

MODELAGEM EM ANÁLISE DE SOBREVIVÊNCIA COM  
EVENTOS RECORRENTES APLICADA A DADOS DA ÁREA  
MÉDICA

Thiago Santos Mota

Dissertação apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de Mestre em Biometria.

BOTUCATU  
São Paulo - Brasil  
Fevereiro - 2013

MODELAGEM EM ANÁLISE DE SOBREVIVÊNCIA COM  
EVENTOS RECORRENTES APLICADA A DADOS DA ÁREA  
MÉDICA

Thiago Santos Mota

Orientadora: Profa. Dra. **Liciana Vaz de Arruda Silveira**

Dissertação apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de Mestre em Biometria.

BOTUCATU  
São Paulo - Brasil  
Fevereiro - 2013

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉCNICA AQUISIÇÃO E TRATAMENTO DA INFORMAÇÃO  
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP  
BIBLIOTECÁRIA RESPONSÁVEL: **ROSEMEIRE APARECIDA VICENTE**

Mota, Thiago Santos.

Modelagem em análise de sobrevivência com eventos recorrentes aplicada a dados da área médica / Thiago Santos Mota. – Botucatu : [s.n.], 2013

Dissertação (mestrado) – Universidade Estadual Paulista, Instituto de Biociências de Botucatu

Orientador: Liciania Vaz de Arruda Silveira

Capes: 10203001

1. Biometria. 2. Diálise. 3. Insuficiência renal crônica.

Palavras-chave: AIC; BIC; Dialítico; Martigale; Modelos condicionais; Modelos marginais; Shoenfeld.

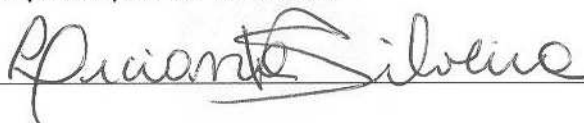
MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DE Thiago Santos Mota, INTITULADA Modelagem em análise de sobrevivência com eventos recorrentes aplicada a dados da área médica, APRESENTADA AO INSTITUTO DE BIOCIÊNCIAS, UNESP, CAMPUS DE BOTUCATU, SÃO PAULO, EM 04 de Fevereiro de 2013.

APROVADA PELA COMISSÃO JULGADORA:

Prof(a) Dr(a) Liciania Vaz de Arruda Silveira

Instituição: Unesp campus de Botucatu


Assinatura:



Prof(a) Dr(a) Lídia Raquel de Carvalho

Instituição: Unesp campus de Botucatu

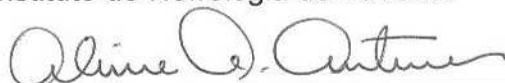
Assinatura:



Prof(a) Dr(a) Aline Araujo Antunes

Instituição: Instituto de Nefrologia de Taubaté

Assinatura:



## **Dedicatória**

Dedico este trabalho aos meus pais, Severina e Antonio (in memoriam), avós, Antonio Alexandre e Maria da Penha (in memoriam), e padrinhos, Dalvinha e Barreto, que proporcionaram uma infância alegre e com muitos ensinamentos.

## Agradecimentos

Primeiramente, agradeço aos meus pais, Severina e Antonio (in memoriam), pela minha educação e criação, sem eles não seria a pessoa que sou hoje.

Aos meus irmãos, Sandro e André, e minha prima Sileide e seu marido Vandson que de uma forma ou de outra me motivaram a estudar através das suas experiências ou pelo apoio para que eu pudesse conquistar meus objetivos.

Aos amigos, Farid, Paulo e Renan, da República, ano de 2012, com os quais, além de compartilhar o espaço de uma casa, compartilhei o dia a dia acadêmico, as alegrias, as angústias e com certeza se tornaram grandes amigos.

Aos amigos da sala 02, Cintia, Farid e Ronaldo, com os quais cursei disciplinas e compartilhei esses dois anos de mestrado o mesmo ambiente.

A todos os professores do curso de mestrado em Biometria e demais funcionários do departamento de Bioestatística que sempre estiveram dispostos a ajudar no que fosse possível.

À banca de qualificação e defesa de mestrado, composta por: profa. Dra. Luzia Aparecida Trinca, Dra. Aline Aline Araujo e a profa. Dra. Lídia Raquel de Carvalho pelas correções e sugestões que foram essenciais para melhoria deste trabalho.

À profa. Dra. Liciania Vaz de Arruda Silveira pela atenciosa orientação neste trabalho e em outros mais que tenho realizado com seu apoio e participação.

À Flávia Priscila Ventura, minha namorada, amiga e companheira de todas as horas.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CNPq) pelo financiamento para a realização dessa pesquisa.

# Sumário

	Página
<b>LISTA DE FIGURAS</b>	<b>vi</b>
<b>LISTA DE TABELAS</b>	<b>vii</b>
<b>RESUMO</b>	<b>viii</b>
<b>SUMMARY</b>	<b>x</b>
<b>1 INTRODUÇÃO</b>	<b>1</b>
<b>2 REVISÃO DE LITERATURA</b>	<b>3</b>
2.1 Análise de Sobrevivência . . . . .	3
2.2 Processos de Contagem . . . . .	7
2.3 Análise de Sobrevivência Multivariada . . . . .	8
2.3.1 Modelagem Marginal . . . . .	9
2.3.2 Modelagem Condicional . . . . .	21
<b>3 MATERIAL E MÉTODOS</b>	<b>28</b>
3.1 Dados . . . . .	28
3.2 Organização do banco de dados e as funções utilizadas no programa R .	30
<b>4 RESULTADOS E DISCUSSÃO</b>	<b>35</b>
<b>5 CONCLUSÕES</b>	<b>44</b>
<b>REFERÊNCIAS BIBLIOGRÁFICAS</b>	<b>45</b>

## APÊNDICES



## Lista de Figuras

Página

1	Risco basal acumulado para ocorrência de novo evento cardiovascular: à esquerda, o modelo de eventos ordenados independentes (AG); à direita, o modelo PWP especificado segundo a ordem de ocorrência dos eventos.	38
2	Resíduos padronizados de Schoenfeld versus os tempos para as covariáveis consideradas no modelo PWP. . . . .	40
3	Gráfico dos resíduos de Martingale para o modelo PWP versus indivíduos.	40

## Lista de Tabelas

	Página
1	Formato do banco de dados para análise de eventos múltiplos dos modelos AG, PWP e modelos condicionais. . . . . 31
2	Formato do banco de dados para análise de eventos múltiplos do modelo WLW. . . . . 32
3	Resultado da seleção das covariáveis para o modelo AG utilizando os métodos: <i>backward</i> , <i>forward</i> e <i>stepwise</i> . . . . . 36
4	Resultado da seleção das covariáveis para o modelo PWP utilizando os métodos: <i>backward</i> , <i>forward</i> e <i>stepwise</i> . . . . . 36
5	Resultado da seleção das covariáveis para o modelo WLW utilizando os métodos: <i>backward</i> , <i>forward</i> e <i>stepwise</i> . . . . . 36
6	Distribuição conjunta das variáveis estado e estrato. . . . . 37
7	Valores de AIC, AIC <sub>c</sub> e BIC para os modelos marginais. . . . . 37
8	Teste de proporcionalidade dos riscos no modelo PWP ajustado. . . . . 39
9	Resultado da seleção das covariáveis obtido para o modelo semi-paramétrico de fragilidade lognormal utilizando os métodos: <i>backward</i> , <i>forward</i> e <i>stepwise</i> . . . . . 41
10	Resultado da seleção das covariáveis obtido para o modelo semi-paramétrico de fragilidade gama utilizando os métodos: <i>backward</i> , <i>forward</i> e <i>stepwise</i> . . . . . 41
11	Valores de AIC e BIC para os modelos condicionais. . . . . 42
12	Estimativas das razões de risco associados às covariáveis do modelo PWP. 42

**MODELAGEM EM ANÁLISE DE SOBREVIVÊNCIA COM  
EVENTOS RECORRENTES APLICADA A DADOS DA ÁREA  
MÉDICA**

Autor: THIAGO SANTOS MOTA

Orientadora: Profa. Dra. LICIANA VAZ DE ARRUDA SILVEIRA

**RESUMO**

Esta dissertação teve como objetivo estudar os modelos marginais e condicionais utilizados em análise de sobrevivência multivariada e avaliar o efeito das covariáveis medidas no tempo, até a ocorrência de eventos cardiovasculares em pacientes prevalentes em tratamento dialítico no hospital das Clínicas da Faculdade de Medicina da UNESP – Campus de Botucatu, SP, com seguimento de 2008 a 2011.

Como os dados tratam de eventos recorrentes do mesmo tipo, foram utilizados na modelagem os modelos marginais – AG (Andersen & Gill, 1982), PWP (Prentice et al., 1981) e WLW (Wei et al., 1989) – e os modelos condicionais – semi-paramétricos com distribuições gama e lognormal para a variável de fragilidade. As covariáveis consideradas neste trabalho foram variáveis clínicas, nutricionais, laboratoriais e dialíticas.

Foram consideradas em todos os modelos as covariáveis significativas ao nível de 5%, utilizando os critérios de seleção de covariáveis: *backward*, *forward* e *stepwise*. Apresentou-se também os valores do critério de AKAIKE (AIC), critério de AKAIKE corrigido ( $AIC_c$ ) e o critério de informação de Bayes (BIC) para esses modelos. Foram discutidas outras questões importantes para selecionar o melhor modelo para o conjunto de dados, como a frequência de eventos e se a ocorrência de novos eventos é influenciada por eventos anteriores. A partir desses critérios, o modelo PWP foi o modelo selecionado para esses dados, e em seguida foi feita a análise de resíduos pelos resíduos de Shoenfeld e Martigale, o que mostrou que esse modelo se ajustou bem aos dados. Sendo esta análise feita no programa estatístico R Development Core Team (2012).

Ao interpretar as covariáveis do modelo, observou-se a associação entre o quociente de massa extracelular corporal por massa celular corporal (MEC/MCC) com a ocorrência de eventos cardiovasculares em pacientes em diálise, sendo que tal exploração até o momento ainda não tinha sido feita considerando a situação de múltiplos eventos por indivíduo.

# MODELING IN SURVIVAL ANALYSIS WITH EVENTS RECURRING APPLIED THE MEDICAL AREA DATA

Author: THIAGO SANTOS MOTA

Adviser: Profa. Dra. LICIANA VAZ DE ARRUDA SILVEIRA

## SUMMARY

The aim is to study the marginal and conditional models used in multivariate survival analysis and to evaluate the effect of the covariates measured over time for occurrence of recurring cardiovascular events in patients under dialysis treatment from clinics hospital of the Medicine Faculty, UNESP - Botucatu, SP, with follow-up time from 2008 to 2011. As the data treat the recurring events of the same type, the marginals models – AG (Andersen & Gill, 1982), PWP (Prentice et al., 1981) and WLW (Wei et al., 1989) – and the conditionals models – frailty models with lognormal and gamma distributions – were used in the modeling. The covariates considered in this study were clinical, nutritional, laboratory and dialytic. The covariates were considered significant at 5% in all models using the covariates selection criteria: backward, forward and stepwise. It also presented the values of the criteria of AKAIKE (AIC), corrected criteria of AKAIKE and criteria of the Bayes (BIC) for these models.

Other issues were discussed to select the best model for the data set, as frequency and the influence of previous events to the occurrence of new events. Based on these criteria, the PWP model was selected to model the data. Residual analysis was done by using the Shoenfeld and Martigale residuals, showed that this model fits the data well. All the results were obtained in R software (R Development Core Team, 2012).

The results showed association between the quotient extracellular mass by body cell mass (MEC/MCC) covariate and the occurrence of cardiovascular events in dialysis patients such exploration such finding is new for multiple events per individual.

# 1 INTRODUÇÃO

A análise de sobrevivência é uma das áreas da estatística que mais tem crescido nos últimos anos, segundo Therneau & Grambsch (2000). Esta área tem como objetivo utilizar técnicas e modelos estatísticos com intuito de analisar dados cuja a variável resposta é o tempo até a ocorrência do evento de interesse. Por questões financeiras e limitação do tempo das pesquisas, muitas vezes esses dados apresentam observações incompletas em que por algum motivo não foi possível observar o evento de interesse, o que é denominado como censura.

Conforme Therneau & Grambsch (2000), com a melhoria da qualidade de vida, reinternações e outros desfechos secundários, uma subárea da análise de sobrevivência que vem crescendo é a análise de dados com múltiplos eventos por indivíduo, que também é denominada na literatura como análise de sobrevivência multivariada. Aplicações desta análise vêm ocorrendo em diversas áreas.

Em especial na área médica, é possível encontrar vários exemplos de aplicação, como por exemplo, tempo até a ocorrência de: acidentes vasculares cerebrais em pacientes em diálise, câncer, ataques epiléticos, diarreia em crianças, doenças respiratórias agudas, entre outros. Nesses tipos de dados observa-se que os eventos podem ser ordenados (múltiplos eventos do mesmo tipo) e não ordenados (eventos de diferentes tipos). Nesta dissertação, ênfase foi dada a múltiplos eventos do mesmo tipo.

Nesta situação foram estudados os modelos marginais – AG (Andersen & Gill, 1982), PWP (Prentice et al., 1981) e WLW (Wei et al., 1989) – e os modelos condicionais – modelos semi-paramétricos com distribuição de fragilidade gama e lognormal – que levam em conta as características dos múltiplos eventos por indivíduo

(ou seja, os tempos de falhas correlacionados) na estimação do vetor de parâmetros associado às covariáveis do conjunto de dados estudado.

Os dados utilizados para analisar esses modelos foram extraídos da tese de doutorado de Antunes (2012), que tratou-se de um estudo coorte prospectivo que incluiu 145 pacientes com doença renal crônica em tratamento dialítico há no mínimo três meses e cuja variável resposta analisada foi o tempo até a ocorrência de eventos cardiovasculares nestes indivíduos. Neste trabalho considerou-se apenas 130 indivíduos devido a dados faltantes no banco de dados.

