entre os estados e o risco ao longo do tempo, pode ser realizada uma modelagem conjunta entre um modelo longitudinal e um modelo multi-estado de Markov com riscos proporcionais (FERRER et al., 2016). Distinguir essas transições entre estados de saúde e o estado absorvente (recaída ou morte) ao longo do tempo e considerando outras covariáveis, é essencial para entender e predizer de forma mais acurada a evolução da doença, e para os médicos e clínicos responsáveis pelo paciente, que têm a necessidade de diagnosticar com maior precisão cada momento da doença, é importante para a adaptação de tratamentos individualizados e mais eficazes (DANTAN et al., 2011).

Geralmente, quando se deseja modelar estatisticamente dados de doenças como o câncer (BRESLOW et al., 1987), usam-se métodos como regressão logística, que é apropriada para variáveis respostas dicotômicas, ou seja, descreverá a relação entre esta resposta, que será usualmente os estados inicial e final da doença, e um conjunto de covariáveis (fatores de risco), e ainda quantifica o grau risco de cada fator através do cálculo da razão de chances ou *odds ratio*. Outro modelo que pode se encaixar no contexto do estudo de câncer e muito difundido é o de sobrevivência de riscos proporcionais, (LEE; WANG, 2003), que tem foco ser aplicado quando o desejo é modelar o tempo até o desfecho da doença, a morte por exemplo, (YU

et al., 2008). Porém, nenhuma dessas duas abordagens permitem calcular a probabilidade de transição entre os vários estágios da doença e o risco de covariáveis para as diversas transições, o que torna a modelagem conjunta com um modelo longitudinal e multi-estado de riscos proporcionais muito atraente.

Este tipo de modelagem conjunta é chamada *joint modelling* ou *joint model* (RIZOPOULOS, 2012), que vem ganhando grande apelo nos últimos anos com muitos artigos com este arcabouço teórico sendo publicados na área da Bioestatística. Alguns deles estudando a progressão do câncer (LANGE et al., 2015) e também considerando a modelagem conjunta com submodelo multi-estado de Markov(FERRER et al., 2016), do qual esta pesquisa está se baseando.

Isto posto, a principal motivação para a elaboração desta dissertação, é compreender progressão de doenças a respeito de fatores de riscos que influenciam a evolução da enfermidade ao longo do tempo aliando informações de como esses fatores de risco atuam nas diversas transições entre todos os possíveis estágios, que a doença pode assumir. Portanto, o principal objetivo deste estudo é ajustar um modelo para monitoramento de doenças, como o câncer, que considera todos os estágios da doença e o tempo até o evento de interesse.

A grande vantagem desta modelagem estatística é possibilitar a inclusão, em um mesmo modelo, de dados com medidas repetidas ao longo do tempo (modelo longitudinal) e dados com transições entre estágios da doença (modelo multi-estado). Geralmente estas duas abordagens são realizadas separadamente e este olhar simultâneo, do *joint model*, se apresenta mais abrangente gerando resultado mais extensivo, ou seja, o pesquisador poderá compreender como a doença se comporta individualmente ao longo do tempo, e compreender como os fatores de riscos agem em seus vários estágios na progressão da enfermidade.

Esta dissertação será apresentada com a seguinte organização: No Capítulo 2 é discorrido sobre a metodologia acerca dos modelos longitudinal e multi-estado em que se detalha todos os conceitos e métodos de estimação, assim como acerca do *joint model* que agregará os dois submodelos anteriores. Na sequência, no Capítulo 3 é feita a apresentação dos modelos ajustados e seus resultados assim como a descrição dos dados utilizados nos ajustes. E por fim no Capítulo 4, as considerações finais são dadas ao leitor.

Capítulo 2

Metodologia

Neste capítulo será apresentado o modelo proposto, o *joint model*, que é composto por dois tipos de modelos ou submodelos. Um modelo longitudinal e outro modelo multi-estado. Para compreensão do submodelo longitudinal, é relevante introduzir conceitos de dados correlacionados ou dados longitudinais e para expor sobre o modelo multi-estado de Markov com riscos proporcionais, que são fundamentados nos princípios da análise de sobrevivência e dos processos estocásticos, é pertinente fazer uma introdução acerca estas duas teorias. E por fim, será descrito como a modelagem conjunta entre eles é simultaneamente realizada.

O recurso computacional utilizado para realizar todas as análises será a linguagem de programação *R version 3.3.3* (R Core Team, 2017). O ajuste da modelagem conjunta será possível devido ao pacote que o R disponibiliza, chamado *JM* (RIZOPOULOS, 2010). O modelo longitudinal será ajustado com a utilização do pacote *nlme* (PINHEIRO et al., 2014) e o multi-estado com o pacote *mstate* (WREEDE et al., 2011).

2.1 Modelo longitudinal

A definição de dados longitudinais basicamente se caracteriza quando a variável resposta, relacionada as medidas dos indivíduos que compõem o estudo, é acompanhada repetidamente ao longo do tempo, prospectivamente ou retrospectivamente, criando uma possível estrutura de dados correlacionados e desta forma, para tais medidas repetidas, a ordem ou o tempo devem ser considerados no planejamento e na análise. Devido a este fato, métodos estatísticos que consideram uma estrutura de covariância são requeridos para modelar dados correlacionados.

Dados longitudinais induzem uma estrutura de covariância, e por meio dos modelos lineares mistos (em inglês LMM), é possível modelar tal peculiaridade com a inclusão no modelo de efeitos fixos e aleatórios (VERBEKE; MOLENBERGHS, 2009). Essa classe de modelos pode ser vista como uma extensão dos modelos de regressão linear (em inglês LM) (YAN; SU, 2009). A principal diferença é que o LMM permite a inclusão dos efeitos fixos e aleatórios

acomodando desta forma a estrutura de covariância inerente aos dados, diferentemente do LM que adota apenas efeitos fixos e parte do pressuposto de independência das observações, ou seja, esta estrutura de covariância considera variância constante. Mais sobre esta e outras estruturas serão detalhadas a seguir, assim como o próprio modelo linear misto.

2.1.1 Modelo linear misto

Atualmente há uma vasta literatura disponível sobre os modelos lineares mistos, algumas com grande destaque na comunidade estatística tais como, Pinheiro e Bates (1978), McCulloch e Searle (2001), Verbeke e Molenberghs (2009).

Quando o modelo possui apenas o erro aleatório e os demais são efeitos fixos, chama-se de modelo fixo ou modelo marginal e este modela características populacionais e seus efeitos são constantes desconhecidas e estimados a partir dos dados. Se o modelo apresentar todos os componentes aleatórios, chama-se de modelo aleatório e modela características individuais e seus efeitos sendo variáveis aleatórias, portanto, são estimados os parâmetros que descrevem a sua distribuição. Se o modelo possuir efeitos fixos e aleatórios, além do erro aleatório, denomina-se modelo misto ou modelo condicional.

A representação matricial do modelo linear misto pode ser escrita da seguinte maneira,

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \epsilon_i \\ b_i &\sim N(0, G) \\ \epsilon_i &\sim N(0, \sigma^2 R_i), \end{aligned} \tag{2.1}$$

em que

- y_i é o vetor n_i x 1 da variável reposta para as observações no i-ésimo grupo.
- X_i a matriz n_i x p associada aos efeitos fixos das observações no i-ésimo grupo.
- β é o vetor p x 1 de coeficientes de efeitos fixos.
- Z_i a matriz n_i x q associada aos efeitos aleatórios das observações no i-ésimo grupo.
- b_i é o vetor q x 1 de coeficientes de efeitos aleatórios para o grupo i.
- ϵ_i é o vetor n_i x 1 de erros das observações no i-ésimo grupo.
- G a matriz q x q das covariâncias dos efeitos aleatórios.
- $\sigma^2 R_i$ a matriz n_i x n_i das covariâncias dos erros para o grupo i.

Ao adicionar o efeito aleatório no modelo, Zb, onde Z, assim como a matriz de efeitos fixos X, é uma matriz conhecida e b o vetor de efeitos aleatórios, é conveniente expressar o modelo em termos de esperança condicional.

$$E[y \mid b] = X\beta + Zb. \tag{2.2}$$

Assim, pela expressão (2.1), temos que $b_i \sim N(0, G)$, ou seja, $E[b] = 0$ e var$(b) = G$.

É definido também que var$(y \mid b) = R$ e com a equação (2.2), é dado a distribuição marginal de y,

$$y \sim (X\beta, ZGZ' + R). \tag{2.3}$$

Pela expressão (2.3) é mostrado que os efeitos fixos estão inseridos na média e os efeitos aleatórios estão inseridos na variância de y e, portanto, atribui-se uma estrutura para esta variância que pode ser denotada como a expressão a seguir.

$$\Sigma = var(y) = ZGZ' + R. \tag{2.4}$$

2.1.1.1 Método de estimação

Há vários métodos de estimação dos componentes de variância para modelos lineares mistos, como por exemplo, o métodos dos momentos (ANOVA) que produz estimadores não viesados e de variância mínima. Sabe-se que a estimação com dados balanceados através da ANOVA, pode ser estendido para dados desbalanceados como é mencionado por Searle et al. (2009). Apesar da estimação de componentes de variância para dados balanceados ser mais simples e ser o alicerce teórico, quando os dados são desbalanceados, segundo McCulloch e Searle (2001), a ANOVA para dados desbalanceados perdeu muita popularidade, pois estimadores da ANOVA, neste caso, nem sempre são baseadas em estatísticas suficientes, ou seja, a consequência é que seu estimadores não possuem propriedades ótimas.

Uma elegante alternativa é o método da máxima verossimilhança restrita (em inglês *Restricted Maximum Likelihood Estimation - REML*) que produz estimadores $BLUP$, ou seja, *best linear unbiased predicted* que significa melhor preditor linear não viesado. Para descrever o método $REML$, o ponto de partida será o conhecido e famoso método de estimação de máxima verossimilhança (*em inglês Maximum Likelihood - LM*).

Considere a função densidade de probabilidade para o caso gaussiano,

$$f(y_i; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\}.$$

A distribuição conjunta é escrita,

$$f(y_1, y_2, ..., y_n; \mu; \sigma^2) = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \right\}^n \exp\left\{ -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} \right\}.$$

Em termos dos parâmetros μ e σ^2, tem-se a chamada verossimilhança,

$$L(\mu; \sigma^2; y_1, y_2, ..., y_n) = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \right\}^n \exp\left\{ -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} \right\},$$

e aplicando o log na verossimilhança, tem-se,

$$l(\mu; \sigma^2; y_1, y_2, ..., y_n) = \log[L(\mu; \sigma^2; y_1, y_2, ..., y_n)] = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}.$$

Como não conhecemos μ, σ^2, então para estimá-los é preciso maximizar a verossimilhança. Para tal, deriva-se a função de log-verossimilhança em função dos parâmetros desconhecidos e iguala-se a zero. Assim, tem-se,

$$\hat{\mu} = \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n},$$

$$\hat{\sigma^2} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu})^2}{n},$$

logo, a ML envolve os parâmetros média (μ) e variância (σ^2) e pode ser maximizada para estimar o parâmetro μ, e esta solução não depende da estimativa de σ^2.

E ainda considerando o caso gaussiano, podemos escrever:

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) + (\bar{y} - \mu)]^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2.$$

Dessa forma podem-se separar as duas verossimilhanças e reescrever a log-verossimilhança para a distribuição gaussiana,

$$\log[L(\mu; \sigma^2; y_1, y_2, ..., y_n)] = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2},$$

$$\log[L(\mu; \sigma^2; y_1, y_2, ..., y_n)] = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{2\sigma^2} - \frac{n(\bar{y} - \mu)^2}{2\sigma^2}.$$

Se for selecionada uma amostra aleatória de tamanho n de uma distribuição gaussiana $N(\mu, \sigma^2)$, então a distribuição amostral de $\bar{y}$ terá, $\bar{y} \sim N(\mu, \frac{\sigma^2}{n})$. Então a verossimilhança para $\bar{y}$ é,

$$L(\bar{y}) = \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{(\bar{y}-\mu)^2}{2\sigma^2/n}} = \frac{n}{2\pi\sigma^2} e^{-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}},$$

e consequentemente, a log-verossimilhança para a média amostral é da forma,

$$\log[L(\bar{y})] = \frac{1}{2}\log(n) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{n(\bar{y}-\mu)^2}{2\sigma^2}. \tag{2.5}$$

Então a log-verossimilhança de uma amostra aleatória que segue uma distribuição gaussiana, pode ser escrita como

$$\log[L(\mu;\sigma^2;y_1,y_2,...,y_n)] = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{2\sigma^2} - \frac{n(\bar{y}-\mu)^2}{2\sigma^2}.$$

Separando a log-verossimilhança para a amostra aleatória, $y_1, y_2, ..., y_n$ tem-se,

$$\begin{aligned}\log[L(\mu;\sigma^2;y_1,y_2,...,y_n)] = &-\frac{1}{2}\log(2\pi\sigma^2) - \frac{n(\bar{y}-\mu)^2}{2\sigma^2} \\ &-\frac{n-1}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{2\sigma^2}.\end{aligned} \tag{2.6}$$

Com isso pode-se ver que a primeira linha da expressão (2.6) é muito similar a log-verossimilhança da média amostral, $\bar{y}$, como visto na expressão (2.5), diferenciando apenas pelo termo constante $\frac{1}{2}\log(n)$. A segunda linha da expressão (2.6) envolve somente o parâmetro σ^2. Logo o nome de máxima verossimilhança restrita se dá devido a referência somente a variância e esta segunda parte da expressão que é chamada de *REML*. Esta verossimilhança é maximizada separadamente da primeira parte e isso produz uma estimativa para σ^2, no qual é chamada de estimativa da variância por *REML*.