Ao estudar esses modelos os objetivos foram identificar as características principais de cada modelo, verificar a possibilidade da utilização de técnicas usuais de seleção de covariáveis, critérios de seleção dos modelos para um determinado conjunto de dados e diagnóstico da qualidade do ajuste dos modelos.

Nesta dissertação o principal objetivo da análise desses dados foi diagnosticar a associação do quociente de massa extracelular por massa celular corporal (MEC/MCC) com a ocorrência de eventos cardiovasculares em pacientes em diálise. Tal exploração foi feita por Antunes (2012), utilizando apenas o modelo de Cox usual com covariáveis dependentes do tempo.

Aqui, abordou-se esta situação em um cenário mais realista do que apresentado no trabalho de Antunes (2012), considerando as características dos dados de múltiplos eventos por indivíduo.



## 2 REVISÃO DE LITERATURA

### 2.1 Análise de Sobrevivência

A análise de sobrevivência é uma das áreas da estatística que mais tem crescido nos últimos anos, segundo Therneau & Grambsch (2000), empregando técnicas e modelos estatísticos com o propósito de analisar dados, cuja variável resposta é o tempo até a ocorrência de um evento de interesse, o qual é denominado muitas vezes como falha. Aplicações desta análise estatística vêm sendo empregadas em diversas áreas como: medicina, engenharia, ciências sociais e financeiras. Vários problemas surgem nestas áreas como, por exemplo, o tempo até a morte de um paciente, o tempo até a falha de um componente eletrônico, o tempo de duração de um casamento e o tempo até a inadimplência de um cliente de um banco, entre outros. Uma das características destes dados é que eles são frequentemente censurados, ou seja, apresentam observações incompletas, que por algum motivo não foi possível observar a ocorrência do evento de interesse. Duas razões justificam considerar a censura na modelagem: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida de pacientes; (ii) a omissão das censuras no cálculo das estatísticas de interesse podem acarretar conclusões viciadas (Colosimo & Giolo, 2006).

Em análise de sobrevivência o tempo de falha é especificado pela função de sobrevivência ou pela função taxa de falha (ou risco). A função de sobrevivência é a probabilidade de o indivíduo não falhar até o tempo  $T$ , ou seja, do indivíduo sobreviver ao tempo  $t$ , podendo ser escrita como:

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{\infty} f(u) du, \quad t \geq 0,$$

observe que  $S(t)$  é uma função contínua e monótona decrescente com  $S(0) = 1$  e  $S(\infty) = \lim_{t \rightarrow \infty} S(t) = 0$ . Ocasionalmente, tem-se que  $S(\infty) > 0$  em situações em que é considerado que alguns indivíduos nunca falham.

A função taxa de falha ou risco  $\lambda(t)$  é definida como:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log \{S(t)\},$$

a função de risco especifica a taxa instantânea de falha no tempo  $t$ , dado que o indivíduo sobreviveu até o tempo  $t$ . Segundo Colosimo & Giolo (2006) esta função é bastante útil para descrever a distribuição do tempo de vida de pacientes.

Outra função importante em análise de sobrevivência é a função de risco acumulado, pois fornece o risco acumulado do indivíduo. De acordo com Colosimo & Giolo (2006) esta função não tem interpretação direta, mas é bastante útil para avaliar a função de risco, principalmente no processo de estimação não paramétrica em que a função risco acumulado apresenta propriedades ótimas e a função de risco é difícil de ser estimada. A função de risco acumulado é definida como:

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log \{S(t)\}.$$

Uma das maneiras para modelar dados de sobrevivência na presença de covariáveis é utilizar o modelo de regressão semi-paramétrico proposto por Cox (1972), que também é denominado modelo de riscos proporcionais, pois a razão da taxa de falha de dois indivíduos distintos é constante no tempo. Assumindo que os tempos  $t_i$  sejam independentes, com  $i = 1, \dots, n$ , o modelo de Cox para o  $i$ -ésimo indivíduo, dado o vetor  $\mathbf{x} = (x_1, \dots, x_p)'$  de covariáveis, é dado por:

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) \exp \{\mathbf{x}_i' \boldsymbol{\beta}\}, \quad (1)$$

em que  $\lambda(t|\mathbf{x}_i)$  representa a função taxa de risco,  $\lambda_0(t)$  é o componente não paramétrico, conhecido como função de base ou basal, pois  $\lambda(t) = \lambda_0(t)$  para  $\mathbf{x} = 0$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  é o vetor dimensão  $1 \times p$  de parâmetros associados às covariáveis e  $\mathbf{x}_i$  é o vetor de dimensão  $p \times 1$  de covariáveis observadas para o  $i$ -ésimo indivíduo. A

representação da razão da taxa de falha de dois indivíduos distintos  $r$  e  $j$  é dada por:

$$\frac{\lambda_r(t)}{\lambda_j(t)} = \frac{\lambda_0 \exp \{ \mathbf{x}'_r \boldsymbol{\beta} \}}{\lambda_0 \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \}} = \exp \{ \mathbf{x}'_r \boldsymbol{\beta} - \mathbf{x}'_j \boldsymbol{\beta} \}.$$

Para estimar o parâmetro  $\boldsymbol{\beta}$ , Cox (1972) propôs uma função de verossimilhança e no artigo seguinte, Cox (1975) introduziu o conceito de verossimilhança parcial obtendo a mesma função de verossimilhança do artigo Cox (1972) e estudou as suas propriedades. Suponha que em uma amostra de censura aleatória  $(t_i, \delta_i)$ ,  $i = 1, \dots, n$ , existam  $k \leq n$  falhas distintas nos tempos  $t_1 < \dots < t_k$ . Seja  $R_i = R(t_i)$  o conjunto de indivíduos que estão sob risco no tempo  $t_i$ . Colosimo & Giolo (2006) mostram uma forma simples de entender a função de verossimilhança parcial sem utilizar um processo de contagem, assim, considerando o seguinte argumento condicional: a probabilidade da  $i$ -ésima observação vir a falhar no tempo  $t_i$  conhecendo quais observações estão sob risco em  $t_i$ , então, tem-se:

$$\frac{P(\text{indivíduo falhar em } t_i \mid \text{sobreviveu a } t_i \text{ e história até } t_i)}{P(\text{uma falha em } t_i \mid \text{história até } t_i)} = \frac{\lambda_i(t | \mathbf{x}_i)}{\sum_{j \in R(t_i)} \lambda_j(t | \mathbf{x}_j)} = \frac{\exp \{ \mathbf{x}'_i \boldsymbol{\beta} \}}{\sum_{j \in R(t_i)} \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \}}. \quad (2)$$

Fazendo o produto de todos os termos representados na equação (2) associados aos tempos distintos de falha, obtém-se a seguinte função de verossimilhança:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \left( \frac{\exp \{ \mathbf{x}'_i \boldsymbol{\beta} \}}{\sum_{j=1}^n Y_j(t_i) \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \}} \right)^{\delta_i}, \quad (3)$$

em que a função indicadora de risco  $Y_j(t_i) = I(T_i \geq t)$ ,  $Y_j(t_i) = 1$  se e somente se o indivíduo  $j \in R(t_i)$ , e  $\delta_i$  é uma função indicadora de falha. Aplicando o logaritmo na equação (3), obtém-se a seguinte equação:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left[ \mathbf{x}'_i \boldsymbol{\beta} - \log \left\{ \sum_{j=1}^n Y_j(t_i) \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \} \right\} \right]. \quad (4)$$

Os valores de  $\boldsymbol{\beta}$  que maximizam a função de verossimilhança parcial,  $L(\boldsymbol{\beta})$ , são obtidos resolvendo o sistema de equações definido por

$U(\boldsymbol{\beta}) = \left( \frac{\partial l}{\partial \beta_1}, \dots, \frac{\partial l}{\partial \beta_p} \right)' = 0$ . Definindo para  $t > 0$  o vetor de dimensão  $p \times 1$ , dado por:

$$\bar{\mathbf{x}}(t, \boldsymbol{\beta}) = \frac{\sum_{j=1}^n Y_j(t) \mathbf{x}'_j \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \}}{\sum_{j=1}^n Y_j(t) \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \}}, \quad (5)$$

em que  $\bar{\mathbf{x}}(t, \boldsymbol{\beta})$  é uma média ponderada do vetor de covariáveis dos indivíduos em risco no tempo  $t$ .

Assim, o vetor escore de derivadas  $U(\boldsymbol{\beta})$ , pode ser representado como:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i [\mathbf{x}'_i - \bar{\mathbf{x}}(t_i, \boldsymbol{\beta})]. \quad (6)$$

A matriz de informação  $\mathcal{I}(\boldsymbol{\beta}) = -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$  de dimensão  $p \times p$  pode ser escrita como:

$$\mathcal{I}(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left\{ \frac{\sum_{j=1}^n Y_j(t_i) \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \} [\mathbf{x}'_j - \bar{\mathbf{x}}(t_i, \boldsymbol{\beta})][\mathbf{x}'_j - \bar{\mathbf{x}}(t_i, \boldsymbol{\beta})]'}{Y_j(t_i) \exp \{ \mathbf{x}'_j \boldsymbol{\beta} \}} \right\}. \quad (7)$$

Segundo Lawless (2002) as equações de verossimilhança,  $U(\boldsymbol{\beta}) = 0$ , podem ser resolvidas utilizando o método iterativo de Newton-Raphson ou outros métodos, sendo que numerosos pacotes dão a estimativa  $\hat{\boldsymbol{\beta}}$  e erro padrão, testes ou intervalos de confiança são baseados na aproximação assintótica normal,

$$\hat{\boldsymbol{\beta}} \simeq N_p \left( \boldsymbol{\beta}, \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} \right).$$

Um problema que surge neste processo de estimação é que a função de verossimilhança parcial, dada pela equação (3), assume que o tempo de sobrevivência é contínuo e com isso não contempla a possibilidade de empates dos tempos de sobrevivência observados. Na prática, empates podem ocorrer nos tempos de falha ou de censura devido à escala de medida, ou seja, se for utilizado uma escala discreta (horas, dias, semanas, meses ou anos), e assim é frequente a observação de tempos de sobrevivência iguais para dois ou mais indivíduos no estudo. Para contornar este problema, Breslow (1972) e Peto (1972) propuseram uma aproximação que é frequentemente usada em programas estatísticos. Assim, seja  $\mathbf{s}_i$  o vetor formado pela soma

das  $p$  covariáveis correspondentes para os indivíduos que falham no mesmo tempo  $t_i$  ( $i=1, \dots, n$ ) e  $d_i$  o número de falhas nestes tempos, com  $i=1, \dots, k$ . A aproximação da equação (3) é escrita como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^k \left[ \frac{\exp \{s'_i \boldsymbol{\beta}\}}{\left[ \sum_{j \in R(t_i)} \exp \{x'_j \boldsymbol{\beta}\} \right]} \right]^{d_i}, \quad (8)$$

sendo esta aproximação adequada quando o número de observações empatadas em qualquer tempo não é muito grande.

## 2.2 Processos de Contagem

Uma forma de tratar situações mais complexas em análise de sobrevivência, tais como covariáveis dependentes do tempo e múltiplos eventos, pode ser feita através dos processos de contagem (processo estocástico). O processo de contagem informa sobre quando o evento ocorreu e também sobre os indivíduos estudados em um tempo  $t$ . Para os dados censurados à direita, tem-se o conhecimento de como os indivíduos tem sido censurados antes do tempo  $t$  e os que morreram neste mesmo tempo ou antes dele.

Na análise de sobrevivência univariada observa-se o par  $(T_i, \delta_i)$  (com  $i=1, \dots, n$ ), em que  $T_i = \min(T_i^*, C_i^*)$ , sendo  $T_i^*$  o tempo de falha e  $C_i^*$  o tempo até ocorrência de censura, e

$$\delta_i = \begin{cases} 1, & \text{se ocorre a falha,} \\ 0, & \text{se ocorre a censura.} \end{cases}$$

Em um cenário de eventos múltiplos, o tempo  $T_i^*$  pode representar mais de um evento de interesse, e assim é necessário uma formulação de processo de contagem que substitui o par  $(T_i, \delta_i)$  pelo par de funções  $(N_i(t), Y_i(t))$ , em que

$N_i(t)$  = o número de eventos observados em  $[0, t]$  para o indivíduo  $i$

$$Y_i(t) = \begin{cases} 1, & \text{se o indivíduo } i \text{ está sob observação e em risco no tempo } t \\ 0, & \text{caso contrário.} \end{cases}$$

Nota-se que  $N_i(t)$  é um processo de contagem que simplesmente conta o número de mortes na amostra no tempo  $t$  ou antes deste tempo.

De acordo com Therneau & Grambsch (2000) esta formulação de processo de contagem generaliza imediatamente o problema para eventos múltiplos e múltiplos intervalos em risco.

### 2.3 Análise de Sobrevida Multivariada

Segundo Therneau & Grambsch (2000) há um crescente interesse e necessidade de aplicar a análise de sobrevivência para um conjunto de dados com múltiplos eventos por indivíduo, isso se deve a ênfase dada sobre a qualidade de vida, reinternações e outros desfechos secundários, e assim, tais análises se tornaram comuns. Estes eventos são caracterizados de duas formas: eventos ordenados (múltiplos eventos do mesmo tipo) ou eventos não ordenados (eventos de diferentes tipos). Neste trabalho será tratado apenas de eventos ordenados. Alguns exemplos em que o evento de interesse pode ocorrer mais de uma vez para o mesmo indivíduo são acidentes vasculares cerebrais, infarto do miocárdio, câncer, ataques epiléticos e infecções respiratórias agudas. Em todas essas aplicações o tempo de falha pode ser um processo repetido várias vezes para o mesmo indivíduo o que faz com que suas observações não sejam necessariamente mutuamente independentes. A suposição de dependência entre os tempos observados é razoável, uma vez que pode haver associação entre estes tempos e estes serem influenciados por fatores não medidos intra-indivíduo. Nestas condições não é recomendado o uso do modelo de Cox na forma usual, mesmo se utilizarmos a formulação por processo de contagem, pois os intervalos de tempo para o mesmo indivíduo se sobrepõem e tempos de falha do mesmo indivíduo podem ser correlacionados (Carvalho et al., 2011). Há outras abordagens na literatura para tratar este problema de múltiplos eventos por sujeito de uma forma mais realista. As abordagens mais utilizadas e inseridas em programas estatísticos padrões são: a modelagem marginal e a modelagem condicional.

### 2.3.1 Modelagem Marginal

A ideia básica dos modelos marginais é determinar as estimativas dos parâmetros do modelo de Cox ignorando a correlação das observações de um mesmo indivíduo, seguido por uma correção na variância dos parâmetros estimados de modo que se obtenha uma estimativa robusta e confiável da variância dos parâmetros. Neste caso estamos interessados na resposta média da população, modelada como uma função das covariáveis (Carvalho et al., 2011). Uma das formas de realizar este tipo de modelagem é utilizando uma estrutura estocástica, sendo a mais plausível, segundo McLain & Peñay (2008), a proposta de Andersen & Gill (1982) a qual inseriram o modelo de Cox em um processo de contagem provando a consistência e a normalidade assintótica. Os modelos mais utilizados nesta abordagem são os modelos AG (Andersen & Gill, 1982), PWP (Prentice et al., 1981) e WLW (Wei et al., 1989). Estes três modelos diferem consideravelmente entre eles e na criação do conjunto de risco. Aplicações desses modelos podem ser encontradas, por exemplo, em Therneau & Grambsch (2000), Colosimo & Giolo (2006), Carvalho et al. (2011) e Kleinbaum & Klein (2011).

**Modelo de Andersen e Gill (AG)** - O modelo AG proposto por Andersen & Gill (1982) considera que o risco basal é igual em todos os intervalos de tempo analisados, sendo que o indivíduo retorna ao grupo de risco após cada evento e assume que os eventos em cada intervalo disjunto são independentes. Dependendo da escala de medida do tempo, a primeira observação poderá ou não começar no zero. Se a primeira observação começar no tempo de entrada, o modelo para o  $i$ -ésimo indivíduo fica representado por:

$$\lambda_i(t) = Y_{mi} \lambda_0(t) \exp \{x'_{mi}(t) \boldsymbol{\beta}\}, \quad (9)$$

em que  $\lambda_i(t)$  é a função de risco para o  $i$ -ésimo indivíduo,  $\lambda_0(t)$  é o componente não paramétrico,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$  é o vetor de dimensão  $p \times 1$  de parâmetros associado às covariáveis,  $x'_{mi}(t)$  é o vetor de dimensão  $1 \times p$  de covariáveis observadas para o

$i$ -ésimo indivíduo no tempo  $t$  e a função indicadora de risco

$$Y_{mi} = \begin{cases} 1, & \text{se o indivíduo } i \text{ estiver sob observação e em risco no tempo } t \\ 0, & \text{caso contrário.} \end{cases}$$

O uso do modelo AG é apropriado no caso em que se tem independência mútua entre as observações em relação a um mesmo indivíduo. Isto significa que o risco é comum para todos os eventos, o que é bastante forte, pois assume que o histórico do indivíduo não afeta o risco presente. Carvalho et al. (2011) sugerem que para testar a suposição de independência, ou seja, que não há uma estrutura de correlação das observações de um mesmo indivíduo (desde que as covariáveis incluídas no modelo expliquem as diferenças entre indivíduos), é recomendado verificar se a variância robusta das estimativas dos parâmetros do modelo AG é um pouco maior do que a do modelo de Cox usual.

**Modelo de Prentice, Williams e Peterson (PWP)** - O modelo PWP proposto por Prentice et al. (1981) pressupõe que o indivíduo só estará em risco de experimentar o  $m$ -ésimo evento depois que tenha experimentado o evento  $m-1$ . O modelo PWP separa a análise em diferentes estratos, assumindo que existe uma dependência entre os tempos de falha de um mesmo indivíduo. O uso de estratos dependentes significa que a função de risco pode variar de um evento para outro, ao contrário do que ocorre no modelo AG. Assim, por exemplo, o risco basal para o segundo evento é zero até que o primeiro evento ocorra, enquanto que o risco do terceiro evento é zero até que o segundo evento ocorra, e assim sucessivamente. O modelo pode ser escrito da seguinte forma:

$$\lambda_{mi}(t) = Y_{mi} \lambda_{0m}(t) \exp \{ \mathbf{x}'_{mi}(t) \boldsymbol{\beta}_m \}, \quad (10)$$

em que  $\lambda_{mi}(t)$  é a função de risco do  $m$ -ésimo evento desde que o  $i$ -ésimo indivíduo tenha experimentado os  $m-1$  eventos,  $\lambda_{0m}(t)$  é a função de base que pode variar de um evento para outro,  $\boldsymbol{\beta}_m = (\beta_1, \dots, \beta_p)$  é o vetor de dimensão  $p \times 1$  de parâmetros associados às covariáveis no  $m$ -ésimo evento,  $\mathbf{x}'_{mi}(t)$  é o vetor de dimensão  $1 \times p$  de covariáveis observadas para o  $i$ -ésimo indivíduo no tempo  $t$  e a função indicadora de



risco  $Y_{mi}$  é zero até ocorrer o evento  $m - 1$ , ocorrendo este evento a função assume o valor 1. De acordo com Lim et al. (2007) o modelo PWP é capaz de avaliar o efeito das covariáveis condicionalmente sobre a história do passado do indivíduo (ocorrência de eventos anteriores) facilitando a predição ou efeito do estudo de covariáveis medidas no tempo. Kelly & Lim (2000) observaram, em seus estudo de simulação, que os erros padrão usual e robustos das estimativas dos parâmetros parecem ser similares independente da correlação intra-indivíduo.

**Modelo de Wei, Lin e Weissfeld (WLW)** - O modelo WLW proposto por Wei et al. (1989) trata as respostas de um conjunto de dados ordenados como se fosse um problema de riscos competitivos com respostas não ordenadas, ou seja, o indivíduo no início do período de observação é considerando estar em risco de sofrer  $m$  eventos e não é utilizado uma estrutura de processo contagem (o tempo é sempre contado a partir do zero). A ordem neste caso é dada pela enumeração de cada evento. Como este modelo pressupõe que todos os indivíduos iniciam o estudo em risco de sofrer  $m$  eventos é necessário criar observações fictícias para aqueles indivíduos que não tenham experimentado este número de eventos. Assim, por exemplo, em um estudo com  $n$  indivíduos e  $m$  eventos, o banco de dados apresentará  $nm$  linhas. O modelo é representado pela seguinte função de risco do  $m$ -ésimo evento do  $i$ -ésimo indivíduo:

$$\lambda_{mi}(t) = Y_{mi}\lambda_{0m}(t) \exp \{ \mathbf{x}'_{mi}(t) \boldsymbol{\beta}_m \}, \quad (11)$$

em que  $\lambda_{mi}(t)$  é a função de risco do  $m$ -ésimo evento do  $i$ -ésimo indivíduo,  $\lambda_{0m}(t)$  é o componente não paramétrico no  $m$ -ésimo evento,  $\boldsymbol{\beta}_m = (\beta_1, \dots, \beta_p)$  é o vetor de dimensão  $p \times 1$  de parâmetros associados às covariáveis no  $m$ -ésimo evento e  $\mathbf{x}'_{mi}(t)$  é o vetor de dimensão  $1 \times p$  de covariáveis observadas para o  $i$ -ésimo indivíduo no tempo  $t$ . A função indicadora de risco  $Y_{mi}$  assume o valor 1 até a ocorrência do  $m$ -ésimo evento, ao menos que, algum fato origine censura.

Note que, no modelo WLW tem-se uma função de risco para cada evento e para cada estrato, como é indicado pelo vetor de parâmetros  $\boldsymbol{\beta}_m$ , o que não ocorre no AG. Em relação ao modelo PWP, a primeira diferença pode ser vista no

conjunto de risco e na definição dos estratos na análise. No modelo PWP o indivíduo necessariamente deverá ter  $m-1$  eventos para poder experimentar o  $m$ -ésimo evento recorrente, enquanto que no modelo WLW o indivíduo está em um conjunto de risco de  $m$  eventos no tempo  $t$ .

Segundo Kleinbaum & Klein (2011) o modelo WLW concentra-se no tempo de sobrevivência total de um estudo até a ocorrência do  $m$ -ésimo evento específico, sendo esta abordagem recomendada quando o pesquisador quer considerar que os eventos ocorreram em diferentes ordens, bem como diferentes tipos de eventos, por exemplo, diferentes condições da doença.

Pode-se observar com isso que os modelos AG e PWP são usados para respostas do mesmo tipo, enquanto que o modelo WLW é usado tanto para respostas do mesmo tipo como para múltiplos eventos de diferentes tipos, mesmo quando não há uma ordem predeterminada. Nos últimos anos, alguns estudos vêm discutindo o uso do modelo WLW em dados de eventos recorrentes. Baseando-se em estudos de simulação, Kelly & Lim (2000) não recomendam o uso do modelo WLW para eventos recorrentes, pois o modelo tem um conjunto de risco que permite que os indivíduos estejam em risco de  $m$  eventos, mesmo que um indivíduo tenha experimentado apenas um evento, o que conduz a superestimação do efeito do tratamento e assim eles acreditam que este modelo é mais apropriado para dados onde há diferentes tipos de eventos para uma mesma pessoa. Therneau & Grambsch (2000), em estudos de simulação, encontraram resultados que sugerem que o modelo WLW pode violar a suposição de riscos proporcionais, mesmo quando isso não ocorre para o conjunto de dados em geral. Eles apontam que a interpretação do modelo não é muito clara e o fato de pensar que o indivíduo é considerado estar em risco de  $m$  eventos, mesmo antes de ele ter experimentado os  $m-1$  eventos é um tanto desconfortável. Outro ponto levantado por eles baseando-se em estudos de simulação em que o modelo é bem especificado (nenhuma covariável importante foi omitida), os modelos AG e PWP fornecem estimativas não viciadas do efeito do tratamento e requerem um tamanho de amostra similar para obter a mesma precisão na estimativa, enquanto

que o WLW fornece estimativas viciadas do efeito do tratamento e necessita de um tamanho de amostra maior.

Metcalf & Thompson (2007) compararam os modelos PWP e WLW em estudos de simulações com dados reais. Do ponto de vista destes autores a aplicação do modelo WLW é justificada em dados de eventos recorrentes, uma vez que o modelo fornece estimativas que requerem distintas interpretações comparadas com as do modelo PWP.

Em relação aos modelos AG e PWP, considerando novamente os estudos de simulação de Therneau & Grambsch (2000) foram observados que o modelo AG fornece estimativas não viciadas com maior precisão, mesmo quando uma covariável importante é omitida. As estimativas são, ainda, corrigidas satisfatoriamente por estimativas robustas. E em relação ao modelo PWP, este modelo na ausência de covariáveis importantes fornece estimativas viciadas, o que deve ser uma preocupação em pesquisas médicas quando covariáveis importantes não são medidas.

Mesmo com todas essas questões levantadas, defensores da abordagem marginal defendem a sua utilização, devido a falta de pressuposto sobre a estrutura de dependência entre os eventos. Segundo Therneau & Grambsch (2000) esses modelos não são perfeitos, mas fornecem importantes informações.

Para a estimação dos parâmetros nos modelos AG, PWP e WLW, o método de máxima verossimilhança parcial é utilizado ignorando a correlação entre as observações. Suponha que há  $n$  indivíduos e cada indivíduo possa experimentar  $m$  falhas. Suponha que a censura é não informativa, ou seja, a censura não fornece nenhuma informação adicional na probabilidade de sobrevivência do indivíduo em um tempo futuro. Seja  $T_{mi}$  o tempo quando a  $m$ -ésima falha ocorre para o  $i$ -ésimo indivíduo e  $C_{mi}$  é o tempo de censura associado ao  $i$ -ésimo indivíduo, medida quando ele inicia o estudo até o último tempo de falha ( $C_{mi}$  é uma censura à direita). O que se observa para o  $i$ -ésimo indivíduo é que  $t = \min(T_{mi}, C_{mi})$  e

$$\delta_{mi} = \begin{cases} 1 & , \text{ se } T_{mi} \leq C_{mi} \\ 0 & , \text{ se } T_{mi} > C_{mi}. \end{cases}$$

A função de verossimilhança parcial para o modelo AG, dado pela equação (9), pode ser representada da seguinte forma:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \prod_{m=1}^{k_i} \left( \frac{Y_{mi}(t) \exp \{ \mathbf{x}'_{mi} \boldsymbol{\beta} \}}{\sum_{j=1}^n \sum_{l=1}^{k_l} Y_{lj}(t) \exp \{ \mathbf{x}'_{lj} \boldsymbol{\beta} \}} \right)^{\delta_{mi}}. \quad (12)$$

Para o modelo PWP e o modelo WLW, dados nas equações (10) e (11) respectivamente, a função de verossimilhança parcial é a mesma e é dada por:

$$L(\boldsymbol{\beta}_m) = \prod_{i=1}^n \prod_{m=1}^{k_i} \left( \frac{Y_{mi}(t) \exp \{ \mathbf{x}'_{mi} \boldsymbol{\beta}_m \}}{\sum_{j=1}^n Y_{mj}(t) \exp \{ \mathbf{x}'_{mj} \boldsymbol{\beta}_m \}} \right)^{\delta_{mi}}. \quad (13)$$

Observe que a diferença da função de verossimilhança parcial do modelo PWP para o modelo WLW está na função indicadora de risco  $Y_{mi}$ .