Assim, para a máxima verossimilhança restrita (ou residual) podem-se fazer as seguintes considerações:

- A *REML* envolve somente o parâmetro de variância (σ^2).
- A *REML* pode ser maximizada para estimar o parâmetro σ^2 e será diferente da estimativa de ML.

Os dois tipos de métodos baseados na verossimilhança, a estimação por máxima verossimilhança completa (ML) e a estimação por máxima verossimilhança restrita $(REML)$, são fornecidos para estimar os parâmetros da regressão β e as componentes de variância Σ. Com base no pressuposto de normalidade multivariada da resposta Y_i, a função de log-verossimilhança completa para o modelo linear misto é,

$$\ell_{ML}(\beta,\Sigma) = \log\ \ L(\beta,\Sigma) = -\frac{1}{2}(\sum_{i=1}^{n}\log\ \ det\Sigma + \sum_{i=1}^{n}(y_i - X_i\beta)^T\Sigma^{-1}(y_i - X_i\beta)).$$

O estimador de máxima verossimilhança segue por meio da maximização da função de verossimilhança sobre β e Σ simultaneamente. É também fácil de mostrar que para o parâmetro de covariância fixo Σ, o estimador de máxima verossimilhança de β coincide com o estimador correspondente de mínimos quadrados generalizados. Sabe-se que o estimador de máxima verossimihança de Σ é viciado para baixo, ou seja, $\hat{\Sigma}$ subestima Σ. Para atenuar este viés, o procedimento de máxima verossimilhança restrita maximiza a seguinte função de log-verossimilhança,

$$\ell_{REML}(\beta,\Sigma) = \ell_{ML}(\beta,\Sigma) - \frac{1}{2}\log\ \ det(\sum_{i=1}^{n}X_i^T\Sigma^{-1}X_i), \tag{2.7}$$

onde o segundo termo, da expressão (2.7), tem a função de atenuar ou corrigir o viés da variância.

Dado as expressões (2.3) e (2.4), o interesse é, simultaneamente, estimar β e Σ. Segundo Filho (2002), os estimadores obtidos pelo método $REML$ com dados balanceados são os mesmos obtidos pelo método da ANOVA e podem ser obtidos, sob normalidade, de forma analítica.

É importante salientar que, com exceção para dados balanceados, não há soluções analíticas para estimar as componentes de variância. Assim, métodos numéricos iterativos são necessários como, por exemplo, o método de Quase-Newton, que é um caso especial do método numérico de Newton, porém mais eficiente computacionalmente pois não necessita de segundas derivadas e tão eficiente quanto o de Newton. O método Quase-Newton mais popular é o método de **BFGS** (*Broyden–Fletcher–Goldfarb–Shanno*) e será adotado para o ajuste do modelo linear misto desta dissertação. Mais detalhes sobre o método de Quase-Newton podem ser visto em Martınez e Santos (1995).

2.1.1.2 Estruturas de covariâncias

Como visto na expressão (2.3), há uma componente de covariância a ser considerada e esta é modelada levando em conta a matriz Z que está associada aos efeitos aleatórios, a matriz G que é associada as covariâncias dos efeitos aleatórios e a matriz R que está associada as covariâncias dos erros. Como os modelos lineares mistos permitem modelar a heterogeneidade existente entre e dentro dos indivíduos, os coeficientes de regressão podem ser diferentes entre indivíduos, ou seja, além de intercepto e inclinação populacional (efeito fixo), mais um componente é descrito com variação no intercepto e inclinação individual (efeito aleatório). Portanto, a escolha de uma estrutura de covariância para os efeitos aleatórios é necessária para que se possa ser considerada uma correlação serial dentro do indivíduo, e assim, esta estrutura

de covariância e relação da variável resposta e um conjunto de covariáveis, pode ser modelada por meios dos modelos lineares mistos. É sabido que em dados longitudinais há uma potencial correlação serial entre as obsevações de um mesmo indivíduos e ignorar essa correlação levará a inferências espúrias sobre a média e sobre a relação entre a resposta e as covariáveis.

Quando é escolhido para a matriz de covariância, definida na expressão (2.3), $\sigma^2 I_{n_i}$, em que I_{n_i} é a matriz identidade com dimensão n_i, é dito que a estrutura de covariância é independente, ou seja, as medidas da variável resposta dos indivíduos são independentes com variâncias iguais como em modelos de regressão com dados independentes. Quando se tem estrutura diferente da encontrada em dados independentes, há outras estruturas de covariância que podem ser escolhidas e para cada tipo de correlação existente, há uma componente que pode ser mais adequada que a outra. A sensibilidade do pesquisador para identificar essa componente é essencial no momento da modelagem pois espera-se uma concordância biológica com o modelo estatístico . É útil dizer que todas as matrizes que representam uma estrutura de variância são matrizes simétricas e para fins didáticos são apresentados alguns exemplos de tais estruturas com matrizes de ordem 4×4,(XAVIER, 2000; FURTADO, 2009).

- Componente de variância, caracterizada por variâncias iguais e covariâncias iguais a zero, que é o caso quando há dados independentes.

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ & \sigma^2 & 0 & 0 \\ & & \sigma^2 & 0 \\ & & & \sigma^2 \end{bmatrix};$$

- Simétrica, caracterizada por variâncias homogêneas e covariâncias constantes.

$$\Sigma = \begin{bmatrix} (\sigma^2 + \sigma_1) & \sigma_1 & \sigma_1 & \sigma_1 \\ & (\sigma^2 + \sigma_1) & \sigma_1 & \sigma_1 \\ & & (\sigma^2 + \sigma_1) & \sigma_1 \\ & & & (\sigma^2 + \sigma_1) \end{bmatrix};$$

- Autorregressiva de primeira ordem ou $AR(1)$, apresenta variâncias homogêneas e correlações que diminuem exponencialmente à medida em que aumenta o intervalo de tempo entre as medidas repetidas. Neste caso o parâmetro autorregressivo é ρ e deve ser $\mid \rho \mid < 1$ para satisfazer propriedade de estacionaridade.

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix};$$

- Toeplitz, apresenta variâncias iguais e covariâncias variáveis a medida em que as distâncias entre os tempos crescem.

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma^2 & \sigma_{12} & \sigma_{13} \\ & & \sigma^2 & \sigma_{12} \\ & & & \sigma^2 \end{bmatrix};$$

- Autorregressiva heterogênea, caracteriza-se pela desigualdade de variâncias e covariâncias e pela maior correlação entre avaliações adjacentes.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ & & \sigma_3^2 & \sigma_2\sigma_4\rho \\ & & & \sigma_4^2 \end{bmatrix};$$

- Não-estruturada, tem todas as variâncias e covariâncias desiguais. Especifica uma matriz completamente geral.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{34} \\ & & & \sigma_4^2 \end{bmatrix};$$

Estas são estruturas de covariância mais utilizadas, porém existem outras mais disponíveis, como é visto em Pinheiro et al. (2014), West et al. (2014).

2.2 Modelo multi-estado

2.2.1 Análise de sobrevivência

A Análise de Sobrevivência, assim como a teoria de processos estocásticos, é a parte da Estatística que permitiu o desenvolvimento dos modelos multi-estado de Markov (ANDERSEN; KEIDING, 2002). Por esta razão é importante fundamentar os principais conceitos desta área de estudo, tais como censura, estado absorvente e modelo de riscos proporcionais que são imprescindíveis para o entendimento mais acurado acerca da teoria e aplicação dos modelos multi-estado bem como *joint model*.

A análise de sobrevivência é aplicada em áreas em que se deseja estudar um evento de interesse até seu tempo de ocorrência, ou seja, a variável resposta é o tempo decorrido até o aparecimento de algum evento, por exemplo, até a morte (ou cura) do paciente quando se trata de um fenômeno da área médica e é chamado de tempo de falha.

Segundo Giolo e Colosimo (2006), a principal característica de dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Uma observação é censurada

quando o paciente, por exemplo, teve o acompanhamento interrompido por alguma razão, mesmo quando este foi devido ao fim do estudo e o evento de interesse ainda não havia ocorrido. Assim como em modelos de regressão, o interesse é modelar a variável resposta em função de um grupo de covariáveis. A diferença essencial é o tempo-até-evento comumente referido como o tempo de sobrevivência o que muda a perspectiva de modelagem sendo necessário conhecer a forma da distribuição do tempo de sobrevivência. Muito frequentemente, essa distribuição pode ser a Exponencial ou Weibull.

Os dados de sobrevivência são caracterizados pelos tempos de falhas e pelas censuras (GIOLO; COLOSIMO, 2006). O tempo de falha é o tempo até a ocorrência do evento de interesse, e no estudo que será desenvolvido nesta dissertação, a falha será a morte do paciente. Dependendo do evento estudado a falha pode ter outras definições como em fenômenos de engenharia. Outro conceito importante é o da censura. Segundo Giolo e Colosimo (2006), censura ocorre quando há a presença de observações incompletas ou parciais e podem ocorrer pela impossibilidade de acompanhamento do paciente no decorrer do estudo. A não ocorrência do evento de interesse até o término do experimento, e a censura pode ser denotada ao se considerar, T uma variável aleatória representando o tempo de falha de um paciente C, uma outra variável aleatória independente de T, representando o tempo de censura do paciente, portanto,

$$t = \min(T, C)$$

$$\delta = \begin{cases} 1, & \text{se} \quad T \leq C, \\ 0, & \text{se} \quad T > C. \end{cases}$$

Ainda, como definido por Martins (2013), há algumas formas de censura, como à direita, esquerda e intervalar. Um tempo de sobrevivência é dito censurado à direita no instante C se o valor exato da observação do evento não é conhecido, mas apenas que este tempo é superior a C. Assim, o tempo, t, de observação de um indivíduo é uma observação da variável aleatória $T = \min(T, C)$. Da mesma forma, um tempo de sobrevivência é dito censurado à esquerda no instante C quando apenas se sabe que o valor dessa observação é inferior a C e neste caso o que se observa é $T = \max(T, C)$. Censura intervalar ocorre quando apenas se sabe que um tempo de sobrevivência pertence ao intervalo C_1, C_2.

Outro conceito importante é a função de sobrevivência que é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja,

$$S(t) = P(T \geq t), \tag{2.8}$$

e sua função de distribuição acumulada é o complementar da equação (2.8), isto é:

$$F(t) = P(T \leq t) = 1 - S(t). \tag{2.9}$$

É oportuno também definir a função de risco. A função de risco é amplamente usada para expressar o risco de morte em algum tempo t, e é definida como a probabilidade de falha durante um intervalo de tempo muito pequeno, supondo que o indivíduo tenha sobrevivido ao início do intervalo ou como o limite da probabilidade que um indivíduo morra em um intervalo muito curto, $t + \Delta t$, dado que o indivíduo sobreviveu no tempo t. Para facilitar a escrita da expressar, considere $\top$ sendo o evento que um indivíduo morra no intervalo de tempo $(t, t+\Delta t)$ dado que o indivíduo sobreviveu em t.

$$h(t) = \frac{\lim_{\Delta t \to 0} P[\top]}{\Delta t}. \tag{2.10}$$

A função de risco também pode ser definida em termos da distribuição acumulada $F(t)$ e da função densidade de probabilidade, $f(t)$:

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{2.11}$$

Por fim, será definido o modelo semiparamétrico de Cox, que também é conhecido com de riscos proporcionais e possibilita a inclusão de efeitos de covariáveis para o cálculo da probabilidade de ocorrer falha. Este modelo é chamado semiparamétrico, porque apenas os efeitos das covariáveis são tratados parametricamente e é mais um conceito importante proveniente de Análise de Sobrevivência para o entendimento dos modelos multi-estado.

A função de risco proporcional é dada da seguinte forma,

$$\lambda(t) = \lambda_0(t) \exp(x_i^\top \beta), \tag{2.12}$$

sendo λ_0 uma função não negativa denominada função de base.

O método de estimação é o da verossimilhança parcial,

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{exp(x_i^\top \beta)}{\sum_{j \in R(t_i)} exp(x_j^\top \beta)} \right\}^{\delta_i}, \tag{2.13}$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i e δ_i é o indicador de falha. Os valores de β, que maximizam a função de verossimilhança parcial, são obtidos por meio da resolução da função escore, $U(\beta)$, aplicando a função de logarítmo em 2.13 e

igualando a função escore a zero, que é o vetor de derivadas de primeira ordem de $\log(L(\beta))$, tem-se,

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \frac{\sum\limits_{j \in R(t_i)} x_j exp(x_j^\top \hat{\beta})}{\sum\limits_{j \in R(t_i)} exp(x_j^\top \hat{\beta})} \right\} = 0. \tag{2.14}$$

Mais detalhes sobre este método e outros conceitos de análise de sobrevivência podem ser vistos em Lee e Wang (2003), Giolo e Colosimo (2006), Carvalho et al. (2011) entre outros.

2.2.2 Processo estocástico

Um processo estocástico é uma coleção de variáveis aleatórias indexadas por um parâmetro de tempo,

$$X = \{X(t), t \in \tau\},$$

tal que, para cada $t \in \tau, X(t)$ é uma variável aleatória definida em um espaço de estados.

A definição de um processo estocástico está associada ao tempo e ao estado. Em relação ao tempo, ele pode ser discreto ou contínuo. Em relação ao estado, este também pode ser discreto ou contínuo. Quando discreto, temos uma cadeia de estados e quando contínuo $x(t)$ o estado é uma sequência.