Em seguida é feita uma correção na variância dos  $\hat{\boldsymbol{\beta}}$  por uma aproximação da estimativa jackknife, para obter uma estimativa robusta. Essencialmente, o jackknife fornecerá uma estimativa não viciada da variância para dados correlacionados sempre que a observação deixada de fora em qualquer passo for independente das observações que entrarem, ou seja, uma estimativa jackknife agrupada por indivíduo deixará de fora um sujeito em um tempo, em vez de uma observação no tempo.

Para obter uma estimativa de jackknife agrupado por indivíduo utilizam-se os resíduos de jackknife que são definidos como:

$$\mathbf{J}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)},$$

em que  $\hat{\boldsymbol{\beta}}_{(i)}$  é o resultado do ajuste que inclui todas as observações exceto o indivíduo  $i$ . Therneau & Grambsch (2000) descrevem uma maneira de calcular os valores dos resíduos de jackknife pelo método de Newton-Raphson, que é reescrito como:

$$\Delta \boldsymbol{\beta} = \mathbf{1}'(\mathbf{U}\mathcal{I}^{-1}) \equiv \mathbf{1}'\mathbf{D},$$

em que  $\mathbf{U}$  é a matriz de escore residual. Assim a mudança em  $\hat{\boldsymbol{\beta}}$  em cada iteração é a soma da coluna da matriz  $\mathbf{D}$ , definida como escore residual dimensionada pela matriz  $\mathcal{I}^{-1}$  (variância dos  $\hat{\boldsymbol{\beta}}$ ).

A estimativa de jackknife agrupada por indivíduo pode ser escrita pela seguinte matriz:

$$\mathbf{V}_j = \frac{n-1}{n}(\mathbf{J} - \bar{\mathbf{J}})'(\mathbf{J} - \bar{\mathbf{J}}),$$

em que  $\bar{\mathbf{J}}$  é a matriz de médias das colunas de  $\mathbf{J}$ . Ignorado o termo  $\frac{n-1}{n}$ , uma aproximação da variância de Jackknife é dado por  $\mathbf{D}'\mathbf{D}$ . Esta variância pode ser escrita como  $\mathbf{D}'\mathbf{D} = \mathcal{I}^{-1}(\mathbf{U}'\mathbf{U})\mathcal{I}^{-1}$ , que pode ser visto como um estimador sanduíche  $\mathbf{A}\mathbf{B}\mathbf{A}$ , em que  $\mathbf{A} = \mathcal{I}^{-1}$  é a matriz de variância usual e  $\mathbf{B} = \mathbf{U}'\mathbf{U}$  é o termo de correção. No software R a chave para inserir esses modelos não alterando a estimativas do modelo de Cox e obtendo a variância robusta é acrescentado o termo `cluster()` na função `coxph()` do pacote `Survival` (Therneau, 2012).

É possível também, nos modelos marginais utilizar testes para comparar os diferentes modelos ajustados. Os testes mais utilizados na literatura são os testes de Wald e o de razão de verossimilhanças.

O teste de Wald, segundo Colosimo & Giolo (2006) é geralmente utilizado para testar hipóteses relativas a um único parâmetro  $\beta_j$ . Considerando a hipótese nula:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0,$$

a estatística para esse teste é dada por:

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \left( \frac{\partial^2 \log \{L(\boldsymbol{\beta}_0)\}}{\partial \boldsymbol{\beta}_0^2} \right)^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

em que, sob  $H_0$ , tem aproximadamente uma distribuição qui-quadrado com  $p$  graus de liberdade ( $\chi_p^2$ ). Para uma única covariável este teste, se reduz ao teste t usual, cuja estatística é  $z = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\widehat{\text{var}}(\hat{\beta}_j)}}$ , que sob a hipótese nula, essa estatística tem distribuição normal padrão.

O teste de razão de verossimilhanças envolve a comparação dos valores do logaritmo da função de verossimilhança  $\log \{L(\hat{\boldsymbol{\beta}})\}$  e  $\log \{L(\boldsymbol{\beta}_0)\}$ . A estatística para este teste é dada por:

$$\text{TRV} = -2 \left[ \log \left\{ \frac{L(\boldsymbol{\beta}_0)}{L(\hat{\boldsymbol{\beta}})} \right\} \right] = 2 \left[ \log \{L(\hat{\boldsymbol{\beta}})\} - \log \{L(\boldsymbol{\beta}_0)\} \right],$$

que, sob  $H_0 : \beta = \beta_0$ , segue aproximadamente uma distribuição qui-quadrado com  $p$  graus de liberdade.

Para fazer a seleção apropriada dos modelos, Lim et al. (2007) levantam alguns pontos importantes que devem ser considerados na seleção, tais como: a distribuição dos tempos de falha, se o interesse da pesquisa se encontra no efeito das covariáveis condicionadas sobre a história do indivíduo, no efeito médio populacional ou na medida dependência da recorrência intra-indivíduo. Eles apontam que quando a dependência de eventos recorrentes anteriores é forte e consistente, o uso do modelo PWP é apropriado. Mas quando há suposição de risco comum para as observações ao longo das recorrências e quando há o interesse em uma taxa de recorrência global, o modelo AG é apropriado. Eles consideram também que a frequência dos eventos é importante como critério de seleção. Assim, quando a frequência de eventos recorrentes por sujeito é pequena pode-se assumir que o risco da ocorrência de um evento pode variar substancialmente, e então o modelo PWP é apropriado. Por outro lado, quando a frequência de eventos recorrentes por sujeito é grande, o risco tenderá a variar menos entre os eventos, e neste caso o modelo AG seria recomendado. Quando se tem o interesse de especificar o efeito da covariável sobre ordem natural do evento recorrente, o modelo WLW não é recomendado, pois permite que o indivíduo esteja em risco de sofrer mais eventos do que ele poderia realmente experimentar. Portanto, neste caso se o interesse é estimar o efeito médio de uma covariável então os modelos AG e PWP são mais apropriados. Por outro lado, se o interesse da pesquisa for medir a correlação entre os eventos recorrentes intra-indivíduo, os modelos condicionais seriam a opção mais apropriada.

Para selecionar os modelos marginais ajustados e os modelos condicionais (que serão abordados na seção 2.3.2), pode-se também utilizar os critérios de Informação de Akaike (AIC), AKAIKE (1974), e o Critério de Informação de Bayes (BIC), Schwarz (1978). A expressão do AIC é dado por:

$$\text{AIC} = -2 \log \{L(\hat{\beta})\} + 2K, \quad (14)$$

em que  $K$  é o número de parâmetros estimáveis do modelo. Quando o AIC é utili-

zado seleciona-se o modelo com o menor valor de AIC, fazendo com que este modelo seja o melhor modelo segundo este critério para ajustar o conjunto de dados estudado. Um outro método baseado no critério de AIC é o Critério de Informação de Akaike Corrigido ( $AIC_c$ ), que tem a mesma expressão do AIC, mas com um termo de correção do vício. Segundo Burnham & Anderson (2002) este critério leva em conta no termo de penalização o tamanho da amostra e a complexidade do modelo (números parâmetros), sendo que seu uso não é recomendado quando a razão  $\frac{n}{K} < 40$ , em que  $n$  é o tamanho da amostra. A expressão do  $AIC_c$  é dada por:

$$AIC_c = -2 \log \{L(\hat{\beta})\} + 2K \frac{n}{n - K - 1}. \quad (15)$$

Uma observação importante é que quando a razão  $\frac{n}{K}$  é suficientemente grande o critério de AIC e  $AIC_c$  são similares e tenderão a selecionar o mesmo modelo.

A expressão do critério de informação de Bayes é dado por:

$$BIC = -2 \log \{L(\hat{\beta})\} + K \log \{n\}. \quad (16)$$

Segundo Burnham & Anderson (2002) o BIC surge do ponto de vista Bayesiano com probabilidade a priori igual sobre cada modelo. Por esse critério seleciona-se o modelo com o menor valor de BIC para ajustar o conjunto de dados estudado.

Uma estratégia para selecionar as covariáveis do modelo é através de rotinas automáticas conhecidas na literatura como métodos *forward*, *backward* ou *stepwise*. Esses métodos estão implementados no programa R, sendo que este programa utiliza o AIC nesses métodos como forma de selecionar o melhor conjunto de covariáveis que possa explicar a variável resposta. A função utilizada no programa R para tratar destes métodos é a função `stepAIC` do pacote `MASS` (Venables & Ripley, 2002).

Outras questões importantes, a serem estudadas nos modelos marginais, são a suposição de riscos proporcionais e a qualidade do ajuste dos modelos. Lin (1994) observa que é importante avaliar a adequação dos modelos marginais, verificando a suposição de riscos proporcionais e utilizando técnicas baseadas nos

resíduos de martingale. Carvalho et al. (2011) afirmam que a verificação das premissas descritas anteriormente, podem ser verificadas exatamente como nos modelos com covariáveis mudando no tempo utilizando uma formulação de processos de contagem. Sendo assim, uma das formas para investigar a suposição de riscos proporcionais é analisar os resíduos de Schoenfeld (1982). Considerando que o  $i$ -ésimo indivíduo com vetor de covariáveis  $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{pi})'$ , é observado falhar, tem-se para este indivíduo um vetor de resíduos de Schoenfeld  $\mathbf{r}_i = (r_{i1}, r_{i2}, \dots, r_{ip})$  em que cada componente  $r_{iq}$ , para  $q = 1, \dots, p$  é definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp \{ \mathbf{x}'_j \hat{\boldsymbol{\beta}} \}}{\sum_{j \in R(t_i)} \exp \{ \mathbf{x}'_j \hat{\boldsymbol{\beta}} \}}, \quad (17)$$

sendo que os resíduos são definidos para cada falha e não definidos para censura, assim, o conjunto de resíduos de Schoenfeld é uma matriz com  $d$  linhas, representando as falhas, e  $p$  colunas, representando as covariáveis consideradas no modelo. Vale a pena notar que,  $\sum_i \mathbf{r}_i = 0$ . Para permitir que a estrutura de correlação dos resíduos seja considerada, uma forma padronizada dos resíduos de Schoenfeld é frequentemente usada e é definida por:

$$s_i^* = [\mathcal{I}(\hat{\boldsymbol{\beta}})]^{-1} \times \mathbf{r}_i, \quad (18)$$

com  $\mathcal{I}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n \int_0^\infty \frac{\sum_i Y_i(s) \exp \{ \mathbf{x}'_i(s) \boldsymbol{\beta} \} [\mathbf{x}_i(s) - \bar{\mathbf{x}}(s_i, \boldsymbol{\beta})] [\mathbf{x}_i(s) - \bar{\mathbf{x}}(s_i, \boldsymbol{\beta})]'}{\sum_i Y_i(s) \exp \{ \mathbf{x}'_i(s) \boldsymbol{\beta} \}} dN_i(s)$ , a matriz de informação observada para o modelo de Cox inserido em um processo de contagem. Note que o vetor  $\bar{\mathbf{x}}(s_i, \boldsymbol{\beta})$  é uma média ponderada do vetor de covariáveis dos indivíduos em risco no tempo  $s_i$ , como foi definido na equação (5), e com as devidas mudanças a formulação de processo de contagem substitui o par  $(T_i, \delta_i)$  pelo par de funções  $(N_i(t), Y_i(t))$  onde:

$$\begin{aligned} N_i(t) &= \text{o número observado de eventos em } [0, t] \text{ para o indivíduo } i; \\ Y_i(t) &= \begin{cases} 1, & \text{se o indivíduo } i \text{ está sob observação e em risco no tempo } t \\ 0, & \text{caso contrário.} \end{cases} \end{aligned}$$

Esta formulação inclui dados de sobrevivência a direita como um caso especial. Pode-se escrever o que foi dito anteriormente em notação matemática da seguinte forma



$N_i(t) = I(\{T_i \leq t, \delta_i = 1\})$  e  $Y_i(t) = I(\{T_i \geq t\})$ .

O uso dos resíduos de Schoenfeld para avaliar a suposição de riscos proporcionais é baseado numa restrição  $\beta(t) = \beta$  do modelo

$$\lambda_i(t) = Y_i \lambda_0(t) \exp \{x_i' \beta(t)\},$$

apresentado por Grambsch & Therneau (1994). A condição  $\beta(t) = \beta$  implica em proporcionalidade do risco e quando  $\beta(t)$  não é constante, o impacto de uma ou mais covariáveis pode variar no tempo. Se a suposição de riscos proporcionais for válida o gráfico de  $\beta_q(t)$  versus  $t$  deve ser uma linha horizontal, com  $q = 1, \dots, p$ . Grambsch & Therneau (1994) sugerem o gráfico  $s_i^*$  versus  $t$ , ou alguma função do tempo,  $g(t)$ , como um método de visualizar a suposição de riscos proporcionais. Inclinação zero mostra evidências a favor da proporcionalidade do risco. Para auxiliar a detecção da suposição de riscos proporcionais, o programa R Development Core Team (2012) fornece uma curva suavizada, com bandas de confiança baseada em *splines*. Segundo Colosimo & Giolo (2006) esta técnica gráfica como qualquer outra fornece conclusões subjetivas, pois depende da interpretação dos gráficos. Uma maneira alternativa, de avaliar a suposição de riscos proporcionais é através dos testes de hipóteses. O teste de hipótese, além de fornecer as estatísticas e o valor  $p$  para cada covariável, ele fornece uma estimativa do coeficiente de correlação de Person ( $\rho$ ) entre os resíduos de Schoenfeld e  $g(t)$ , em que valores de  $\hat{\rho}$  próximos de zero mostram não haver evidência contra a suposição de riscos proporcionais. Um teste para testar a hipótese global de proporcionalidade dos riscos sobre todas as covariáveis no modelo, assumindo  $g_q(t) = g(t)$ , pode ser realizado utilizando a seguinte estatística:

$$T = \frac{(g(t) - \bar{g}(t))' S^* \mathcal{I} S^{*'} (g(t) - \bar{g}(t))}{d \sum_{k=1}^d (g(t)_k - \bar{g}(t))^2}, \quad (19)$$

em que  $\mathcal{I}$  é a matriz de informação observada,  $d$  é o número de falhas e  $S^* = dR\mathcal{I}^{-1}$ , sendo  $R$  a matriz  $d \times p$  dos resíduos de Schoenfeld não padronizados. Sob a hipótese nula de proporcionalidade dos riscos,  $T$  tem aproximadamente distribuição qui-quadrado ( $\chi^2$ ) com  $p$  graus de liberdade e os valores  $T > \chi_{p,1-\alpha}^2$  mostram evidências contra a suposição de riscos proporcionais.