Ainda podemos classificar os processos estocásticos em estacionários e independentes. É estacionário quando um processo mantém seu comportamento dinâmico invariante no tempo. Se os valores de $X(t)$ são independentes, isto é, o valor assumido por $X(t_j)$ não depende do valor assumido por $X(t_i)$ se $i \neq j$, então o processo é independente.

Uma definição importante na teoria de processos estocásticos e para os modelos multi-estado, é o processo de Markov. Processo de Markov nada mais é que um processo estocástico em que o próximo estado depende apenas do estado atual. Assim um processo markoviano, $X(t)$ é formalmente definido da seguinte maneira,

$$P[X(t_{k+1}) \leq x_{k+1} \mid X(t_k) = x_k, X(t_{k-1}) = x_{k-1}, ..., X(t_1) = x_1, X(t_0) = x_0] =$$
$$P[X(t_{k+1}) \leq x_{k+1} \mid X(t_k) = x_k],$$

para todo $t_0 \leq t_1 \leq \ldots t_k \leq t_{k+1}$.

Portanto, o processo definido acima é um processo estocástico estacionário com pressuposto markoviano ou ainda falta de memória. Podemos resumir essa ideia dizendo que

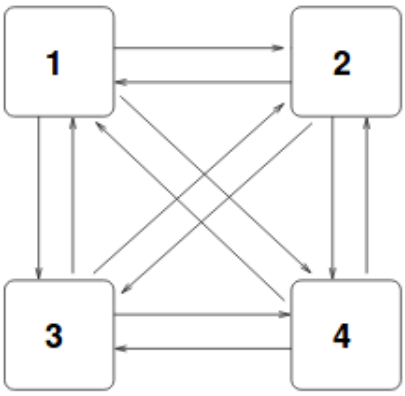
para um processo de Markov as informações de estados passados e o tempo que o processo permanece no estado atual, são irrelevantes. Sabe-se igualmente, pela teoria de probabilidade, que a distribuição exponencial, similarmente, possui a característica da falta de memória. Dado esta similaridade, a distribuição dos tempos entre eventos de uma cadeia de Markov, segue uma distribuição exponencial. Essas características fazem deste tipo de processo estocástico mais utilizado por ser de fácil interpretação.

Nesta dissertação será aplicado o modelo multi-estado em tempo contínuo e portanto, define-se uma cadeia de Markov em tempo contínuo, ou seja, $X(t), t \geq 0$. Como a variável tempo é contínua, esta representa instantes ou momentos do tempo e não períodos no tempo, como no caso do tempo discreto. Neste contexto o espaço de estados S é finito e enumerável, a probabilidade de o processo estar no estado j num momento futuro depende apenas do estado presente e não dos estados visitados em qualquer momento passado,

$$P_{ij}(t) = P[X(t+s) = j \mid X(s) = i].$$

Logo, para completar uma Cadeia de Markov em tempo contínuo deve-se definir seus estados e as probabilidades de transição entre os estados em um instante.

Pode-se representar uma cadeia de Markov pelo diagrama de transições ou pela matriz de transição, sendo que esta matriz Q apresenta as intensidades de transição em um período.



$$Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{41} & q_{42} & q_{43} & q_{44} \end{pmatrix}$$

Figura 1 – Representação - Cadeia de Markov

Na Figura 1, além da matriz de transição Q, há a representação da transição entre os estados que uma doença ou fenômeno pode atingir. Esta matriz de transição é geral, ou seja, considera que todas as transições são possíveis e ainda não determina um estado absorvente.

As transições são descritas seguidamente no respectivo diagrama. A transição $1 \to 2$ corresponde a transição do estado 1 para o estado 2, $1 \to 3$ a transição do estado 1 para o estado 3 e assim por diante até a transição do estado 3 para o estado 4, enquanto a transição contrária pode ocorrer também, ou seja, $2 \to 1$ corresponde à volta do estado 2 para o estado 1. É possível, ainda, que em um próxima obervação, o indivíduo permanecer no mesmo estado que havia sido observado anteriormente e na matriz Q, a intensidade dessa transição é representada pelos parâmetros, $q_{11}, q_{22}, q_{33}, q_{44}$. E os demais parâmetros, $q_{12}, ..., q_{21}, ..., q_{34}, ..., q_{43},$, representam as intensidades das demais transições.

Como um processo multi-estado é um processo estocástico, nesta seção, alguns processos estocásticos são definidos,(SEN et al., 2010), com o intuito do leitor entender em que contexto da teoria de processos estocásticos o processo multi-estado se encaixa.

- Série temporal com tempo discreto

Considere o exemplo ilustrativo sobre o registro da temperatura máxima diária em uma região num período de tempo com o objetivo de verificar a existência de tendência linear e sazonalidade na série. Um possível modelo para a temperatura máxima diária é dado da seguinte maneira,

$$Y_t = \alpha + \beta_t + e_t, \tag{2.15}$$

em que,

- Y_t : Temperatura máxima diária em uma região em um período de tempo.
- α e β são constantes desconhecidas.
- $e_t = \rho e_{t-1} + \mu_t$, sendo que $\mid \rho \mid < 1$, $\mu_t \sim N(0, \sigma^2)$ e $t = 1, ..., n$.

Nota-se que a coleção de variáveis aleatórias, $Y = \{Y_t, t = 1, ..., n\}$ são dependentes e formam um processo estocástico chamado de série temporal com tempo discreto.

- Processo estocástico espacial, com parâmetro espacial discreto

Neste segundo exemplo, suponha que é observado a mortalidade ou morbidade registrados em um país dado um período de tempo com o objetivo de avaliar a distribuição espacial da prevalência do fenômeno considerado, denotado por,

$$Y_s = \mu_s + e_s, s = (s_1, s_2)^t \in S \subset \mathbb{R}^2, \tag{2.16}$$

em que,

Y_s: Contagem de mortalidade observada para um país com coordenada s.
μ_s: vetor esperado da prevalência.
e_s: termo do erro aleatório.

Nesta especificação, a medida que a distância entre s_1 e s_2 diminui, a correlação, Cor(e_{s_1}, e_{s_2}), aumenta.

Aqui, $Y = \{Y_s, S \subset \mathbb{R}^2\}$, também são dependentes e este processo estocástico é chamado de processo estocástico espacial com parâmetro espacial discreto.

- Processo estocástico espaço-temporal

Considere agora, para uma área metropolitana de um número de municípios, registros do número diário de acidentes automobilísticos, $N(s,t)$ em um período de tempo. Este fenômeno pode ser modelado usando as definições dos exemplos 2.2.2 e 2.2.2 e denotado como $N = \{N(s,t); s \in S, t \in \tau\}$ onde,

S é o conjunto de municípios e τ é o conjunto de índices do tempo.
Y é Contagem de acidentes de carros registrados diariamente em área urbana.

Dessa forma, N constitui um processo estocástico espaço-temporal.

- Processo da morte com parâmetro contínuo

Neste exemplo, suponha que um pesticida é lançado em uma colônia com n insetos para avaliar sua efetividade, ou seja, avaliar o processo de morte dos insetos. Seja, n_t número de insetos vivos no tempo $t > 0$. Pode-se notar que n_t jamais será incrementado e n_t e n_r, $r \neq t$, são variáveis aleatórias dependentes. Resumindo,

- n_t é número de insetos vivos no tempo $t > 0$.
- n_t não incrementa, ou seja, será constante ou decrescente.
- n_t e n_r, $r \neq t, \Longrightarrow$ são variáveis aleatórias dependentes.

Portanto, este processo estocástico é conhecido comumente como processo da morte com parâmetro temporal contínuo.

- Processo de soma parcial

Suponha, que $X_n, n \geq 1$, seja uma amostra aleatória independente e identicamente distribuída,com média μ e variância finita σ^2 e seja $S_n = X_1 + ... + X_n, \forall_n \geq 1$.Então o processo estocástico $\{S_n; n \geq 0\}$, é conhecido como processo de soma parcial e possui propriedades de dependência markoviana, ou seja, podemos resumir com a seguinte expressão:

$$P(S_{n+1} = K \mid S_n, ..., S_0) = P(S_{n+1} = k \mid S_n), \forall k, n \geq 1. \tag{2.17}$$

- Processo multi-estado

Finalmente, considere o exemplo de uma amostra de trabalhadores de uma fábrica de tabaco e os estados,

1. A_1 é o estado em que o trabalhador está livre de doença respiratória.
2. A_2 é o estado em que o trabalhador está com possibilidade de infectado mas não diagnosticado.
3. A_3 é o estado em que o trabalhador está com problema respiratório diagnosticado.
4. A_4 é o estado em que o trabalhador está com necessidade de cuidados ambulatoriais.
5. A_5 é o estado de morte do trabalhador devido a problemas respiratório.
6. A_6 é o estado em que o trabalhador saiu da fábrica ou morte por outras causas.

Seja X_n representando um dos estados de saúde dos trabalhadores, descritos acima, no mês n e assuma que $X_0 = A_1$. Então, $X_n = A_i, (i_n = 1, ..., 6)$, X_{n+1} pode permanecer em X_n ou transitar para um dos outros estados, teremos 36 possibilidades de transição e é representado da seguinte maneira,

$$A_{i_n} \Longrightarrow A_{i_{n+1}}, \qquad i_n, i_{n+1} = 1, ..., 6,$$

e os estados A_5 e A_6, são os estados absorventes.

Este é um exemplo de processo estocástico multi-estado com dois estados absorventes e com isto pode-se definir o que é um modelo multi-estado.

2.2.3 Modelo multi-estado de Markov

Modelo multi estado (em inglês *Multi State Model - MSM*), é visto como um processo estocástico que descreve como um indivíduo se move entre uma série de estados em algum instante no tempo e nos últimos anos vem sendo muito utilizado para monitorar progressões de doenças e estimar as probabilidades de transição entre os estágios das doenças (JACKSON et al., 2003; BEYERSMANN et al., 2011; HUSZTI et al., 2012).

Por isso são utilizados na área médica para estimar as probabilidades de transição entre vários estágios da doença, como por exemplo, os estágio de câncer, diabetes, doenças renais, HIV ou outros tipos de doenças que são caracterizadas por vários estados ou estágios de progressão da doença.

A intensidade de transição entre os estados pode ser representada com a seguinte expressão,

$$q_{rs}(t, z(t)) = \lim_{\delta t \to 0} P(S(t + \delta t) = s \mid S(t) = r)/\delta t, \tag{2.18}$$

em que $z(t)$ é o conjunto de variáveis explicativas.

Na Figura 2 há a representação de uma matriz de transição para um modelo geral para progressão de doença com um estado absorvente, que é o tipo de processo estocástico que interessa nesta dissertação. Nota-se que para a transição entre os estados, o paciente precisa passar necessariamente pelo estado conseguinte, ou seja, não há a possibilidade de estar no estado 1 e ir diretamente para o estado 3, ele passará antes pelo estado 2, com a possibilidade e retornar ao seu estado anterior. Uma vez que o paciente chega no último estado da doença ele não mais transitará entre outros estados, por essa razão dá-se o nome de estado absorvente.

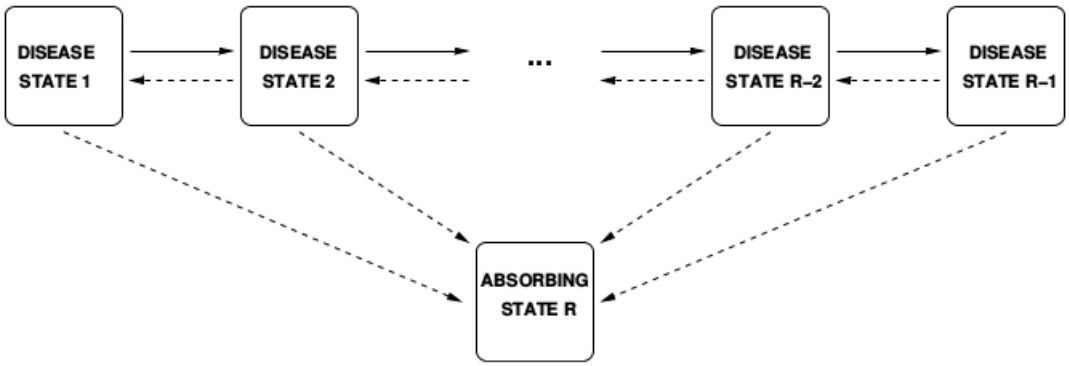
2.2.3.1 Função de verossimilhança

Será descrito nessa seção, um resumo da função de verossimilhança para um modelo multi-estado em tempo contínuo.

A verossimilhança é calculada a partir da matriz de probabilidades de transição $P(t)$. Para um processo homogêneo no tempo, o elemento (r, s) de $P(t)$ é a probabilidade de estar no estado s no tempo $t + u$ no futuro, dado que o estado no tempo t é r.

A série de tempos $(t_{i1}, t_{i2}, ..., t_{ij})$ e os correspondentes estados $(S_i(t_{i1}), S_i(t_{i2}), ..., S_i(t_{ij}))$ são os dados para o indivíduo i no tempo j. Considere um modelo multi-estado geral, com um par de sucessivos estados observados da doença $S(t_j), S(t_{j+1})$ nos tempos t_j, t_{j+1}. A contribuição deste par de estados para a verossimilhança é,

$$L(Q) = \prod_i L_i = \prod_{ij} L_{i,j} = P_{S(t_{ij})S(t_{i,j+1})}(t_{i,j+1} - t_{ij}). \tag{2.19}$$



$$Q = \begin{pmatrix} q_{11} & q_{12} & 0 & 0 & \dots & q_{1R} \\ q_{21} & q_{22} & q_{23} & 0 & \dots & q_{2R} \\ 0 & q_{32} & q_{33} & q_{34} & \ddots & q_{3R} \\ 0 & 0 & q_{43} & q_{44} & \ddots & q_{4R} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

Figura 2 – Modelo multi-estado

Cada $L_{i,j}$ é a entrada da matriz $P(t)$ na linha $S(t_{ij})$ e na coluna $S(t_{i,j+1})$ avaliado no tempo, $t_{i,j+1} - t_{ij}$.