A hipótese de riscos proporcionais para a  $q$ -ésima covariável,  $q = 1, \dots, p$ , pode ser testada utilizando a estatística:

$$T_q = \frac{\left[ \sum_{k=1}^d (g_k(t) - \bar{g}(t)) s_{qk}^* \right]^2}{d \mathcal{I}_q^{-1} \sum_{k=1}^d (g_k(t) - \bar{g}(t))^2}, \quad (20)$$

em que  $\mathcal{I}_q^{-1}$  é o  $q$ -ésimo elemento da diagonal do inverso da matriz de informação observada. Sob a hipótese nula de riscos proporcionais para  $q$ -ésima covariável,  $T_q$  tem aproximadamente distribuição qui-quadrado com 1 grau de liberdade. Valores  $T_q > \chi_{1,1-\alpha}^2$  mostram evidências contra a suposição de riscos proporcionais para a covariável  $q$ .

Segundo Carvalho et al. (2011), quando é utilizado um processo de contagem, a função  $g(t) = t$  é normalmente usada, pois sob o processo de contagem, em geral os eventos estão melhor distribuídos ao longo do tempo porque seguem o tempo calendário, não havendo necessidade de uma transformação, por exemplo,  $\log(t)$ .

Outra premissa que pode ser avaliada nesses modelos marginais são os resíduos de martigale. Por esses resíduos pode-se avaliar a presença de pontos aberrantes no modelo e a forma funcional. Os resíduos de martigale são definidos por:

$$M_i = N_i - E_i \quad i = 1, \dots, n,$$

tal que  $N_i$  é igual ao número de eventos observados no intervalo  $[0, \infty)$  e  $E_i$  é o número de eventos esperados sob o modelo ajustado no intervalo  $[0, \infty)$ . No caso especial em que os dados apresentam censura à direita e as covariáveis são fixadas no início do estudo, ou seja, não são dependentes do tempo, os resíduos de martigale podem ser escrito como:

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i) \exp \left\{ \sum_{k=1}^p \mathbf{x}_i \hat{\boldsymbol{\beta}} \right\} = \delta_i - \hat{e}_i, \quad (21)$$

em que  $\hat{\Lambda}_0(t_i) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j=1}^n Y_j(t_i) \exp \{ \mathbf{x}_j' \hat{\boldsymbol{\beta}} \}}$  (estimador proposto por Breslow (1972)) e  $\hat{e}_i$  é denominado resíduos de Cox-Snell. Mais informações desses resíduos podem ser encontradas em Therneau & Grambsch (2000) e Carvalho et al. (2011).

Segundo Colosimo & Giolo (2006) os resíduos de martingale são frequentemente usados para verificar a presença de pontos atípicos (*outliers*) e também para verificar a forma funcional das covariáveis, ou seja, se as covariáveis devem ser usadas como, por exemplo,  $\log(x_i)$ ,  $x_i^2$ , ao invés de usar  $x_i$  ou mesmo categorizada. Assim para uma variável contínua  $x_q$ , o gráfico  $M_i$  versus  $x_{i,q}$  é utilizado para que se possa avaliar a forma funcional desta covariável. Outro uso deste resíduo, como já foi dito, é para verificar a presença de pontos atípicos que pode ser feito fazendo o gráfico  $M_i$  versus índice do indivíduo. Caso haja o interesse em calcular um único resíduo para cada indivíduo é necessário utilizar o argumento `collapse=id` no comando `resid(modelo,type=martingale)` do pacote `Survival` (Therneau, 2012) do programa R, em que `id` é a variável de identificação do indivíduo.

### 2.3.2 Modelagem Condicional

Outra forma de tratar a associação das observações de um mesmo indivíduo em análise de sobrevivência é através dos modelos condicionais. O objetivo destes modelos está em tratar os tempos que apresentam uma possível associação intra-indivíduo como independentes condicionalmente às variáveis de fragilidade. A fragilidade ou efeito aleatório tenta explicar a correlação das observações intra-indivíduo e é frequentemente assumida, constante sobre o tempo e comum para todas as observações do mesmo indivíduo, e assim responsável por criar a dependência entre os eventos intra-indivíduo. Segundo Kleinbaum & Klein (2011) o componente aleatório é designado para explicar a variabilidade que não é explicada por outros preditores no modelo. Os modelos que podem ser usados neste contexto são os de riscos proporcionais e os modelos paramétricos com fragilidade compartilhada. Neste trabalho será tratado apenas dos modelos de risco proporcionais. Aplicações desses modelos e outras generalizações dos modelos condicionais podem ser encontradas em Therneau & Grambsch (2000), Kleinbaum & Klein (2011), Carvalho et al. (2011), Colosimo & Giolo (2006), Wienke (2010) e Hanagal (2011).

#### Modelos de riscos proporcionais com fragilidade Comparti-

**lhada** - Os modelos de fragilidade compartilhada têm como objetivo explicar a correlação dos dados devido a fatores não observáveis intra-indivíduo. Assumindo que os dados para  $m$ -ésima observação do indivíduo  $i$  segue um modelo de fragilidade compartilhada de riscos proporcionais, tem-se o seguinte modelo:

$$\lambda_{mi}(t) = z_i \lambda_0(t) \exp \{ \mathbf{x}'_{mi} \boldsymbol{\beta} \}, \quad (22)$$

para  $i = 1, \dots, n$  e  $m = 1, \dots, k_i$ , tem-se que  $\lambda_{mi}(t)$  é a função de risco para o tempo  $t$  da  $m$ -ésima observação do  $i$ -ésimo indivíduo condicionalmente ao valor do efeito aleatório  $z_i$  para o indivíduo  $i$ ,  $\mathbf{x}'_{mi}$  é o vetor de dimensão  $p$  de covariáveis que contém a informação das covariáveis no tempo  $t$  da  $m$ -ésima observação do  $i$ -ésimo indivíduo,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$  é o vetor de dimensão  $p \times 1$  de parâmetros associados às covariáveis. Note que os valores da fragilidade  $z_i$  são considerados serem amostras independentes de variáveis aleatórias  $Z_i$ , assumidas não variarem com o tempo, com distribuição de probabilidade conhecida com média 1 e alguma variância desconhecida, digamos  $\psi$ .

O modelo apresentado na equação (22) pode ser reescrito da seguinte forma:

$$\lambda_{mi}(t) = \lambda_0(t) \exp \{ \mathbf{x}'_{mi} \boldsymbol{\beta} + \omega_i \}, \quad (23)$$

em que  $z_i = \exp \{ \omega_i \}$ . Neste modelo é pressuposto que  $\omega_i$ 's são uma amostra independente de alguma distribuição com média 0 e variância  $\zeta$ , de modo que quando  $\zeta = 0$ , tem-se o modelo de riscos proporcionais, proposto por Cox (1972).

As distribuições mais utilizadas e que são inseridas no programa R para a variável de fragilidade  $z_i$  do modelo apresentado nas equações (22) e (23) são as distribuições gama e lognormal. A escolha da distribuição de probabilidade da variável de fragilidade é feita por razões práticas. Do ponto de vista computacional e analítico a distribuição gama se ajusta muito bem como uma mistura de distribuições para os dados de sobrevivência. Assim, para as variáveis aleatórias  $Z_i$ ,  $i = 1, \dots, n$ , seguindo uma distribuição gama, ou seja,  $Z_i \sim \Gamma(\nu, \epsilon)$  variáveis aleatórias independentes com  $\nu, \epsilon \geq 0$ , temos que a função densidade de probabilidade, com

$z_i \geq 0$  é dada por:

$$f(z_i) = \frac{1}{\Gamma(\iota)} z_i^{\iota-1} \exp\{-\epsilon z_i\}, \quad (24)$$

tomando  $\iota = \epsilon = \zeta^{-1}$  e substituindo na equação (24), obtém-se a seguinte equação:

$$f(z_i) = \frac{1}{\Gamma(\frac{1}{\zeta})} z_i^{\frac{1}{\zeta}-1} \exp\left\{-\frac{z_i}{\zeta}\right\}, \quad (25)$$

como  $E(Z_i) = \frac{\iota}{\epsilon} = 1$  e  $\text{Var}(Z_i) = \frac{\iota}{\epsilon^2} = \zeta$ , se  $\zeta = 0$ , então todas as variáveis de fragilidade serão iguais a 1, e assim tem-se o modelo de Cox da forma usual.

Segundo Wienke (2010) quase todos os argumentos em favor da distribuição gama são baseados em aspectos matemáticos e computacionais, não havendo razões biológicas que façam com que essa distribuição seja mais preferível que a outras distribuições.

Colosimo & Giolo (2006) fazem uma revisão dos métodos que são usados para estimar os parâmetros do modelo de riscos proporcionais com fragilidade compartilhada utilizando a distribuição gama para a variável de fragilidade. Os métodos abordados são: O algoritmo EM, estimação via verossimilhança penalizada e estimação bayesiana MCMC (Markov Chain Monte Carlo).

O algoritmo EM é usado para maximizar a função de verossimilhança sobre  $\beta$  e  $\Lambda_0$  com variância da variável de fragilidade  $\zeta$  fixa. Em seguida, integra-se a função de verossimilhança completa (baseada no produto da verossimilhança condicional e na função densidade de probabilidade da variável de fragilidade) para obter um perfil de verossimilhança em  $\zeta$ , e partir daí, otimizá-lo para a obtenção conjunta dos estimadores de máxima verossimilhança de  $(\beta, \Lambda_0, \zeta)$ . A suposição da distribuição gama é usada no passo E para fornecer um mecanismo que auxilia na obtenção dos estimadores e no passo M atualizam-se as estimativas até a convergência do algoritmo. Este algoritmo de acordo com Colosimo & Giolo (2006) e Hanagal (2011) é um algoritmo lento e podem ocorrer problemas de convergência. Sendo que mais detalhes deste algoritmo pode ser visto em Nielsen et al. (1992).

Uma abordagem alternativa para este método é o uso da função de verossimilhança penalizada no processo de estimação. Neste caso, o termo de fragi-

lidade é tratado como um coeficiente de regressão adicional que é restrito por uma função de penalidade acrescentada para o logaritmo da verossimilhança. Segundo Hanagal (2011) este método converge rapidamente e produz os mesmos resultados do algoritmo EM, no caso do modelo de fragilidade gama. A função de penalidade é introduzida para evitar diferenças grandes entre as fragilidades para diferentes grupos (Colosimo & Giolo, 2006). Utilizando a equação (23) e a função de verossimilhança parcial proposta por Cox, dada pela equação (3), temos que a função de verossimilhança parcial penalizada (PPL) é expressa por:

$$\text{PPL}(\boldsymbol{\beta}, \boldsymbol{\omega}, \zeta) = \log \{L(\boldsymbol{\beta}, \boldsymbol{\omega}_i)\} - h(\omega_i, \zeta),$$

em que  $h(\omega_i, \zeta)$  é a função de penalidade. Note que:

$$L(\boldsymbol{\beta}, \boldsymbol{\omega}) = \prod_{i=1}^n \prod_{m=1}^{k_i} \left( \frac{\exp \{\mathbf{x}'_{mi} \boldsymbol{\beta} + \omega_i\}}{\sum_{j \in R(t_m)} \exp \{\mathbf{x}'_{ji} \boldsymbol{\beta} + \omega_{ji}\}} \right)^{\delta_{mi}}.$$

Aplicando o logaritmo nesta expressão, obtém-se:

$$\log \{L(\boldsymbol{\beta}, \boldsymbol{\omega})\} = \sum_{i=1}^n \sum_{m=1}^{k_i} \delta_{mi} \left[ \mathbf{x}'_{mi} \boldsymbol{\beta} + \omega_i - \log \left\{ \sum_{j \in R(t_m)} \exp \{\mathbf{x}'_{ji} \boldsymbol{\beta} + \omega_{ji}\} \right\} \right].$$

Utilizando a função densidade de probabilidade da variável aleatória de fragilidade  $Z_i$  dada na equação (25), com média 1 e variância conhecida  $\zeta = \frac{1}{\theta}$ , e considerando a transformação  $\omega_i = \log \{z_i\}$  então a função densidade de probabilidade da variável aleatória  $\omega_i$  é dada por:

$$f(\omega_i) = \frac{1}{\Gamma(\theta)} \exp \{\omega_i\}^{\theta-1} \exp \{-\theta \exp \{\omega_i\}\} \exp \{\omega_i\}, \quad (26)$$

aplicando o logaritmo na equação (26), tem-se a seguinte expressão:

$$\log \{f(\omega_i)\} = \omega_i - \Gamma(\theta) + \theta \omega_i - \omega_i - \theta \exp \{\omega_i\},$$

observando que o logaritmo da densidade de  $\omega_i$  é  $\frac{\omega_i - \exp \{\omega_i\}}{\zeta}$  mais uma função de  $\zeta$ , obtém-se a seguinte expressão da função de verossimilhança parcial penalizada:

$$\text{PPL}(\boldsymbol{\beta}, \boldsymbol{\omega}, \zeta) = \log \{L(\boldsymbol{\beta}, \boldsymbol{\omega}_i)\} - \frac{1}{\zeta} \sum_{i=1}^n (\omega_i - \exp \{\omega_i\}). \quad (27)$$

Therneau & Grambsch (2000) provam que a solução do modelo com a função de penalidade  $\frac{1}{\zeta} \sum_{i=1}^n (\omega_i - \exp \{\omega_i\})$ , coincide com a solução do EM com fragilidade gama para qualquer valor de  $\zeta$  fixado. Esta função é utilizada no pacote `Survival` (Therneau, 2012) do R para a estimação.