Usando um algoritmo de otimização, a log- verossimilhança é maximizada e as estimativas de $q_{r,s}$ são obtidas, assim como seus erros padrão por meio da matriz hessiana.

2.3 *Joint model*

Geralmente as modelagens de dados longitudinais e de sobrevivência são feitas separadamente, segundo Wu et al. (2011), tais dados contém, na prática, ambas informações e analisá-los separadamente pode levar a resultados ineficientes e tendenciosos. Isso pode ocorrer se o objetivo for explicar, concomitantemente, a influência de variáveis explicativas na remissão de uma doença ao longo do tempo, e ainda, qual o risco de morte considerando o tempo entre diagnóstico da doença e morte do paciente.

Com a metodologia do *joint model* todas essas informações podem ser incorporadas em um único modelo e as inferências serão mais eficientes. Segundo Martins (2013), o fato de existirem processos individuais latentes que variam ao longo do tempo, isto é, que são inerentes ao indivíduo, e que contribuem, tanto para o modelo longitudinal, como para o de sobrevivência, justifica a modelagem conjunta. Este modelo permite considerar, simultaneamente, as possíveis correlações entre as medidas repetidas em um indivíduo e o seu tempo de sobrevivência.

O *joint model*, inicialmente surgiu com o intuito de estudar a relação entre a resposta de um processo longitudinal e a resposta de um processo que leva em consideração o tempo até a ocorrência do evento de interesse (HENDERSON et al., 2000; TSIATIS; DAVIDIAN, 2004). Na literatura, há alguns modelos propostos para a modelagem conjunta entre biomarcadores longitudinais e modelos para dados de sobrevivência como o de riscos proporcionais (ELASHOFF et al., 2008) e este tipo de modelagem continua sendo desenvolvida e aplicada (PROUST-LIMA; TAYLOR, 2009; WU et al., 2011; RIZOPOULOS, 2011). Há também estudos sendo publicados com o enfoque Bayesiano (MARTINS et al., 2016) e quando a distribuição de probabilidade atribuída para a variável resposta é diferente da gaussiana (JUAREZ-COLUNGA et al., 2017) entre outros.

No entanto, ainda há poucos modelos propostos combinando modelos multi-estado de Markov e modelos mistos para dados longitudinais (LANGE et al., 2015), com a finalidade de estudar os vínculos entre a evolução dos biomarcadores e o risco de transição entre os diferentes estados da doença e esta é a proposta central desta dissertação.

A especificação do *joint model* será feita seguindo o estudo de Ferrer et al. (2016), que o aplicou em uma modelagem conjunta entre um processo longitudinal e um processo multi-estado com riscos proporcionais em dados clínicos de uma amostra real de pacientes com câncer de próstata.

Para o modelo linear misto para dados longitudinais considere a expressão em (2.1) e apenas para facilitar será reescrito da seguinte maneira,

$$Y_{ij} = Y_i^*(t) + \epsilon_{ij}. \tag{2.20}$$

Para o modelo multi-estado com riscos proporcionais, para modelar as transições entre os estados r para s, considere a equação (2.18) e reescrevendo tem-se a intensidade no tempo t,

$$\begin{aligned} \lambda_{rs}^i(t \mid b_i) &= \lim_{\delta t \to 0} P(S(t+\delta t) = s \mid S(t) = r; b_i)/\delta t \\ &= \lambda_{rs,0}(t) exp(X_{rs,i}^S \beta_{rs} + W_{rs,i}(b_i, t)^\top \eta_{rs}). \end{aligned} \tag{2.21}$$

Nota-se que na primeira linha na expressão (2.21), a nova expressão leva em conta os efeitos aleatórios b_i que serão compartilhados entre os dois submodelos. Ainda $\lambda_{rs,0}(.)$ é a intensidade de referência paramétrica, que no caso será a B-splines, e $X_{rs,i}^S$ é o vetor de covariáveis associado ao r-vetor de coeficientes β_{rs}. A função multivariada $W_{rs,i}(b_i, t)$, define a estrutura de dependência entre os dois submodelos, o longitudinal e o multi-estado. A função multivariada $W_{rs,i}(b_i, t) = Y_i^*(t)$, pode ter apenas efeito aleatório no intercepto ou na inclinação ou em ambos. Portanto, o s-vetor de coeficientes η_{rs} quantifica o impacto longitudinal na intensidade da transição entre os estados r e s.

Desta forma, este modelo permite, por exemplo, explicar a ligação entre as medidas

de um biomarcador ao longo tempo e o tempo de recorrência clínica, distinguindo as várias transições entre os estados de saúde, ao estudar o risco de morte. O impacto, para o risco de morte, de covariáveis na trajetória do biomarcador, após a transição entre um desses estados, também é incorporado no modelo.

2.3.1 Função de verossimilhança

O método de estimação para a modelagem conjunta, adotado por Ferrer et al. (2016), é o da máxima verossimilhança e é definido da seguinte forma,

$$L(\theta) = \prod_{i=1}^{N} \int_{\mathbb{R}} f_Y(Y_i \mid b_i; \theta) f_S(S_i \mid b_i; \theta) f_b(b_i; \theta) db_i, \tag{2.22}$$

em que θ é o vetor de parâmetros especificados nas equações, (2.1) e (2.21) e $f(.)$ é uma função densidade de probabilidade. A função $f_Y(Y_i \mid b_i; \theta)$, se refere ao modelo misto. Portanto, considera-se, para o caso gaussiano multivariado, o modelo linear misto pela expressão que segue,

$$f_Y = (Y_i \mid b_i; \theta) = \frac{1}{(2\pi\sigma^2)^{n_i/2}} exp\left\{-\frac{||Y_i - X_i^\top \beta - Z_i^\top b_i||}{2\sigma^2}\right\}, \tag{2.23}$$

onde $||x||$ é a norma euclidiana do vetor x e os outros elementos da expressão (2.23), são equivalentes aos descritos na expressão (2.1).

Para o modelo multi-estado, considere a definição dada na Seção 2.2.3.1 e a expressão (2.19) e, ainda, considerando a intensidade entre as transições como $\lambda^i_{S(t_{ij}),S(t_{i,j+1})}(t_{i,j+1} \mid b_i)$, então a contribuição individual para a verossimilhança ficará,

$$\begin{aligned} f_S(S_i \mid b_i; \theta) &= \prod_{j=0}^{m_i-1} \left\{ P_{S(t_{ij}),S(t_{ij})}(t_{ij}, t_{i,j+1} \mid b_i) \lambda^i_{S(t_{ij}),S(t_{i,j+1})}(t_{i,j+1} \mid b_i)^{\delta_{i,j+1}} \right\} \\ &= \prod_{j=0}^{m_i-1} \left\{ exp\left[\int_{t_{ij}}^{t_{i,j+1}} \lambda^i_{S(t_{ij}),S(t_{ij})}(u \mid b_i) du \right] \lambda^i_{S(t_{ij}),S(t_{i,j+1})}(t_{i,j+1} \mid b_i)^{\delta_{i,j+1}} \right\}. \end{aligned} \tag{2.24}$$

Ligando os dois submodelos, longitudinal e multi-estado, os efeitos aleatórios b_i, seguem uma distribuição gaussiana multivariada, dada por,

$$f_b(b_i; \theta) = \frac{1}{(2\pi^{q/2}) det(G)^{1/2}} exp\left\{-\frac{b_i^\top G^{-1} b_i}{2}\right\}, \tag{2.25}$$

em que G é a matriz de covariância dos efeitos aleatórios definido em 2.1.

Com o intuito de encontrar as estimativas de máxima verossimilhança os autores, Self

e Pawitan (1992), propuseram a estimação do *joint model* considerando dois estágios. No primeiro estágio é estimada a parte fixa e predito os efeitos aleatórios. No segundo estágio essas estimativas são usadas para imputar um valor apropriado para $Y_i^*(t)$ que é usado para definir a função multivariada $W_{rs,i}(b_i, t)$. Esse procedimento se inicia com o algoritmo EM para um número fixo de iterações até que haja convergência. Se não houver convergência é utilizado o algoritmo de Quase-Newton.

Além disso, as integrações em 2.22 e 2.24 não apresentam solução analítica, por essa razão, algoritmos de integração e otimização são utilizados. A função desenvolvida por Ferrer et al. (2016), que está disponível no R, $JM::JMstateModel()$, realiza essa integração numérica. As integrais para o modelo multi-estado são aproximadas usando quadraturas Gauss-Kronrod, e as integrais sobre os efeitos aleatórios usando quadrados pseudo-adaptativos Gauss-Hermite. Mais detalhes sobre o procedimento de otimização, o algoritmo EM e as integrações numéricas são encontrados em Rizopoulos (2012) e Ferrer et al. (2016).

2.3.2 Otimização

O procedimento de otimização descrito no parágrafo acima, requer em muitos casos, um controle dos parâmetros de convergência por parte do analista. Assim como definido por Rizopoulos (2012), este procedimento começa com um número fixo de iterações do algoritmo *EM*, se a convergência não for alcançada, o processo muda para o algoritmo *Quase-Newton* até que a convergência seja atingida. Evidentemente que os valores iniciais dos parâmetros para o *Joint model*, são retirados dos submodelos longitudinal e multi-estado para estabelecer o início do ajuste. Durante as iterações do algoritmo *EM*, a convergência é definida quando um dos dois seguintes critérios é satisfeito:

$$\max\{\mid \theta^{(it)} - \theta^{it-1} \mid /(\mid \theta^{it-1} \mid +\epsilon_1)\} < \epsilon_2,$$
$$\ell(\theta^{it}) - \ell(\theta^{it-1}) < \epsilon_3\{\mid \ell(\theta^{it-1}) \mid +\epsilon_3\},$$

em que θ^{it} denota os valores dos parâmetros na itth iteração, $\ell(\theta) = \sum_i \log p(T_i, \delta_i, y_i; \theta)$.

Quando é necessário utilizar as iterações pelo algoritmo de *Quase-Newton*, apenas o último critério é usado.

2.3.3 Integração numérica

Segundo Rizopoulos (2012), uma dificuldade computacional para o ajuste de um *Joint model*, ocorre devido a integral com respeito ao tempo na definição da função de sobrevivência. Analogamente, para o caso de um submodelo multi-estado, esta dificuldade também surge

com relação a integral definida em 2.24 assim como a integral dos efeitos aleatórios definidas em 2.22 e 2.25. Este recurso de otimização numérica, faz com que o ajuste de modelos da classe *Joint model*, seja uma tarefa computacionalmente intensiva.

A aproximação das integrais sobre efeitos aleatórios, é feita pela quadratura de Gauss-Hermite e para aproximar a integral sobre tempo do processo multi-estado, usa-se quadratura de Gauss-Kronrod. Mais detalhes sobre a estas integrações numéricas, podem ser encontrados em Rizopoulos (2012).

2.3.4 Problemas de convergência

Devido as características de otimização e integração numérica, o ajuste de um *Joint model*, se torna computacionalmente intensivo e isso pode implicar problemas de convergência. Segundo Rizopoulos (2012), apesar da função que ajusta o *Joint model* no *R*, permitir controlar a otimização dos algoritmos de convergência e suas quantidades de iterações e ainda os números de pontos das quadraturas de inetgraçção numéricas e a tolerância de convergência para ambos os casos, a plena convergência do modelo pode não ser garantida para todos os conjuntos de dados. Portanto, ao ajustar um *Joint model* a intuição geral sobre o procedimento de otimização e integração numérica, alterando adequadamente valores dos argumentos de controle destes procedimentos, é indispensável para o sucesso no ajuste com conjuntos de dados mais "difíceis". Na maioria dos casos, as mudanças mais úteis estão nos valores que indicam o número de iterações do algoritmos EM e $Quase - Newton$, número de pontos paras as funções de integração numérica e o tipo de derivada numérica (aproximação *foward* ou *central difference*) que calcula a matriz Hessiana baseada na função *score*. Mais detalhes sobre estas parametrizações, também podem ser encontrados em Rizopoulos (2012).

2.3.5 Testes da Razão da verossimilhança e de *Wald*

Como descrito na seção 2.3.1, os parâmetros em um *Joint model* são estimados pela verossimilhança e por esta razão inferências com o teste da razão da verossimilhança podem ser feitas.

Considere o interesse em testar a seguinte hipótese nula:

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta \neq \theta_0,$$

Então, aplica-se o teste da razão da verossimilhança para testar a hipótese nula declarada acima,

$$TRV = -2\{\ell(\hat{\theta}_0) - \ell(\hat{\theta})\},$$

em que $\hat{\theta}_0$ e $\hat{\theta}$ são as estimativas de máxima verossimilhança sobre a hipótese nula e alternativa respectivamente. E ainda sob a hipótese nula, aplica-se o teste de *Wald* com estatística de teste definida como,

$$W = (\hat{\theta} - \theta_0)^\top \tau(\hat{\theta})(\hat{\theta} - \theta_0).$$

Sob a hipótese nula, a distribuição assintótica para estes testes é a *Qui-quadrado* com p graus de liberdade, com p sendo o número de parâmetros testados. Se o interesse fosse testar apenas um parâmetro, o teste de $Wald$ seria equivalente a $(\hat{\theta}_j - \theta_{0j})/e\hat{.}p.(\hat{\theta}_j)$, no qual, sob a hipótese nula segue, assintoticamente, uma distribuição Normal padrão.

Essas estatísticas de teste são assintoticamente equivalentes. Contudo, na prática, quando estamos lidando com amostras finitas, elas geralmente diferem. Neste caso, o teste de razão de verossimilhança é geralmente considerado o mais confiável e o teste de $Wald$ o menos confiável. O teste de $Wald$ exige o ajuste do modelo apenas sob hipóteses nulas e alternativas, respectivamente, enquanto que o teste de razão de verossimilhança requer o ajuste do modelo sob ambas as hipóteses, e, portanto, é computacionalmente um pouco mais intensivo (RIZOPOULOS, 2012).

Para realizar o teste da razão da verossimilhança e teste de $Wald$ para um *Joint model* ajustado no *R*, basta usar as funções $anova()$ e $summary()$, assim como, por exemplo, em outros modelos lineares de regressão ajustados no *R*. Ao definir na função $anova()$ para o teste de Wald, o nome do modelo e o processo longitudinal (com o argumento, *process="Longitudinal")*, a saída será a estatística de teste *chi-quadrado* e seus respectivos p-valores, para cada parâmetro estimado para os efeitos fixos. O que corresponde testar,

$$H_0 : \beta_j = 0$$

versus

$$H_1 : \beta_j \neq 0,$$

em que, β_j, são os parâmetros estimados para cada efeito fixo do modelo longitudinal.

Para verificar se as covariáveis no submodelo multi-estado contribuem na expicação da variabilidade no risco de morte para cada transição, pode-se testar a hipótese nula equivalentemente como no modelo longitudinal, a diferença é que deve-se especificar na função $anova()$ o argumento *process="Event"*, e o resultado disponibilizará a estatística *Qui-quadrado* e seu respectivo p-valor. Ao utilizar a função $summary()$, basta fornecer o nome do modelo e todos

os resultados dos testes hipóteses para todos os parâmetros, tanto para o modelo longitudinal quanto para o modelo multi-estado, serão apresentados na saída do comando.

Um problema com o teste de *Wald* para testar os efeitos fixos no clássico modelo linear misto é que ele é baseado em erros padrão que subestimam a verdadeira variabilidade em β porque não leva em consideração a variabilidade introduzida ao estimar as componentes de variância. Por esse motivo, tipicamente uma distribuição aproximada F com graus de liberdade apropriados é usada em lugar da distribuição *Qui-quadrado*. Segundo Rizopoulos (2012), este problema pode ser exagerado para os modelos da classe *Joint model*, pois não é só ignorado o fato de que é estimada as componentes da variância, mas também que é necessário estimar o processo multi-estado. Portanto, geralmente é aconselhável aplicar o teste da razão da verossimilhança.

Assim como o teste de *Wald*, o TRV é aplicável com o uso da função $anova()$ e para isto basta informar os dois modelo aos quais se deseja comparar, ressaltando que o primeiro é sempre o modelo sob a hipótese nula, ou seja, o modelo sem a covariável. A saída do comando disponibilizará informações como os critérios de informação de ajuste de AIC e BIC (AKAIKE, 1974; EMILIANO et al., 2010), o $\log$ da verossimilhança dos modelos, a estatística TRV, graus de liberdade e o p-valor. Desta forma será possível inferir sobre a contribuição da covariável no modelo. Importante salientar que ambos os testes aqui discutidos são apropriados para comparar dois modelos aninhados, já que o modelo sob a hipótses nula é um caso especial do modelo sob a hipótese alternativa. Se houver interesse em comparar dois modelos não aninhados, pode-se utilizar os critérios de informação AIC e BIC, por exemplo.

2.3.6 Diagnóstico de ajuste

É sabido que, na prática, ao se escolher um modelo dentre alguns possíveis candidatos, é necessário verificar a qualidade do ajuste dos modelos. Para modelos da classe do *joint model* esta tarefa, assim como outros modelos estatísticos, pode ser feito por meio de análise gráfica dos resíduos. Em Rizopoulos (2010), é definido como esta análise de resíduos pode ser obtida para um *joint model*.

Os possíveis gráficos de resíduos são, os já consagrados, valores preditos versus resíduos padronizados ou o gráfico de probabilidade normal dos resíduos. Tais gráficos são facilmente e diretamente disponibilizados com uma função básica do R, $plot()$, que também se aplica a modelos da classe *joint model*, e suas definições são encontradas em vários livros que abordam a o ajuste de modelos de regressão tais como, Sheather (2009), Montgomery et al. (2015), Faraway (2016) entre outros. Esses resíduos são aplicáveis ao submodelo longitudinal. Para o submodelo multi-estado, pode-se aplicar os tipos de resíduos utilizados em diagnóstico de ajuste para modelos de sobrevivência, como por exemplo, os resíduos de Martingale. Sua definição, e outros tipos de resíduos para modelos de sobrevivência, podem ser encontrados em Carvalho et al. (2011).

Não obstante a isso, ainda segundo, Rizopoulos (2010), embora as análises de resíduos

dos submodelos possam ser usadas separadamente para diagnosticar o ajuste de um modelo *joint model*, ocorre um problema com a distribuição de referência dos resíduos para o processo longitudinal. Esta é afetada pela perda de informações quando o indivíduo alcança seu estado absorvente ou há uma censura e este não pode mais ser observado até o último tempo, assim sua distribuição muda após o evento e os resíduos são calculados não mais de forma aleatória, como é justificado em Kenward e Molenberghs (1998). Isso implica que os resíduos não constituem mais uma amostra aleatória da população alvo, então, espera-se que não apresentem mais média zero e independência.

Para contornar este problema, Rizopoulos (2010), propôs aumentar os dados observados com imputação aleatória das respostas longitudinais sob o modelo de dados completos, criando uma correspondência aos dados como se não tivesse dados de indivíduos não observados antes do tempo final. Usando essas respostas longitudinais imputadas, os resíduos então calculados para os dados completos, e um gráfico de valores ajustados versus resíduos padronizados imputados é produzido.

Para o modelo ajustado nesta dissertação, não houve a necessidade de realizar esta técnica de imputação de dados para a análise de resíduos pois a amostra final selecionada, todos os indivíduos possuiam o mesmo número de obervações do nível do biomarcador, como é justificado por Rizopoulos (2012).

Capítulo 3

Ajuste do *Joint model*

3.1 Apresentação dos Dados

Para o ajuste do modelo proposto nesta dissertação, foram utilizados dados sobre câncer de próstata oriundos da Finlândia, e que fazem parte de um estudo denominado *European Randomized Study of Screening for Prostate Cancer (ERSPC)* (FINNE et al., 2003). A amostra é composta por indivíduos que nasceram na finlândia entre 1929 e 1944, de acordo com *Population Registry of Finland*, e vivem nas áreas metropolitanas das cidades de Helsinki e Tampere. Homens com incidência de câncer de próstata foram identificados por meio dos registros do *Finnish Cancer Registry*. Os dados foram, gentilmente, disponibilizados pelo professor epidemilogista e *PhD* Anssi Auvinen, da *Faculty of Social Sciences*, *University of Tampere*.

Todos os anos, entre 1996 e 1999, os indivíduos eram aleatoriamente selecionados. A seleção foi realizada em primeiro de março de 1996 e primeiro de janeiro nos outros anos, 1997 a 1999. Homens com 55, 59, 63 e 67 anos foram amostrados anualmente. Por exemplo, em 1996 foram selecionados do *Finnish Cancer Registry*, homens que nasceram em 1929, 1933, 1937 e 1941; em 1997, as seleções eram de indivíduos nascidos em 1930, 1934, 1938 e 1942; 1998 os anos de nascimentos considerados foram, 1931, 1935, 1939 e 1943; e por fim, no ano de 1999, a seleção dos participantes foi realizada com aqueles que nasceram em 1932, 1936, 1940 e 1944.

Como é visto na Tabela 1, para cada ano foram selecionados 8000 indivíduos. Totalizando 32000 homens. Porém, apenas 30403 estavam disponíveis no momento do convite para fazerem parte da seleção. Os motivos de indisponibilidade foram morte, mudança de endereço para fora da área de estudo ou recusa. Cada um dos participantes teve o nível do biomarcador do câncer de próstata medido em até três vezes. Os que, por alguma razão, não puderam ter as três medidas coletadas foram excluídos da amostra para o ajuste do modelo, desta forma a amostra final foi composta por 808 indivíduos. Estas exclusões foram necessarias para convergência

do modelo.. Este biomarcador é a medida do antígeno prostático específico, que em inglês é *Prostate Specific Antigen* com a sigla *PSA*. Segundo Slawin et al. (1995) e Sobreiro (2013), *PSA* é uma proteína sérica, produzida pelo epitélio prostático e pelas glândulas periuretrais do homem, cuja função é liquefazer o coágulo seminal e principal marcador do câncer de próstata.

Tabela 1 – Número de indivíduos selecionados

	Ano de seleção			
Ano de nascimento	1996	1997	1998	1999
1929	1505			
1930		1656		
1931			1559	
1932				1476
1933	1616			
1934		1778		
1935			1754	
1936				1783
1937	1981			
1938		2301		
1939			2098	
1940				1998
1941	2898			
1942		2265		
1943			2589	
1944				2743
Total	8000	8000	8000	8000
Disponíveis	7337	7699	7696	7671

Além do *PSA*, outras variáveis foram observadas:

- A idade do paciente no momento da seleção - (55, 59, 63 e 67 anos);
- O tempo em que ele permaneceu no estudo;
- Se havia histórico de pessoas com câncer de próstata na família do participante;
- O tempo em que cada uma das três medidas de *PSA* foram coletadas após o início do estudo;
- A extensão do tumor - localizado (órgão confinado) - regional (estendendo-se para a pelve fora da próstata ou nódulos linfáticos) - metastático (disseminação distante, mais frequentemente nos ossos);
- Se o indivíduo morreu ou não e se morreu, quanto tempo após o início do estudo.

Os dados foram disponibilizados sem nenhum tratamento da forma como foram coletados e para o ajuste do modelo, foi necessário preparar dois conjuntos de dados. O primeiro foi disposto no formato longo para, obviamente, ser aplicado ao submodelo longitudinal. Na Tabela 2 são apresentadas as 6 primeiras linhas deste conjunto de dados no formato longo, onde ID é o código identificador do paciente, *age* é a idade do paciente, *Diag_time* é tempo de permanência na seleção do estudo, *PC_TNM* indica a extensão do tumor, *famhist1* indica se havia histórico de câncer de próstata na família, *psa* indica a ordem da medida do biomarcador (psa1 - primeira medida, psa2 - segunda medida e psa3 - terceira medida), *value* é o valor do biomarcador *PSA* na escala original, *times* é o tempo que levou para a coleta do *PSA* desde no início da seleção do indivíduo, *Death* indica se o indivíduo atingiu o estado absorvente de morte e *Death_time* é o tempo que levou até a morte. Quando estas duas últimas variáveis apresentam valores faltantes, significa que o indivíduo não morreu. Todas as variáveis que se referem a tempo, estão medidas em anos.

Tabela 2 – Dados no formato longitudinal

ID	age	Diag_time	PC_TNM	famhist1	Death_time	psa	value	times	Death
1	55	10.77	metastatico	1		psa1	1.39	0.33	
1	55	10.77	metastatico	1		psa2	1.32	4.19	
1	55	10.77	metastatico	1		psa3	1.68	8.21	
2	55	10.32	localizado	0		psa1	1.47	0.34	
2	55	10.32	localizado	0		psa2	1.54	4.15	
2	55	10.32	localizado	0		psa3	5.05	8.19	

O segundo conjunto de dados foi preparado em formato multi-estado e posteriormente ajuste deste submodelo. Os estados da doença foram definidos segundo a Figura 3, ou seja, o estado inicial foi chamado de 1 ou estado normal, estado da doença 2 ou estado anormal e o estado absorvente, que é a morte, foi chamado de 3. As transições foram formatadas da seguinte maneira. A transição 1 é a saída do indivíduo do estado um para o estado dois, ou seja, $1 \rightarrow 2$, transição 2, saída do indivíduo do estado um para o estado três, $1 \rightarrow 3$ e transição 3, é a saída o indivíduo do estado dois para o estado três, $2 \rightarrow 3$. Note que a definição dos estados não contempla a possibilidade de haver uma transição do paciente estar no estado intermediário 2 e retornar ao estado inicial da doença pois tal evento não foi observado nesta amostra.

Na Tabela 3, são mostradas as 6 primeiras linhas do conjunto de dados no formato multi-estado. Note que as variáveis ID e *Diag_time* são as mesmas do conjunto de dados longitudinal. A variável *Diag_time* foi a única, dentre todas as testadas, que apresentou significância estatística no modelo longitudinal e por esta razão as outras não aparecem no conjunto de dados no formato multi-estado.

Para explicar este conjunto, considere o indivíduo com ID igual a 2. A coluna $from$ indica o estado inicial e a coluna to o estado seguinte. Ele partiu do estado 1, com risco de transições 1, 2 e 3, indicado pela coluna $trans$. Isso significa que ele pode mudar para os

estados 2 e 3. No momento 4.15, o paciente se move para estado 2 (anormal), de onde ele corre risco para uma transição adicional para o estado 3 (isto é, transição 3). Como a coluna $status$ indica valor 0, ou seja, não ocorreu esta última transição, significa que o paciente é censurado no tempo 8.19 que é indicado pela coluna $Tstop$. A coluna $time$ é simplesmente a diferença entre $Tstop$ e $Tstart$, onde o refere-se ao tempo gasto no estado atual.