Outro método que é utilizado para estimação dos parâmetros do modelo de riscos proporcionais com fragilidade gama é o MCMC. Este método tem como propósito simular valores de fragilidade à medida que os passos vão avançando em cada iteração, baseando-se na distribuição da fragilidade. O algoritmo relaciona um passo com simulações de fragilidade fundamentadas nos parâmetros atuais e na distribuição condicional de fragilidade e um passo que os parâmetros são atualizados baseados nos valores de fragilidade (Colosimo & Giolo, 2006).

A distribuição Lognormal é outra distribuição utilizada para a variável de fragilidade  $z_i$ , que está inserida no pacote `Survival` (Therneau, 2012) do programa R. O uso desta variável é importante por descrever situações clínicas. Para a variável aleatória  $Z_i$ , com  $i = 1, \dots, n$ , seguindo uma distribuição Lognormal, a função de densidade de probabilidade é dada por:

$$f(z_i) = \frac{1}{\sqrt{2\pi}z_i\sigma} \exp \left\{ -\frac{1}{2} \frac{(\log \{z_i\} - \mu)^2}{\sigma^2} \right\}, \quad (28)$$

em que  $\mu$  é um parâmetro de locação e  $\sigma^2$  é um parâmetro de escala. Observe que  $E(Z_i) = \exp \left\{ \mu + \frac{\sigma^2}{2} \right\}$  e  $\text{Var}(Z_i) = \exp \{2\mu + \sigma^2\} (\exp \{\sigma^2\} - 1)$ . Assim, quando  $E(Z_i) = 1$  e  $\text{Var}(Z_i) = \zeta$ , utiliza-se a seguinte parametrização:  $\mu = -\frac{\sigma^2}{2}$  e  $\sigma^2 = \log(\zeta + 1)$ .

O método de estimação neste caso é realizado utilizando a função de verossimilhança penalizada, cuja a função de penalidade é dada por:

$$h(\boldsymbol{\omega}, \zeta) = \frac{1}{2\zeta} \sum_{i=1}^n \omega_i^2,$$

note que, neste caso, para  $z_i = \exp \{\omega_i\}$ , o efeito aleatório  $\omega_i$  tem distribuição gaussiana. O pacote `Survival` (Therneau, 2012) do programa R emprega a distribuição gaussiana para ajustar esse modelo utilizando a função `coxph`. A estimação da

variância neste caso é feita pelo estimador de máxima verossimilhança restrita aproximada (RMLE), detalhes desse método estão descritos em Therneau & Grambsch (2000).

Colosimo & Giolo (2006) apresentam os testes de hipóteses para testar a associação entre as observações do mesmo grupo, ou seja, testar a hipótese nula  $H_0 : \zeta = 0$ . As estatísticas utilizadas são a estatística de Wald e da razão de verossimilhança.

A estatística de Wald pode ser escrita como:

$$W_\zeta = (\hat{\zeta} - \zeta_0)' \mathcal{I}(\hat{\zeta})(\hat{\zeta} - \zeta_0),$$

em que  $\mathcal{I}(\hat{\zeta})$  é a matriz de informação observada. Sob  $H_0$  e para  $\hat{\zeta}$  de dimensão 1, tem-se que

$$W_\zeta = \frac{\hat{\zeta}^2}{\text{Var}(\hat{\zeta})},$$

sob  $H_0 : \zeta = 0$ , segue uma distribuição qui-quadrado com 1 grau de liberdade.

Para o teste de razão de verossimilhanças, tem-se a seguinte estatística:

$$\text{TRV}_\zeta = 2 \log \left\{ \frac{L(\hat{\lambda}_0, \hat{\beta}, \hat{\zeta})}{L(\hat{\lambda}_0^*, \hat{\beta}^*)} \right\} = 2 \left[ \log \{L(\hat{\lambda}_0, \hat{\beta}, \hat{\zeta})\} - \log \{L(\hat{\lambda}_0^*, \hat{\beta}^*)\} \right],$$

em que

$$L(\hat{\lambda}_0, \hat{\beta}, \hat{\zeta}) = \prod_{i=1}^n \left[ z_i^{\frac{1}{\hat{\zeta}} - 1} \left( \frac{1}{\hat{\zeta}} \right)^{\frac{1}{\hat{\zeta}}} \exp \left\{ -z_i \frac{1}{\hat{\zeta}} \right\} \frac{1}{\Gamma\left(\frac{1}{\hat{\zeta}}\right)} \right. \\ \left. \prod_{m=1}^{k_i} \exp \left\{ - \int_0^t z_i \lambda_0(u) \exp \{ \mathbf{x}'_{mi} \boldsymbol{\beta} \} du \right\} (z_i \lambda_0(t) \exp \{ \mathbf{x}'_{mi} \boldsymbol{\beta} \})^{\delta_{mi}} \right],$$

é a função de verossimilhança completa, sendo  $\hat{\lambda}_0$ ,  $\hat{\beta}$  e  $\hat{\zeta}$  as estimativas obtidas utilizando o modelo de fragilidade dado pela equação (22) e a função de verossimilhança,  $L(\hat{\lambda}_0^*, \hat{\beta}^*)$ , considerando o modelo da equação (22) com todos os  $z_i$ ,  $i = 1, \dots, n$ , iguais a 1.

Se o interesse for testar o efeito das covariáveis, é possível testar a hipótese do tipo  $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ . As estatísticas de Wald e da razão da verossimilhança novamente podem ser utilizadas, e são dadas da seguinte forma:

$$W_{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$



e

$$\text{TRV}_{\boldsymbol{\beta}} = 2 \log \left\{ \frac{L(\hat{\lambda}, \hat{\boldsymbol{\beta}}, \hat{\zeta})}{L(\hat{\lambda}, \boldsymbol{\beta}_0, \hat{\zeta})} \right\} = 2 \left[ \log\{L(\hat{\lambda}_0, \hat{\boldsymbol{\beta}}, \hat{\zeta})\} - \log\{L(\hat{\lambda}, \boldsymbol{\beta}_0, \hat{\zeta})\} \right], \text{ respectivamente.}$$

Assintoticamente,  $W_{\boldsymbol{\beta}}$  e  $\text{TRV}_{\boldsymbol{\beta}}$  têm distribuição  $\chi_p^2$ , em que  $p$  é a diferença do número de parâmetros que estão sendo comparados.

## 3 MATERIAL E MÉTODOS

### 3.1 Dados

O conjunto de dados em que foram aplicados os modelos marginais e condicionais trata-se de um estudo de coorte prospectivo que incluiu pacientes com doença renal crônica em tratamento dialítico por no mínimo três meses, maiores de 18 anos e que foi observado a ocorrência de eventos cardiovasculares. Os dados fazem parte da tese de Doutorado de Antunes (2012), sendo considerados neste trabalho 130 pacientes com doença renal crônica sendo 104 pacientes tratados por hemodiálise (HD) e 26 pacientes por diálise peritoneal (DP), no Hospital das Clínicas da Faculdade de Medicina da UNESP - Campus Botucatu/SP. Os pacientes foram acompanhados à partir de 2008 por um tempo de seguimento médio de 38 meses.

A diálise é uma terapia que visa substituir a função renal de indivíduos que não conseguem eliminar água e outros produtos que normalmente são filtrados pelo rim e eliminados na urina, podendo ser realizada tanto na forma de hemodiálise quanto na forma de diálise peritoneal. “A desnutrição e sobrecarga líquida têm sido apontadas como fatores que contribuem para o pior prognóstico cardíaco da população em diálise ” (Thomasset (1962) e Dumler & Kilates (2000), apud Antunes (2012), p.34) e sendo a doença cardiovascular a causa principal de óbito desses pacientes (Antunes, 2012).

A bioimpedância (BIA) em diálise, é uma ferramenta para avaliar a composição corporal, e monitorar o estado nutricional. A BIA é um método não invasivo e relativamente barato para avaliação da composição corporal (Antunes, 2012). A análise por BIA consiste em aplicar uma corrente elétrica de pequena am-

peragem no corpo do indivíduo e a partir daí observar um comportamento do fluxo dessa corrente, percebido como um valor de impedância. Esta impedância por sua vez, exprime dois componentes: a resistência (tecido muscular e água corporal oferecem baixa resistência e, por outro lado, tecido adiposo e tecido ósseo oferecem alta resistência) e a reactância (representa o fluxo da corrente através da célula). A partir dessas duas medidas podem-se estimar os valores de água corporal, massa magra e adiposa, ângulo fase, massa celular corporal, massa intracelular e extracelular, para mais detalhes ver Antunes (2012).

As covariáveis consideradas em seu trabalho foram variáveis clínicas, nutricionais, laboratoriais e dialíticas: sexo, método de diálise (diálise peritoneal e hemodiálise), tempo em diálise, idade, diagnóstico de diabetes mellitus (DM), triglicerídeos (TG), pressão arterial sistólica (PAS) e diastólica (PAD), colesterol total (CT) e HDL colesterol, risco relativo de ocorrência de doença cardíaca por Framingham, paratormônio, cálcio, fósforo, creatinina, depuração fracional da uréia (Kt/V), proteína C-reativa (PCR), albumina e índice de massa corpórea (IMC). Na análise de bioimpedância (Biodynamics®), modelo 450, 800  $\mu A$ , 50 KHz) a variável analisada foi o quociente de massa extracelular corporal por massa celular corporal (MEC/MCC). Antunes (2012) expõe que o quociente MEC/MCC passou a ser utilizado recentemente como marcador de prognóstico em diálise, sendo interpretado como: quanto maior o valor do quociente MEC/MCC pior o estado nutricional e/ou de hidratação, repercutindo em pior prognóstico do paciente. Apesar do quociente MEC/MCC ser utilizado como marcador prognóstico, não havia estudo que indique sua associação com a ocorrência de eventos cardiovasculares em diálise tendo sido objetivo do trabalho de Antunes (2012) investigar essa relação. Sendo objetivo de seu trabalho investigar a relação entre os parâmetros de bioimpedância e prognóstico cardiovascular em diálise. Para o estudo dessa relação ela utilizou o modelo de Cox usual com covariáveis dependentes do tempo.

Os tipos de eventos cardiovasculares considerados no trabalho de Antunes (2012) foram:

- Ataques isquêmicos transitórios (Acidente Vascular Cerebral - AVC);
- Infarto do miocárdio (ocorrência da morte tecidual do músculo cardíaco);
- Arritmia (alteração do ritmo normal do coração);
- Eventos trombóticos (ocorrência de coagulação do sangue no interior de vaso sanguíneo ou no coração);
- Angina instável (dores frequentes no lado esquerdo do peito que podem evoluir para um infarto do miocárdio);
- Morte súbita (ocorre repentinamente, sem previsão, trauma ou violência, em adultos ou crianças);
- Emergência Hipertensiva (quando a pressão arterial é superior a  $180 \times 120$  mmHg com lesões no órgão alvo – por exemplo: coração, cérebro e pulmão).

Nesta dissertação tratou-se esses eventos cardiovasculares como eventos do mesmo tipo, uma vez que, todos esses eventos contribuem para pior prognóstico do paciente, podendo levá-lo a morte ou deixá-lo vulnerável a outros eventos cardiovasculares.

### **3.2 Organização do banco de dados e as funções utilizadas no programa R**

Os dados foram organizados em uma planilha eletrônica, sendo configurados de acordo com os pressupostos dos modelos marginais e condicionais. Para exemplificar a estruturação do banco de dados, considere três indivíduos, o 5º, 6º e 74º, dos 130 pacientes do banco de dados e as covariáveis idade e método de diálise (0: diálise peritoneal e 1: hemodiálise). Para os modelos AG, PWP e os modelos condicionais o banco de dados é o mesmo, seguindo as características de uma estrutura de processo de contagem, sendo a estruturação exemplificada na Tabela 1.

Tabela 1. Formato do banco de dados para análise de eventos múltiplos dos modelos AG, PWP e modelos condicionais.

id	início	fim	estado	estrato	idade	método
5	0	8	1	1	73	0
5	8	26	0	2	73	0
6	0	9	1	1	56	0
6	9	24	1	2	56	0
6	24	26	0	3	56	0
74	0	12	1	1	49	1
74	12	19	1	2	49	1
74	19	26	1	3	49	1
74	26	34	0	4	49	1

id: identificação do indivíduo; início: tempo de entrada no estudo; fim: tempo de acompanhamento ou ocorrência do evento; estado: ocorrência de eventos (1: sim e 0: não); estrato: ordenação dos tempos de ocorrência dos eventos.

Observe que o 5º indivíduo do banco de dados apresentou um evento no tempo de 8 meses e foi acompanhado por 26 meses, o 6º indivíduo sofreu o mesmo evento duas vezes e também foi acompanhado até o tempo de 26 meses, enquanto o 74º sofreu três eventos do mesmo tipo sendo acompanhado por 34 meses. O modelo AG e os modelos condicionais utilizam a estrutura de processo de contagem mas não necessitam da variável estrato. Por outro lado, o modelo PWP necessita dessa variável para separar a análise do risco basal em diferentes estratos.