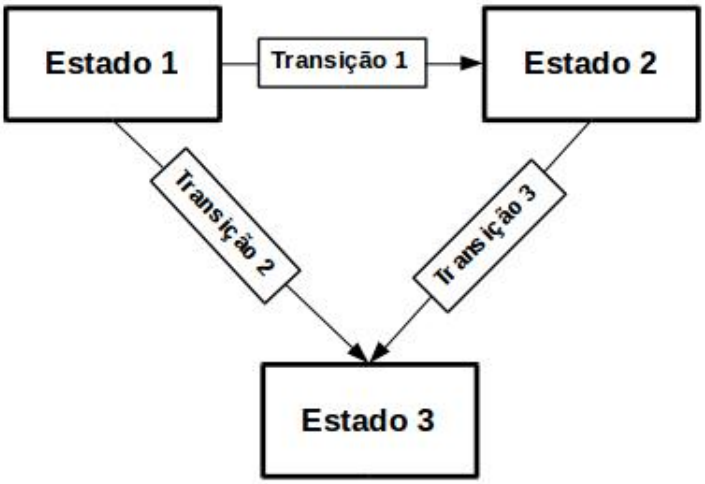


Figura 3 – Estados e transições do processo multi-estado

Tabela 3 – Dados no formato multi-estado

ID	from	to	trans	Tstart	Tstop	time	status	Diag_time
1	1	2	1	0.00	8.21	8.21	0	10.77
1	1	3	2	0.00	8.21	8.21	0	10.77
2	1	2	1	0.00	4.15	4.15	1	10.32
2	1	3	2	0.00	4.15	4.15	0	10.32
2	2	3	3	4.15	8.19	4.04	0	10.32
3	1	2	1	0.00	8.21	8.21	0	12.38

Na Figura 4, é apresentado um gráfico de perfil individual do $\log(PSA)$ ao longo do tempo. Cada linha cinza contida no gráfico representa um indivíduo, desta forma pode-se observar o comportamento de cada elemento da amostra. A linha azul, representa o comportamento médio estimado por meio de método de suavização $loess$, (CLEVELAND; DEVLIN, 1988).

Ainda é possível, por meio da Figura 4, notar que as observações são longitudinais e cada observação da amostra parece ter um comportamento individual diferente, o que pode justificar o uso de um submodelo longitudinal com efeito misto.

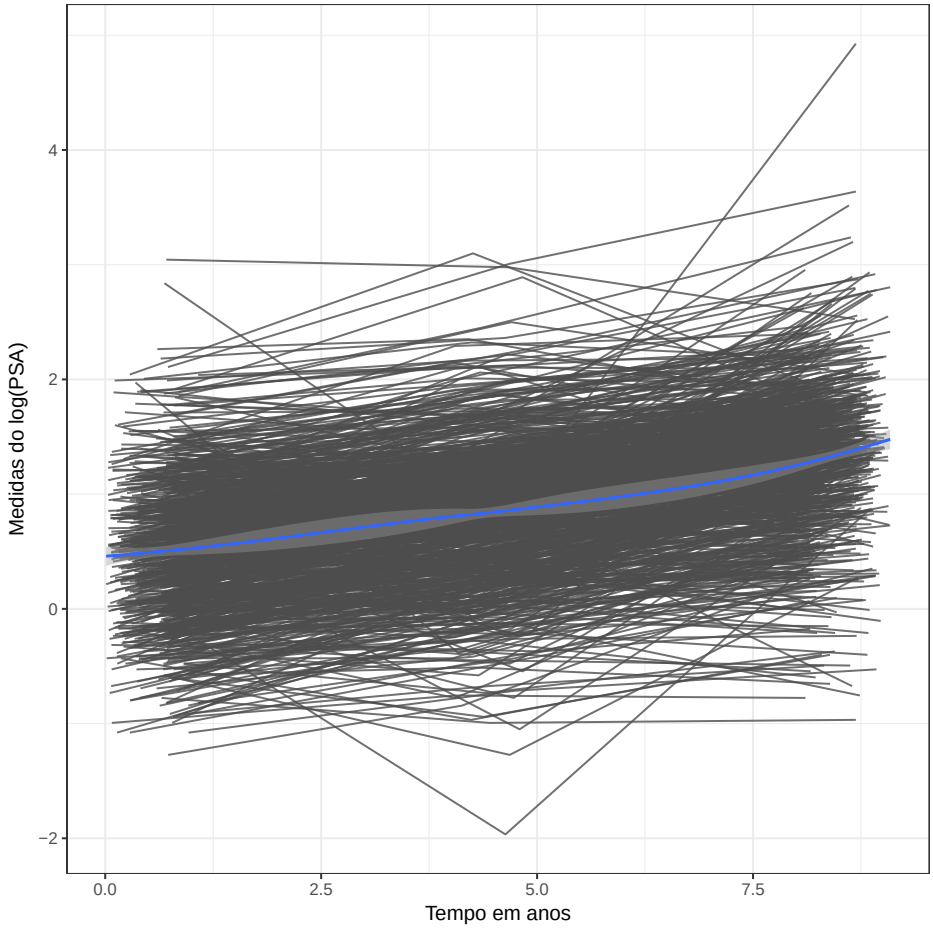


Figura 4 – Medidas do log(PSA) ao longo do tempo

3.2 Modelando os dados

3.2.1 Ajuste do submodelo longitudinal

Ressaltando que a variável resposta, $Y_{ij} = \log[PSA_{i(t_{ij})}]$, que indica o $\log$ do PSA para o indivíduo i no tempo t_{ij}, esta transformação logarítmica faz com que a resposta apresente uma densidade empírica estimada com formato muito próximo à distribuição gaussiana, como é visto na Figura 5.

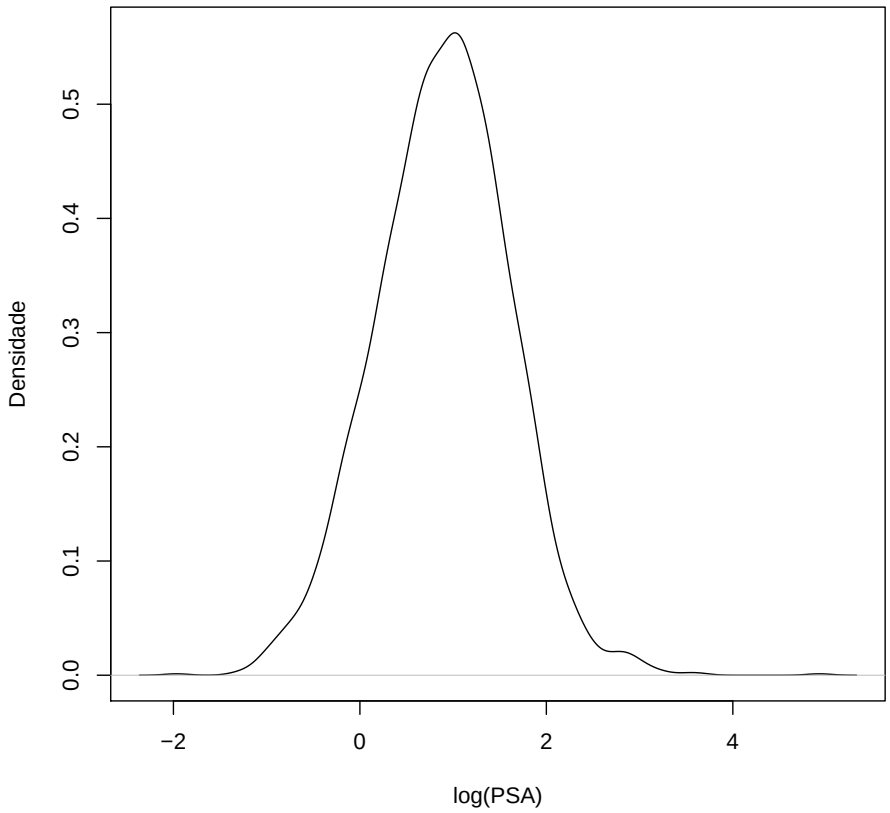


Figura 5 – Densidade empírica estimada do log(PSA)

Assim, o modelo longitudinal foi considerado ter reposta que segue uma distribuição gaussiana e com dois efeitos aleatórios, um no intercepto e outro na inclinação. As covariáveis para efeitos fixos são o tempo de ocorrência do biomarcador do PSA e a covariável $Diag_time$ que é o tempo em que o paciente permaneceu na seleção do estudo.

Dada a expressão em 2.1, o modelo pode escrito como segue,

$$Y_{ij} = \beta_0 + b_{i0} + (b_{i1} + \beta_1)times + \beta_2 Diag\_time + \epsilon_{ij}, \tag{3.1}$$

onde $b_i \sim N(0, G)$, com $b_i = (b_{i0}, b_{i1})$ e G é a matriz de covariância dos efeitos aleatórios. Este modelo considera dois efeitos aleatórios e assume que a medida PSA, muda de paciente para paciente ao longo do tempo e ainda pode-se escrever, $b_i \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}G_{11} & G_{12}\\G_{21} & G_{22}\end{pmatrix}\right)$. Em relação ao erros, estes terão distribuição gaussiana com média zero e matriz de covariância com variância constante e denotado por, $\epsilon_{ij} \sim N(0, \sigma^2 I)$.

Na Tabela 4 são apresentados os parâmetros estimados, erro padrão e p-valor dos efeitos fixos do modelo linear misto ajustado, ou seja, do modelo marginal. Nota-se que todos os coeficientes das covariáveis consideradas no modelo apresentaram efeitos significativos,

segundo p-valores apresentado sendo menores que 0.05. A medida que o tempo passa o nível do biomarcador tende a aumentar, devido ao efeito positivo apresentado pela estimativa da covariável $times$, 0.1076. A covariável $Diag\_time$ apresentou efeito negativo, -0.0747, ou seja, a medida que o tempo em observação aumenta o valor do biomarcador diminui, o que faz sentido pois quanto mais tempo o paciente é acompanhado, pressupõe que por mais tempo ele receberá cuidados médicos. Assim como as outras covariáveis testadas, a interação entre $times$ e $Diag\_time$, também não foi significativa.

Tabela 4 – Modelo marginal

	Coeficiente	Erro padrão	P-valor
Intercepto	1.2885	0.0665	$< 0.001^{***}$
times	0.1076	0.0026	$< 0.001^{***}$
Diag_time	-0.0747	0.0053	$< 0.001^{***}$

A matriz de efeitos aleatórios, G, obtida pelo ajuste do modelo é,

$$G = \begin{pmatrix} 0.3280 & -0.0176 \\ -0.0176 & 0.0028 \end{pmatrix},$$

em que 0.3280 é a variância do intercepto, 0.0028 é a variância do tempo e o valor -0.0176 é a covariância dos dois termos e nos mostra que a correlação dos valores iniciais da variável resposta e com a inclinação do tempo é negativa, ou seja, o valor do PSA tende a diminuir ao longo do tempo após a intervenção do tratamento. Ainda, é informado pelo modelo, que a variância residual estimada é 0.0855, que define a matriz $\sigma^2 I$ para o ϵ_{ij}. Como a variância residual é constante e conforme em 2.1, isto quer dizer que, não existe correlação entre os indivíduos, apenas dentro dos indivíduos.

3.2.2 Ajuste do submodelo multi-estado

Antes de apresentar o modelo multi-estado com riscos proporcionais, será inspecionada as transições entre os estados que os dados simulados possuem. Para isso, na Tabela 5, há o resumo das transições. Verifica-se que, entre os 808 indivíduos que estavam no estado 1, 709 transitaram para o estado 2, 9 transitaram para o estado 3 e 90 permaneceram no estado 1. Entre os que estavam no estado 2, 112 atingiram o estado absorvente 3 e 597 permaneceram em seu estado atual. A última linha da Tabela 5, é composta de zeros pois uma vez que o paciente atinge o estado 3 que é o absorvente, não pode mais atingir outro estado. Nota-se também que das 808 obervações, 121 $(9 + 112)$ chegaram ao estado absorvente, que neste caso é a morte.

Para o ajuste do modelo multi-estado com riscos proporcionais, a covariável $Diag\_time$ foi considerada para se estimar seu impacto em cada uma das três possíveis transições. Portanto, na saída do modelo teremos os coeficientes de risco estimados do tempo que cada paciente

Tabela 5 – Resumo das transições entre os estados

	Estado 1	Estado 2	Estado 3	Sem Transição	Total
Estado 1	0	709	9	90	808
Estado 2	0	0	112	597	709
Estado 3	0	0	0	0	0

permaneceu no estudo para cada transição. Desta forma, o modelo é escrito, de acordo com a expressão (2.21),

$$\lambda(t) = \lambda_{rs,0}(t)\exp(Diag\_time_{12}\beta_{12} + Diag\_time_{13}\beta_{13} + Diag\_time_{23}\beta_{23}), \tag{3.2}$$

porém ainda sem o compartilhamento do efeito aleatório entre os submodelos. O $\lambda_{rs,0}(t)$ é definido como o risco basal, os índices das covariáveis e dos coeficientes estimados, indicam as transições, ou seja, $Diag\_time_{12}\beta_{12}$ dará o impacto da covariável $Diag\_time$ na transição entre os estados 1 e 2, que de acordo com a Figura 3, é a primeira transição do processo. E essa interpretação também vale para as outras duas transições.

Na Tabela 6, são apresentados os resultados do modelo multi-estado com riscos proporcionais. É importante salientar que os dados no formato multi-estado foram preparados usando a função do R *mstate()*, do pacote de mesmo nome, e os parâmetros do modelo foram ajustados com a função *coxph()* do pacote *survival*.