O modelo WLW tem uma estrutura de entrada dos dados diferente. Suponha que num estudo tem-se  $n$  indivíduos, e estes ao entrarem no estudo estão sujeitos a sofrerem  $m$  eventos, que é o número máximo de eventos no estudo. O banco de dados apresentará então  $nm$  linhas e mesmo que alguns indivíduos não sofram os  $m$  eventos, observações fictícias são criadas para esses indivíduos. Para exemplificar,

neste trabalho, os dados apresentam 130 indivíduos e o número máximo de três eventos, então, o banco de dados terá 390 linhas. Esse modelo não considera a estrutura de processo de contagem (o tempo é sempre contado a partir do zero) e a Tabela 2 exemplifica a configuração do banco de dados para o modelo WLW, considerando novamente os mesmos indivíduos e as covariáveis consideradas na Tabela 1. Observe

Tabela 2. Formato do banco de dados para análise de eventos múltiplos do modelo WLW.

id	fim	estado	estrato	idade	método
5	8	1	1	73	0
5	26	0	2	73	0
5	26	0	2	73	0
6	9	1	1	56	0
6	24	1	2	56	0
6	26	0	3	56	0
74	12	1	1	49	1
74	19	1	2	49	1
74	26	1	3	49	1

id: identificação do indivíduo; fim: tempo de acompanhamento ou ocorrência do evento; estado: ocorrência de eventos (1: sim e 0: não); estrato: ordenação dos tempos de ocorrência dos eventos.

que o 5º indivíduo apresentou apenas um evento no estudo, mas foi criada mais duas observações deste indivíduo para completar as suas três linhas no banco de dados, independentemente dele ter sofrido apenas um evento. O mesmo raciocínio vale para o 6º e 74º, com a ressalva que um sofreu dois e outro três eventos, respectivamente. A variável estrato neste modelo serve para separar o risco para cada evento.

Em seguida, os dados foram importados no programa R, utilizando o comando `read.csv2("nome dos dados.csv")`. Como algumas covariáveis no conjunto de dados apresentaram valores omissos, resolveu-se antes de estimar os parâmetros,

utilizar a função `na.omit(nome dos dados)` para retirar esses valores, pois as funções `stepAIC` (utilizada para selecionar as covariáveis) e a função `residuals` com o argumento `collapsed` (utilizada para calcular resíduos por indivíduo) não são executadas com esses valores omissos. Vale lembrar que, o programa R elimina automaticamente essas linhas do banco de dados quando os valores omissos são indicados por *NA*, mas para calcular os resíduos e selecionar as covariáveis utilizando os comandos mencionados anteriormente é necessário realizar essa tarefa de eliminar os valores omissos antes da estimação do modelo.

Para ajustar os modelos marginais e condicionais foi utilizada a função `coxph` do pacote `Survival` (Therneau, 2012), sendo que nos modelos marginais acrescenta-se o argumento `cluster` para estimar a variância robusta dos parâmetros estimados e nos modelos condicionais o argumento `frailty` para estimar a variância da variável de fragilidade.

Como forma de selecionar as covariáveis nos modelos marginais e condicionais utilizou-se a função `stepAIC` e foram utilizados os métodos *forward*, *backward* ou *stepwise* para seleção de covariáveis.

Em seguida, calculou-se os valores de AIC utilizando a função `extractAIC(modelo)` para os modelos marginais e condicionais. Para calcular os valores BIC e  $AIC_c$  para esses modelos basta utilizar as funções `extractAIC(modelo, k = log(número de observações))` e `extractAIC(modelo, k = 2 × número de observações / (número de observações - (número parâmetros) - 1))`, respectivamente, sendo  $k$  o termo de correção. Após comparar os modelos por esses critérios, optou-se pelo modelo com menores valores desses critérios, como forma de selecionar um modelo parcimonioso para o conjunto de dados de eventos cardiovasculares. Outras questões discutidas para selecionar o melhor modelo para o conjunto de dados, sugeridas por Lim et al. (2007), são a frequência de eventos e se a ocorrência de novos eventos são influenciados por eventos anteriores. Assim, avaliou-se os modelos marginais através do gráfico da função de risco basal acumulado (segundo a ordem da ocorrência de eventos) e de uma tabela de contingência dada pela ordem

de ocorrência dos eventos e da função indicadora de falha.

Para avaliar a qualidade do ajuste dos modelos marginais analisou-se os resíduos de Shoenfeld através da função `cox.zph` e também foram analisados os resíduos de Martingale utilizando a função `residuals` com os argumentos `type=martingale` e `collapsed`. Para mais detalhes do código para o ajuste desses modelos ver o Apêndice.

Por fim, como forma de interpretar os parâmetros estimados associados às covariáveis do modelo selecionado, utilizou-se as razões de risco em relação aos parâmetros estimados  $\hat{\beta}$  associados às covariáveis do modelo selecionado, que pode ser obtido pela função `summary(modelo)`, como forma de verificar se há associação das covariáveis com a ocorrência de eventos cardiovasculares.



## 4 RESULTADOS E DISCUSSÃO

Inicialmente, foi realizada a seleção das covariáveis dos modelos marginais utilizando os métodos: *backward*, *forward* e *stepwise*. Para os modelos marginais após feita seleção pela função `stepAIC` selecionou-se as covariáveis ao nível de significância de 5% sendo esta significância analisada pelo teste de Wald robusto.

Para os modelos marginais – AG, PWP e WLW – os métodos de seleção mencionados anteriormente, selecionaram as mesmas covariáveis e esses resultados estão apresentados nas Tabelas 3, 4 e 5. Observando a Tabela 3, notou-se que o erro padrão robusto difere muito pouco do erro padrão usual do modelo de Cox, sugerindo que o modelo AG seja um modelo marginal apropriado para esses dados, pois indicou não haver uma estrutura de dependência entre os eventos de um mesmo indivíduo. Assim, se isso for verdade, a história de eventos cardiovasculares anteriores não altera o risco presente, e com isso, a função de risco basal seria a mesma para todos os indivíduos. Como forma de selecionar melhor o modelo marginal para esses dados, o próximo passo foi analisar a frequência de eventos recorrentes por sujeito.

Na Tabela 6 exibiu-se a distribuição conjunta das variáveis estado e estrato, o que mostrou não haver uma frequência muito grande dos eventos quando separados pela ordem ocorrência dos mesmos. Como houve apenas um número de duas observações para os indivíduos que apresentaram três eventos cardiovasculares, resolveu-se agrupá-los com os indivíduos que apresentaram dois eventos. Analisando esta tabela observa-se que o risco da ocorrência de um evento pode variar substancialmente, e assim, é sugerido por Lim et al. (2007) que o modelo PWP passa a ser o modelo mais apropriado para tratar dos eventos cardiovasculares recorrentes. O gráfico da Figura 1 reforça o que foi argumentado anteriormente, tendo em vista

Tabela 3. Resultado da seleção das covariáveis para o modelo AG utilizando os métodos: *backward*, *forward* e *stepwise*.

Covariável	Estimativas	Erro Padrão	Erro padrão (Robusto)	Wald	Valor p
PAS	0,01694	0,00839	0,00835	2,02800	0,04253
Cálcio	0,46178	0,21808	0,20290	2,27600	0,02285
Fósforo	0,41322	0,14369	0,16392	2,52100	0,01171
MEC/MCC	3,25116	0,89540	0,84873	3,83100	0,00013

Tabela 4. Resultado da seleção das covariáveis para o modelo PWP utilizando os métodos: *backward*, *forward* e *stepwise*.

Covariável	Estimativas	Erro Padrão	Erro padrão (Robusto)	Wald	Valor p
Cálcio	0,38460	0,21620	0,19320	1,99000	0,04656
Fósforo	0,40840	0,14550	0,16400	2,49100	0,01275
MEC/MCC	2,54030	0,90170	0,81200	3,12800	0,00176

Tabela 5. Resultado da seleção das covariáveis para o modelo WLW utilizando os métodos: *backward*, *forward* e *stepwise*.

Covariável	Estimativas	Erro Padrão	Erro padrão (Robusto)	Wald	Valor p
Fósforo	0,42650	0,12290	0,19330	2,20700	0,02730
MEC/MCC	2,49640	0,69770	0,97080	2,57200	0,01010

Tabela 6. Distribuição conjunta das variáveis estado e estrato.

Estrato	Estado		Total
	Falhas	Censuras	
1	37 (28,46%)	93 (71,54%)	130 (100%)
2	14 (46,47%)	16 (53,33%)	30 (100%)
Total	51 (31,88%)	109 (68,13%)	160 (100%)

que o risco aumenta conforme a ocorrência de um novo evento cardiovascular. Isto sugere que as covariáveis sejam modeladas condicionadas sob a história de eventos cardiovasculares dos indivíduos.

Para esses modelos marginais, obteve-se também o valores de AIC,  $AIC_c$  e BIC. Os valores obtidos por esses critérios são apresentados na Tabela 7 e apontam que o modelo PWP foi o melhor modelo marginal para ajustar o conjunto de dados de eventos cardiovasculares recorrentes, pois apresentou os menores valores desses critérios analisados. A partir destes resultados, verificou-se a qualidade

Tabela 7. Valores de AIC,  $AIC_c$  e BIC para os modelos marginais.

Critério	Modelos Marginais		
	AG	PWP	WLW
AIC	438,6385	383,5298	722,0768
$AIC_c$	438,8966	383,6836	722,1533
BIC	450,9392	392,7553	728,2272

do ajuste do modelo PWP utilizando os resíduos de Schoenfeld e os resíduos de Martingale. Na Tabela 8, observa-se que as estimativas do coeficiente de correlação de Pearson ( $\hat{\rho}$ ) são todos próximos de zero. Além disso, observou-se que tanto o teste global quanto os testes para cada covariável não apresentaram evidências para

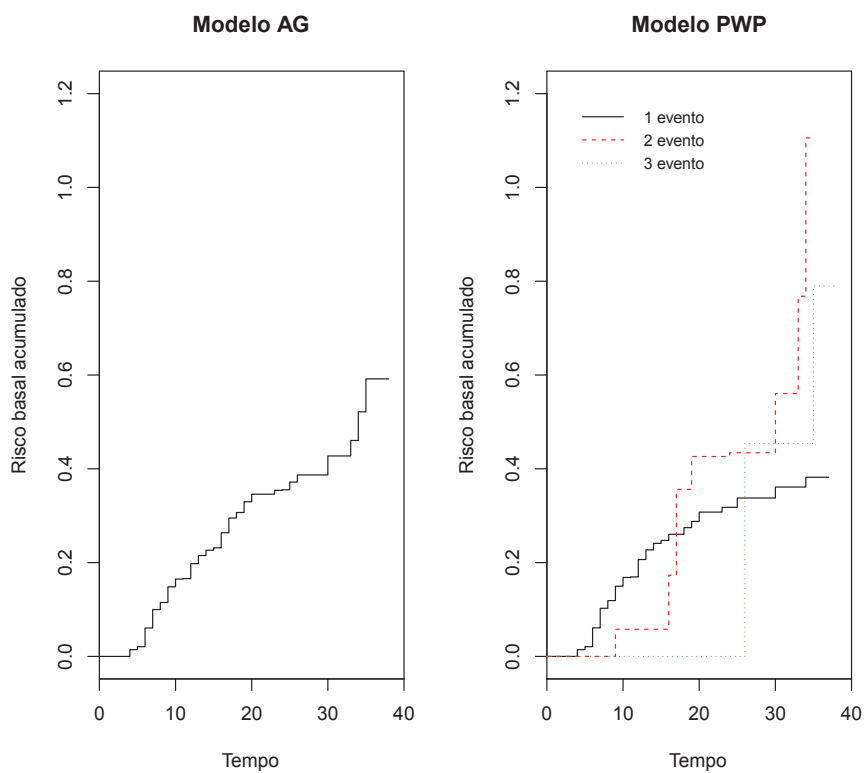


Figura 1 - Risco basal acumulado para ocorrência de novo evento cardiovascular: à esquerda, o modelo de eventos ordenados independentes (AG); à direita, o modelo PWP especificado segundo a ordem de ocorrência dos eventos.

rejeitar a hipótese nula, ou seja, a hipótese de riscos proporcionais.

Tabela 8. Teste de proporcionalidade dos riscos no modelo PWP ajustado.

Covariável	$\hat{\rho}^a$	$T^b$	valor p
Cálcio	0,11600	0,57500	0,44800
Fósforo	-0,14800	2,18700	0,13900
MEC/MCC	-0,13800	1,04300	0,30700
GLOBAL	–	2,62700	0,45300

<sup>a</sup> Coeficiente de correlação de Pearson estimado entre os resíduos padronizados de Schoenfeld e a variável resposta tempo até a ocorrência do evento;

<sup>b</sup> Estatística do teste com distribuição aproximadamente qui-quadrado com 3 graus de liberdade.

Analisando os gráficos dos resíduos padronizados de Schoenfeld versus os tempos para as covariáveis consideradas no modelo PWP (Figura 2), confirmou-se o que foi exposto no teste de proporcionalidade, tendo em vista que, não foi possível visualizar tendências evidentes ao longo do tempo. Pela análise do gráfico de resíduos de Martingale para o modelo PWP versus indivíduos (Figura 3), notou-se que os resíduos estão relativamente bem distribuídos acima e abaixo da reta que passa pelo zero e não tem nenhum padrão detectável. Desse modo, o modelo PWP é uma opção satisfatória para análise desses dados.

Em seguida, os modelos condicionais com distribuições gama e lognormal para a variável de fragilidade foram investigados. Um primeiro passo foi realizar a seleção de covariáveis utilizando novamente os métodos de seleção: *backward*, *forward* e *stepwise*. Considerando também para esses modelos as covariáveis significativas ao nível de 5% e esta significância avaliada pelo teste de Wald. Notou-se que as covariáveis selecionadas pelos métodos de seleção mencionadas foram as mesmas para ambos os modelos e os resultados são apresentados nas Tabelas 9 e 10. Observou-se, pelas Tabelas 9 e 10, que em ambos os modelos a correlação intra-indivíduo não é significativa. Isto significa que a variável de fragilidade não altera significativamente

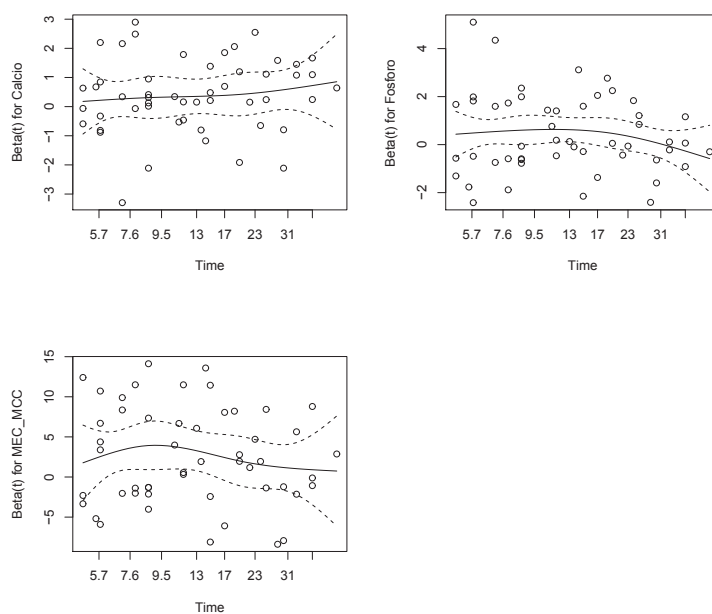


Figura 2 - Resíduos padronizados de Schoenfeld versus os tempos para as covariáveis consideradas no modelo PWP.