Tabela 6 – Modelo multi-estado com riscos proporcionais

	Coeficiente	Exp(Coeficiente)	Erro padrão	P-Valor
β_{12}	-0.0696	0.9327	0.0123	$< 0.001^{***}$
β_{13}	-0.2038	0.8156	0.1867	0.2750
β_{23}	-0.1252	0.8823	0.0367	$< 0.001^{***}$

A coluna coeficientes da Tabela 6 apresenta os valores de β estimados. Valores positivos para β indicam que a covariável $Diag\_time$ aumenta o risco de óbito dado a transição considerada e valores negativos para β indica o contrário, ou seja, diminui o risco. A coluna $Exp(coeficientes)$, é o risco relativo, ou seja, valores acima de 1 indicam sobrerrisco e valores entre 0 e 1 indicam proteção. Os intervalos com (95%) de confiança, para estas estimativas, são apresentados na Tabela 7. A Tabela 6, também divulga os erros padrão das estimativas e sua significância por meio do p-valor.

Nota-se na Tabela 6, que a covariável $Diag\_time$ impacta significativamente as transições de número 1 e 3, ou seja, o o tempo em que o indivíduo permaneceu em observação influencia significativamente o risco de óbito nas transições do estado 1 para o 2 e do estado 2 para o 3. E este risco diminui na primeira transição em $1.072(\exp(-0.9327))$ vezes a cada unidade de tempo, podendo esta estimativa,ao nível de 95% de confiança, variar entre 1.046 e

1.098. Para a transição entre os estados 2 e 3, o risco diminui $1.133(\exp(-0.8823))$ vezes, a cada unidade de tempo, podendo esta estimativa,ao nível de 95% de confiança, variar entre 1.055 e 1.218.

A não influência de $Diag\_time$ na transições 2, fica evidenciada não apenas pelos p-valores mostrados na Tabela 6, mas também pelos intervalos de confiança da Tabela 7. Para esta estimativa, pode ser observado que os seu respectivo intervalo de confiança possui o valor 1 contido indicando que as estimativa de risco não é significativa. Uma explicação para a não influência significativa, se deve ao baixo número de indivíduos que estavam no estado normal da doença e transitaram ao estado absorvente, como é visto na Tabela 5, apenas 9 casos observados. É interessante, ainda, salientar que destes, apenas 2 deles atingiram a morte devido ao câncer de próstata. Os restantes, morreram de outras causas.

Tabela 7 – Intervalo de confiança(95%) para os riscos proporcionais

	Exp(Coeficientes)	Limite Inferior	Limite Superior
β_{12}	0.9327	0.9106	0.9555
β_{13}	0.8156	0.5657	1.1760
β_{23}	0.8823	0.8210	0.9482

3.2.3 Ajuste do joint model

Finalmente, nesta seção, o modelo *joint model* ajustado é apresentado. Com base na expressão (2.21), e a covariável $Diad\_time$ que foi considerada nos submodelos linear misto e multi-estado com riscos proporcionais, nas equações dos modelos (3.1) e (3.2) respectivamente.

Portanto, o modelo será escrito da seguinte forma,

$$\lambda^i_{rs}(t \mid b_i) = \lambda_{rs,0}(t)\exp(\beta_{rs}Diag\_time_i + \eta_{rs,intercepto}Y^*_i(t) + \eta_{rs,inclinação}Y^*_i(t)), \quad (3.3)$$

onde o processo multi-estado inclui os três estados, $(r, s \in \{1, 2, 3\})$, e as três transições descritas na Figura 3.

A matriz dos efeitos aleatórios G, obtida pelo ajuste do joint model é,

$$G = \begin{pmatrix} 0.3118 & -0.0232 \\ -0.0232 & 0.005 \end{pmatrix},$$

em que 0.3118 é a variância do intercepto, 0.005 é a variância do tempo, ou seja da inclinação, e o valor -0.0232 é a covariância dos dois termos. E as interpretações feitas para matriz G dos efeitos aleatórios para o submodelo linear misto, valem para a matriz G do *joint model*. Ainda, é informado pelo modelo, que a variância residual estimada é 0.3329, que define a matriz de covariância dos erros, $\sigma^2 I$ para o ϵ_{ij}.

Tabela 8 – Joint model

	Processo multi-estado				Processo longitudinal		
	Coef.	Erro padrão	p-valor		Coef.	Erro padrão	p-valor
β_{12}	0.03	0.02	0.0557	Intercepto	1.33	0.07	< 0.0001
β_{13}	0.91		NaN	Diag_time	-0.08	0.01	< 0.0001
β_{23}	-0.24	0.23	0.2992	times	0.11	0.00	< 0.0001
$\eta_{12,intercepto}$	1.20	0.11	< 0.0001	σ	0.3329		
$\eta_{13,intercepto}$	9.26		NaN				
$\eta_{23,intercepto}$	14.52	2.99	< 0.0001	D_{11}	0.3118		
$\eta_{12,inclinação}$	1.53	1.38	0.2674	D_{12}	-0.0232		
$\eta_{13,inclinação}$	-21.81	14.80	0.1406	D_{22}	0.005		
$\eta_{23,inclinação}$	11.05	12.39	0.3724				

Pode-se observar na Tabela 8 que as estimativas para os efeitos fixos do processo longitudinal no *joint model* permaneceram significativas e com estimativas e erros padrão muito semelhantes aos do modelo linear misto (ver Tabela 4) confirmando que o nível de *PSA* e sua trajetória ao longo do tempo está associado com o tempo de permanência do paciente em obervação.

Para o processo multi-estado, a covariável $Diag\_time$, que foi identificada como um fator de prognóstico, mostrou que o tempo de permanência do indivíduo no estudo, não está associado com a intensidade de transição entre os estados de saúde após o ajuste do modelo levando em conta a dinâmica longitudinal para explicar a evolução do biomarcador. Nota-se tal fato pela não-significância dos parâmetros $\beta's$.

Com relação aos parâmetros $\eta's$, que quantificam a associação entre a dinâmica longitudinal de evolução do *PSA* (para valores atuais de intercepto e inclinação) e as progressões clínicas (indicadas pelas transições entre os estados), houve efeitos significativos sobre o nível atual(intercepto) do biomarcador entre as transições 1(estado 1 para o 2) e 3(estado 2 para o 3). Por exemplo, depois de ajustado o modelo com covariável e para o verdadeiro valor da inclinação do biomarcador, um incremento ao verdadeiro valor do biomarcador$(\log(PSA)$ sem medida do erro), induz ao um risco, em $3.32(\exp(1.20))$ vezes maior do paciente que está no estado 1 (normal) transitar para o estado 2(anormal). Analogamente, a cada unidade aumentada ao verdadeiro valor do biomarcador, aumenta o risco de morte para os pacientes que estão no estado 2, aumenta drasticamente $(\exp(14.52))$. Para todas as transições não houve efeito significativo da atual inclinação do biomarcador em relação as intensidades das transições.

Na Figura 6, é mostrado gráficos de diagnóstico de resíduos para o modelo ajustado. Inclui o gráficos de valores ajustados versus os resíduos observados.

Observa-se que a curva ajustada pelo método *Loess* não mostram uma tendência sistemática, sugerindo que os pressupostos do modelo não estão sendo violados e justifica a não necessidade de imputação de dados para a análise de resíduos verificação de indepência dos resíduos (RIZOPOULOS, 2012).

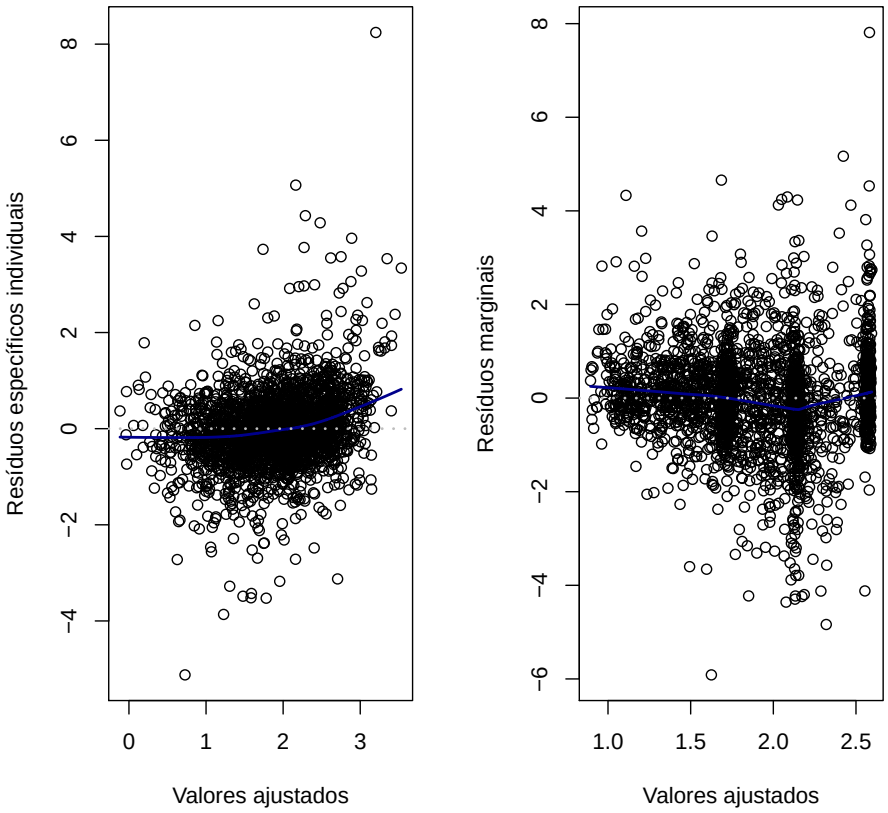


Figura 6 – Análise de resíduos do joint model

Capítulo 4

Considerações finais

O *joint model*, para o biomarcador PSA observado ao longo do tempo, disponibilizou um modelo que pode auxiliar na progressão da doença do câncer de próstata e outras doenças com biomarcadores com repetição ao longo do tempo, levando em conta, para o prognóstico, fatores para gerenciar o risco de óbito ao longo tempo e a influência desses fatores na transição entre os estados da doença. Com o processo longitudinal pode-se identificar quais são os fatores de prognóstico e a junção do processo multi-estado confirma se tais fatores influenciam as intensidades entre todas as transições.

Apesar da implementação de tal modelo ser acessível, por já existir os pacotes necessários no *CRAN* do *R*, *mstate* e *JM*, a convergência do modelo não é tão simples. Para o ajuste do processo multi-estado, os dados precisam estar em uma estrutura específica. Esta fase não é difícil pois o pacote *mstate* dá todas as funções necessárias para isso. Este é apenas um procedimento detalhado é imprescindível pois a qualidade do ajuste está diretamente relacionada as especificações dos estados e seus tempos, portanto esta preparação requer muito cuidado e excelência. Com relação a convergência do modelo, notou-se, para este conjunto de dados, que o segundo estágio, para estimação dos parâmetros do modelo, com o algoritmo Quase-Newton, realmente é necessário, pois ao tentar ajuste apenas com o estágio inicial com o algoritmo *EM* para as mais diversas quantidades de iterações (valores baixos, altos e muito altos), ou poucas iterações para o método Quase-Newton, o modelo não convergiu. Outra especificação importante, e já verificada em outro estudo (FERRER et al., 2016), é a não convergência quando se utiliza poucos pontos para a resolução das integrais sobre os efeitos aleatórios usando quadraturas Gauss-Hermite. É recomendado que se utilize, no mínimo, nove pontos. A dificuldade do modelo estimar os erros padrão, referentes aos parâmetros relacionados a transição 2, pode ser justificada pela pouca quantidade de observações nesta transição, apenas 9 como pode-ser na Tabela 5. Essa quantidade pequena de dados pode ter causado um problema na matriz Hessiana.

O modelo confirmou que o acompanhamento médico mais duradouro influencia a evolu-

ção do biomarcador e quais fatores podem ser ou não de risco para as intensidades de transição entre os estados da doença, quando considerada a dinâmica longitudinal do biomarcador.

Além disso, a atual inclinação do biomarcador não teve impacto significativo para nenhuma das transições consideradas, ou seja, o valor da inclinação do biomarcador de cada paciente não influenciou o risco para nenhuma das transição entre os estados.

Outros estudos sobre câncer de próstata encontraram uma forte associação entre a inclinação do $\log(PSA)$ e transições entre qualquer estados de recorrência clínica e terapia hormonal (TAYLOR et al., 2013; SÈNE et al., 2014). Porém, neste estudo não havia a informação de recorrências clínicas e terapia hormonal no conjunto de dados e estes possíveis estados não foram incorporados, o que pode explicar a não significância da inclinação sobre as intensidades de todas as transições.

Como em estudos considerando um processo de sobrevivência (RIZOPOULOS, 2012), predições individualizadas sobre o risco da progressão da doença podem ser feitas para quantificação deste risco e considerando o processo multi-estado, ainda, há a possibilidade de verificar a progressão deste risco para todas as transições entre os estados considerados no estudo. Desta forma obtêm-se um modelo mais completo para análise da progressão da doença levando em conta todo o dinamismo longitudinal e das transições entre os estado que a doença pode apresentar.

Devido ao pedido de sigilo, em relação aos dados, feito pelo professor da *University of Tampare*, eles não poderão ser disponibilizados para reprodução das análises, porém no Apêndice, é apresentado código reproduzível com dados simulados. Os mesmos dados utilizados na qualificação desta dissertação.

Referências

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974.

ANDERSEN, P. K.; KEIDING, N. Multi-state models for event history analysis. **Statistical methods in medical research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 11, n. 2, p. 91–115, 2002.

BEYERSMANN, J.; ALLIGNOL, A.; SCHUMACHER, M. **Competing risks and multistate models with R**. [S.l.]: Springer Science & Business Media, 2011.

BRESLOW, N. E.; DAY, N. E.; DAVIS, W. **Statistical methods in cancer research**. [S.l.]: International Agency for Research on Cancer Lyon, 1987. v. 2.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de Sobrevivência: teoria e aplicações em saúde**. [S.l.]: SciELO-Editora FIOCRUZ, 2011.

CLEVELAND, W. S.; DEVLIN, S. J. Locally weighted regression: an approach to regression analysis by local fitting. **Journal of the American statistical association**, Taylor & Francis Group, v. 83, n. 403, p. 596–610, 1988.

DANTAN, E.; JOLY, P.; DARTIGUES, J.-F.; JACQMIN-GADDA, H. Joint model with latent state for longitudinal and multistate data. **Biostatistics**, Oxford University Press, v. 12, n. 4, p. 723–736, 2011.

ELASHOFF, R. M.; LI, G.; LI, N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. **Biometrics**, Wiley Online Library, v. 64, n. 3, p. 762–771, 2008.

EMILIANO, P. C.; VEIGA, E. P.; VIVANCO, M. J.; MENEZES, F. S. Critérios de informação de akaike versus bayesiano: análise comparativa. **19º Simpósio Nacional de Probabilidade e Estatística**, 2010.

FARAWAY, J. J. **Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models**. [S.l.]: CRC press, 2016. v. 124.

FERRER, L.; RONDEAU, V.; DIGNAM, J.; PICKLES, T.; JACQMIN-GADDA, H.; PROUST-LIMA, C. Joint modelling of longitudinal and multi-state processes: application to

clinical progressions in prostate cancer. **Statistics in medicine**, Wiley Online Library, v. 35, n. 22, p. 3933–3948, 2016.

FILHO, J. A. C. **Modelos lineares mistos: estruturas de matrizes de variâncias e covariâncias e seleção de modelos.** Tese (Doutorado) — Universidade de São Paulo, 2002.

FINNE, P.; STENMAN, U.-H.; MÄÄTTÄNEN, L.; MÄKINEN, T.; TAMMELA, T.; MARTIKAINEN, P.; RUUTU, M.; ALA-OPAS, M.; ARO, J.; KARHUNEN, P. et al. The finnish trial of prostate cancer screening: where are we now? **BJU international**, Wiley Online Library, v. 92, n. s2, p. 22–26, 2003.

FRYDMAN, H.; SZAREK, M. Nonparametric estimation in a markov "illness–death" process from interval censored observations with missing intermediate transition status. **Biometrics**, Wiley Online Library, v. 65, n. 1, p. 143–151, 2009.

FURTADO, S. M. T. Uso de modelo misto para a análise de dados longitudinais de um experimento com bovinos em lactação. UNIVERSIDADE FEDERAL DE LAVRAS, 2009.

GIOLO, S. R.; COLOSIMO, E. A. **Análise de sobrevivência aplicada**. [S.l.]: Edgard Blucher, 2006.

HENDERSON, R.; DIGGLE, P.; DOBSON, A. Joint modelling of longitudinal measurements and event time data. **Biostatistics**, Oxford University Press, v. 1, n. 4, p. 465–480, 2000.

HUSZTI, E.; ABRAHAMOWICZ, M.; ALIOUM, A.; BINQUET, C.; QUANTIN, C. Relative survival multistate markov model. **Statistics in medicine**, Wiley Online Library, v. 31, n. 3, p. 269–286, 2012.

JACKSON, C. H. et al. Multi-state models for panel data: the msm package for r. **Journal of Statistical Software**, v. 38, n. 8, p. 1–29, 2011.

JACKSON, C. H.; SHARPLES, L. D.; THOMPSON, S. G.; DUFFY, S. W.; COUTO, E. Multistate markov models for disease progression with classification error. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 52, n. 2, p. 193–209, 2003.

JUAREZ-COLUNGA, E.; SILVA, G.; DEAN, C. Joint modeling of zero-inflated panel count and severity outcomes. **Biometrics**, Wiley Online Library, 2017.

KENWARD, M. G.; MOLENBERGHS, G. Likelihood based frequentist inference when data are missing at random. **Statistical Science**, JSTOR, p. 236–247, 1998.

LANGE, J. M.; HUBBARD, R. A.; INOUE, L. Y.; MININ, V. N. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. **Biometrics**, Wiley Online Library, v. 71, n. 1, p. 90–101, 2015.

LEE, E. T.; WANG, J. **Statistical methods for survival data analysis**. [S.l.]: John Wiley & Sons, 2003. v. 476.

MARTINEZ, J. M.; SANTOS, S. A. Métodos computacionais de otimização. **Colóquio Brasileiro de Matemática, Apostilas**, v. 20, 1995.

MARTINS, R.; SILVA, G. L.; ANDREOZZI, V. Bayesian joint modeling of longitudinal and spatial survival aids data. **Statistics in medicine**, Wiley Online Library, v. 35, n. 19, p. 3368–3384, 2016.

MARTINS, R. M. d. C. Métodos bayesianos aplicados à modelagem conjunta de dados longitudinais e de sobrevivência. 2013.

MCCULLOCH, C. E.; SEARLE, S. R. Generalized linear mixed models (glmms). **Generalized, Linear, and Mixed Models**, Wiley Online Library, 2001.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2015.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D. R core team (2014) nlme: linear and nonlinear mixed effects models. r package version 3.1-117. **Available at h ttp://CRAN. R-project. org/package= nlme**, 2014.

PINHEIRO, J. C.; BATES, D. M. Mixed-effects models in s and s-plus. **Statistics and computing**, 1978.

PROUST-LIMA, C.; TAYLOR, J. M. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. **Biostatistics**, Oxford University Press, v. 10, n. 3, p. 535–549, 2009.

PUTTER, H. Tutorial in biostatistics: Competing risks and multi-state models analyses using the mstate package. 2016.

PUTTER, H.; FIOCCO, M.; GESKUS, R. B. Tutorial in biostatistics: competing risks and multi-state models. **Statistics in medicine**, Wiley Online Library, v. 26, n. 11, p. 2389–2430, 2007.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

RIZOPOULOS, D. Jm: An r package for the joint modelling of longitudinal and time-to-event data. **Journal of Statistical Software (Online)**, v. 35, n. 9, p. 1–33, 2010.

RIZOPOULOS, D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. **Biometrics**, Wiley Online Library, v. 67, n. 3, p. 819–829, 2011.

RIZOPOULOS, D. **Joint models for longitudinal and time-to-event data: With applications in R**. [S.l.]: CRC Press, 2012.

SAINT-PIERRE, P.; COMBESCURE, C.; DAURES, J.; GODARD, P. The analysis of asthma control under a markov assumption with use of covariates. **Statistics in Medicine**, Wiley Online Library, v. 22, n. 24, p. 3755–3770, 2003.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance components**. [S.l.]: John Wiley & Sons, 2009. v. 391.

SELF, S.; PAWITAN, Y. Modeling a marker of disease progression and onset of disease. In: **AIDS epidemiology**. [S.l.]: Springer, 1992. p. 231–255.

SEN, P. K.; SINGER, J. M.; LIMA, A. C. P. D. **From finite sample to asymptotic methods in statistics**. [S.l.]: Cambridge University Press, 2010. v. 28.

SÈNE, M.; BELLERA, C. A.; PROUST-LIMA, C. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. **Journal de la Société Française de Statistique**, v. 155, n. 1, p. 134–155, 2014.

SHEATHER, S. **A modern Approach to Regression with R**. [S.l.]: Springer Science & Business Media, 2009.

SLAWIN, K. M.; OHORI, M.; DILLIOGLUGIL, O.; SCARDINO, P. T. Screening for prostate cancer: an analysis of the early experience. **CA: a cancer journal for clinicians**, Wiley Online Library, v. 45, n. 3, p. 134–147, 1995.

SOBREIRO, B. P. Câncer da próstata estadio t1c tratado por prostatectomia radical. 2013.

TAYLOR, J. M.; PARK, Y.; ANKERST, D. P.; PROUST-LIMA, C.; WILLIAMS, S.; KESTIN, L.; BAE, K.; PICKLES, T.; SANDLER, H. Real-time individual predictions of prostate cancer recurrence using joint models. **Biometrics**, Wiley Online Library, v. 69, n. 1, p. 206–213, 2013.

TSIATIS, A. A.; DAVIDIAN, M. Joint modeling of longitudinal and time-to-event data: an overview. **Statistica Sinica**, JSTOR, p. 809–834, 2004.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. [S.l.]: Springer, 2009.

WEST, B. T.; WELCH, K. B.; GALECKI, A. T. **Linear mixed models: a practical guide using statistical software**. [S.l.]: CRC Press, 2014.

WREEDE, L. C. de; FIOCCO, M.; PUTTER, H. et al. mstate: an r package for the analysis of competing risks and multi-state models. **Journal of Statistical Software**, v. 38, n. 7, p. 1–30, 2011.

WU, L.; LIU, W.; YI, G. Y.; HUANG, Y. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. **Journal of Probability and Statistics**, Hindawi Publishing Corporation, v. 2012, 2011.

XAVIER, L. Modelos univariado e multivariado para analise de medidas repetidas e verificação da acurácia do modelo univariado por meio de simulação. Piracicaba, SP (Brazil), 2000.

YAN, X.; SU, X. **Linear regression analysis: theory and computing**. [S.l.]: World Scientific, 2009.

YU, M.; TAYLOR, J. M. G.; SANDLER, H. M. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. **Journal of the American Statistical Association**, Taylor & Francis, v. 103, n. 481, p. 178–187, 2008.

Apêndice

Abaixo, são disponibilizados os códigos do *R*, que foram utilizados nas análises com dados simulados.

```
#####################################################################
###  Baixando e instalando o pacote mstate com versão compatível
#####################################################################

packageurl <- "https://cran.r-project.org/src/contrib/Archive/mstate/mstate_0.2.7.tar.gz"
install.packages(packageurl, repos=NULL, type="source")


#######################################################
### Carregando todos os pacotes necessários
#######################################################
sapply(c('mstate','JM','nlme'), char=T, library)

#######################################################
### Versão do pacote: mstate
#######################################################
R.version$version.string
packageDescription("mstate", fields = "Version")

#######################################################
### Carregando função JMstateModel.R e dados
### <https://github.com/LoicFerrer/JMstateModel/>
#######################################################
source("JMstateModel.R")
load("data.RData")

#####################################################################
### Gráfico de perfil individual da resposta ao longo do tempo
#####################################################################

library(ggplot2)
```

```
plot_long <- (ggplot(data_long) +
               geom_line(aes(x = times, y = Y, group = id),
                         color = "grey30", alpha = 0.8) +
               stat_smooth(aes(x = times, y = Y),
                           method = "loess", size = 0.75) + theme_bw() +
               xlab("Time") + ylab("Marker value"))
plot_long

############################################################
### Gráfico de densidade empírica da variável resposta
############################################################
ggplot(data_long, aes(x=Y), color = "grey30", alpha = 0.8)
  + geom_density() + theme_bw()
  + xlab("log(PSA)") + ylab("Densidade")

#########################
### Modelo linear misto
#########################
lmeFit <- lme(Y ~ times* X, data = data_long,
              random = ~ times | id, method = "REML",
              control = list(opt = "optim"))

summary(lmeFit)

#########################
#### Modelo multi-estado
#########################
# Construção da matriz 3*3 da possíveis transições
tmat <- matrix(NA, 3, 3)
tmat[1, 2:3] <- 1:2
tmat[2, 3] <- 3

dimnames(tmat) <- list(from = c("State 0", "State 1", "State 2"),
                       to = c("State 0", "State 1", "State 2"))

# Define a covariável no modelo multi-estado:
covs <- "X"

# A função mstate::msprep() prepara os dados de sobrevivência no formato multi-estado
data_mstate <- msprep(time = c(NA,"t_State_1","t_State_2"),
                      status = c(NA,"State_1", "State_2"),
                      data = data_surv, trans = tmat,
                      keep = covs,
                      id = "id")

# Resume as transiçoes entre os estados
events(data_mstate)
```

```
# mstate::expand.covs(), permite definir quais  covariáveis impactam cada transição
data_mstate <- expand.covs(data_mstate, covs, append = TRUE,
                           longnames = FALSE)

#############################################################
#### Ajusta o modelo multi-estado com riscos proporcionais
#############################################################
coxFit <- coxph(Surv(Tstart, Tstop, status) ~
                    X.1 + X.2 + X.3 + strata(trans),
                data = data_mstate, method = "breslow",
                x = TRUE, model = TRUE)
summary(coxFit)

############################
#### Ajusta o Joint model
############################
# Define a derivada dos efeitos fixos e aleatórios no modelo misto,
# e indica quais covariáveis são mantidas, para a dependência na inclinação
# do biomarcador:
dForm <- list(fixed = ~1 + X, indFixed = c(2, 4),
              random = ~1,indRandom = 2)


# Joint model:
jointFit <- JMstateModel(lmeObject = lmeFit, survObject = coxFit,
                         timeVar = "times", parameterization = "both",
                         method = "spline-PH-aGH",
                         interFact = list(value = ~strata(trans) - 1,
                                          slope = ~strata(trans) - 1,
                                         data = data_mstate),
                         derivForm = dForm,
                         Mstate = TRUE,
                         data.Mstate = data_mstate, ID.Mstate = "id",
                         control = list(GHk = 3, lng.in.kn = 3))
summary(jointFit)
```