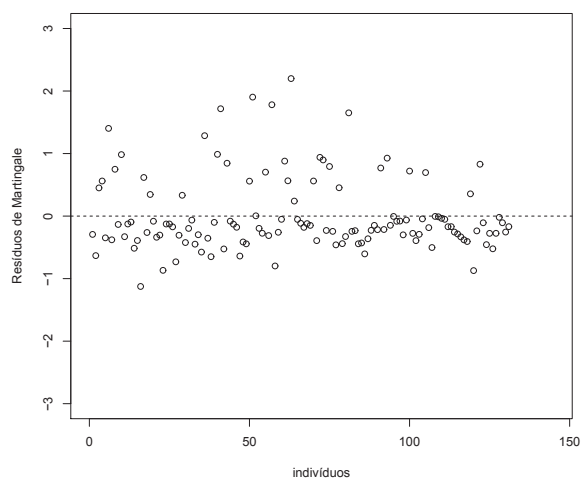


Figura 3 - Gráfico dos resíduos de Martingale para o modelo PWP versus indivíduos.

os efeitos e as interpretações das covariáveis do modelo de Cox usual. As variâncias estimadas das variáveis de fragilidade para os modelos com distribuição lognormal e gama foram  $\hat{\zeta} = 0,419$  e  $\hat{\zeta} = 0,485$ , respectivamente.

Tabela 9. Resultado da seleção das covariáveis obtido para o modelo semi-paramétrico de fragilidade lognormal utilizando os métodos: *backward*, *forward* e *stepwise*.

Covariável	Estimativas	Erro Padrão	Wald	Valor p
Fósforo	0,41900	0,15100	7,68000	0,00560
MEC/MCC	2,94200	0,93500	9,90000	0,00160
fragilidade lognormal (indivíduos)	–	–	20,50000	0,23000

Tabela 10. Resultado da seleção das covariáveis obtido para o modelo semi-paramétrico de fragilidade gama utilizando os métodos: *backward*, *forward* e *stepwise*.

Covariável	Estimativas	Erro Padrão	Wald	Valor p
Fósforo	0,419	0,152	7,58	0,0059
MEC/MCC	2,990	0,949	9,93	0,00160
fragilidade gama (indivíduos)	–	–	22,64	0,2700

Para os modelos condicionais também foram obtidos os valores do AIC e BIC, não sendo calculado os valores do  $AIC_c$  pois para os modelos lognormal e gama as razões  $\frac{n}{K} = \frac{160}{18,27411} = 8,75 < 40$  e  $\frac{n}{K} = \frac{160}{20,98008} = 7,63 < 40$ , respectivamente. Esses valores são apresentados na Tabela 11 e mostram que dos modelos condicionais, o modelo lognormal foi o melhor modelo para ajustar o conjunto de dados desse trabalho.

Tabela 11. Valores de AIC e BIC para os modelos condicionais.

Critério	Modelos Condicionais	
	Lognormal	Gama
AIC	438,65692	443,39750
BIC	494,85300	507,91490

Comparando os valores dos critério de AIC e BIC do modelo marginal PWP e do modelo condicional lognormal apresentados nas Tabelas 7 e 11, respectivamente, observou-se ainda que o modelo marginal PWP continuou sendo o melhor modelo para ajustar o conjunto de dados de eventos cardiovasculares.

O próximo passo, teve como objetivo fornecer as interpretações para os parâmetros associados às covariáveis estimadas pelo modelo PWP. Assim, inicialmente foram fornecidas as razões de risco dos indivíduos em relação as covariáveis, sendo as estimativas apresentadas na Tabela 12. A partir desses resultados observou-

Tabela 12. Estimativas das razões de risco associados às covariáveis do modelo PWP.

Covariáveis	Razão de risco	Razão de risco I.C (95%)
Cálcio	1,469	(1,006 ; 2,145)
Fósforo	1,504	(1,091 ; 2,075)
MEC/MCC	12,684	(2,582 ; 62,294)

se que os indivíduos que apresentam valores elevados das variáveis – cálcio, fósforo e MEC/MCC – tiveram um risco maior de sofrer eventos cardiovasculares comparados com os indivíduos que apresentam menores valores. Assim, por exemplo, em relação a covariável MEC/MCC, dado a história de eventos anteriores, no aumento de unidade na escala exponencial tem-se um risco de evento cardiovascular aproximadamente 13 vezes maior, para os indivíduos que apresentaram maiores valores



desta covariável. Além disso, pode-se dizer com 95% de confiança que esse risco é superior a 2,582 e inferior a 62,294.

## 5 CONCLUSÕES

A partir dos resultados apresentados conclui-se que o modelo PWP com as covariáveis – cálcio, fósforo e MEC/MCC – foi o melhor modelo para tratar o conjunto de dados analisados, baseando-se nos critérios de seleção do modelo, análise de diagnóstico da qualidade de ajuste do modelo e as características do evento estudado. Outra questão importante neste trabalho é a associação encontrada entre a covariável MEC/MCC e a ocorrência de eventos cardiovasculares nos pacientes em diálise, sendo o acompanhamento desse parâmetro útil na predição de eventos cardiovasculares.

## REFERÊNCIAS BIBLIOGRÁFICAS

- AKAIKE, H. A. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v.19, n.6, p.716–723, 1974.
- ANDERSEN, P. K.; GILL, R. D. Cox's regression model for counting processes: a large sample study. **Annals of Statistical**, v.10, n.1, p.1100–1120, 1982.
- ANTUNES, A. A. Impacto do acompanhamento com bioimpedância na predição de eventos cardiovasculares em diálise. Botucatu, 2012. 53p. Tese (Doutorado) - Faculdade de Medicina da Universidade Estadual Paulista "Júlio de Mesquita Filho".
- BRESLOW, N. Contribution to the discussion of the paper by D.R Cox. **Journal of the Royal Statistical Society B**, v.34, n.1, p.216–217, 1972.
- BURNHAM, K. P.; ANDERSON, D. R. **Model selection and multimodel inference a practical information-theoretic approach**. New York: Springer, 2002. 347p.
- 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**. Rio de Janeiro: Fiocruz, 2011. 432p.
- COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Edgard Blücher, 2006. 392p.
- COX, D. R. Regression models and life-tables. **Journal of the Royal statistical Society-B**, v.34, n.1, p.187–220, 1972.
- COX, D. R. Partial likelihood. **Biometrika**, v.62, n.2, p.269–276, 1975.

- DUMLER, F.; KILATES, C. Use of bioelectrical impedance techniques for monitoring nutritional status in patients on maintenance dialysis. **J Ren Nutr**, v.10, n.1, p.116–124, 2000.
- GRAMBSCH, P. M.; THERNEAU, T. M. Partial residuals for the proportional hazards regression model. **Biometrika**, v.81, n.3, p.515–526, 1994.
- HANAGAL, D. D. **Modeling survival data using frailty models**. New York: Chapman & Hall/CRC, 2011. 312p.
- KELLY, P. J.; LIM, L. L.-Y. Survival analysis for recurrent event data: an application to childhood infectious diseases. **Statistics in Medicine**, v.19, n.1, p.13–33, 2000.
- KLEINBAUM, D. G.; KLEIN, M. **Survival analysis: A self-learning text**. New York: Springer, 2011. 700p.
- LAWLESS, J. F. **Statistical models and methods for lifetime data**. New Jersey: Wiley, 2002. 621p.
- LIM, H. J.; LIU, J.; MELZER-LANGE, M. Comparison of methods for analyzing recurrent events data: application to the emergency department visits of pediatric firearm victims. **Accident Analysis and Prevention**, v.39, n.2, p.290–299, 2007.
- LIN, D. Y. Cox regression analysis of multivariate failure time data: the marginal approach. **Statistics in Medicine**, v.13, n.21, p.2233–2247, 1994.
- MCLAIN, A.; PEÑAY, E. Some issues in marginal recurrent event Cox type models. **Institute of Statistical Mimeo Series**, v.2618, n.1, p.1–30, 2008.
- METCALFE, C.; THOMPSON, S. G. Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? **Statistical Methods in Medical Research**, v.16, n.2, p.103–122, 2007.

NIELSEN, G. G.; GILL, R. D.; ANDERSEN, P. K.; SØRENSEN, T. I. A. A counting process approach to maximum likelihood estimation in frailty models. **Scandinavian Journal of Statistics**, v.19, n.1, p.25–43, 1992.

PETO, R. Contribution to the discussion of the paper by D.R Cox. **Journal of the Royal Statistical Society B**, v.34, n.1, p.205–207, 1972.

PRENTICE, R. L.; WILLIAMS, B. J.; PETERSON, A. V. On the regression analysis of multivariate failure time data. **Biometrika**, v.68, n.1, p.373–379, 1981.

R DEVELOPMENT CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-08-9.

SCHWARZ, G. Estimating the Dimension of a Model. **Annals of Statistics**, v.6, n.2, p.461–464, 1978.

SHOENFELD, D. Partial residuals for the proportional hazards regression model. **Biometrika**, v.69, n.1, p.239–241, 1982.

THERNEAU, T. **A Package for Survival Analysis in S**, 2012. R package version 2.36-14.

THERNEAU, T. M.; GRAMBSCH, P. M. **Modeling survival data: extending the Cox model**. New York: Springer, 2000. 350p.

THOMASSET, M. A. Bioelectric properties of tissue. Impedance measurement in clinical medicine. Significance of curves obtained. **Lyon Med** 1962, v.94, n.1, p.107–118, 1962.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. 4. ed. New York: Springer, 2002. ISBN 0-387-95457-0.

WEI, L. J.; LIN, D. Y.; WEISSFELD, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. **Journal of the American Statistical Association**, v.84, n.1, p.1065–1073, 1989.

WIENKE, A. **Frailty models in survival analysis**. New York: Chapman & Hall/CRC, 2010. 299p.

## APÊNDICES

Rotinas utilizadas no programa estatístico R

Análise dos modelos marginais – Modelos AG, PWP e WLW

Entrada dos dados:

```
dados<-read.csv2("dados1.csv")
dados2<-read.csv2("dados2.csv")
dat1<-na.omit(dados)
dat2<-na.omit(dados2)
require(survival)
require(MASS)
```

Seleção de variáveis:

Modelos nulos

```
agn<-coxph(Surv(início,fim,estado)~MEC_MCC+cluster(id),
data=dat1,method="breslow")
summary(agn)
pwpn<-coxph(Surv(início,fim,estado)~MEC_MCC+cluster(id)+strata(estrato),
data=dat1,method="breslow")
summary(pwpn)
wlwn<-coxph(Surv(fim,estado)~ MEC_MCC+cluster(id)+strata(estrato),
data=dat2,method="breslow")
summary(wlwn)
```

## Modelos Completos

```
agc<-coxph(Surv(inicio,fim,estado)~sexo+metodo+tempo_dial+idade+
DM+TG+PAS+PAD+CT+HDL+RR_fram_inicial+PTH+Calcio+Fosforo+
Creatinina+Kt_V+PCR+albumina+IMC+MEC_MCC+
cluster(id), data=dat1, method="breslow")
summary(agc)
```

```
pwpc<-coxph(Surv(inicio,fim,estado)~sexo+metodo+tempo_dial+idade+
DM+TG+PAS+PAD+CT+HDL+RR_fram_inicial+PTH+Calcio+Fosforo+
Creatinina+Kt_V+PCR+albumina+IMC+MEC_MCC+
cluster(id)+strata(estrato), data=dados, method="breslow")
summary(pwpc)
```

```
wlwc<-coxph(Surv(fim,estado)~sexo+metodo+tempo_dial+idade_inicial+
DM+TG+PAS+PAD+CT+HDL+RR_fram_inicial+PTH+
Calcio+Fosforo+Creatinina+Kt_V+PCR+albumina+IMC+MEC_MCC
+cluster(id)+strata(estrato),data=dat2,method="breslow")
summary(wlwc)
```

Seleção modelo AG:

```
result1<-stepAIC(agc,method='backward')
summary(result1)
result2<-stepAIC(agn,scope=list(upper=agc,lower=agn),method='forward')
summary(result2)
result3<-stepAIC(agn,scope=list(upper=agc,lower=agn),method='both')
summary(result3)
```

Seleção modelo PWP:

```
result1<-stepAIC(pwpc,method='backward')
```



```
summary(result1)
result2< -stepAIC(pwpn,scope=list(upper=pwpc,lower=pwpn),method='forward')
summary(result2)
result3< -stepAIC(pwpn,scope=list(upper=pwpc,lower=pwpn),method='both')
summary(result3)
```

Seleção modelo WLW:

```
result1< -stepAIC(wlwc,method='backward')
summary(result1)
result2< -stepAIC(wlwn,scope=list(upper=wlwc,lower=wlwn),method='forward')
summary(result2)
result3< -stepAIC(wlwn,scope=list(upper=wlwc,lower =wlwn),method='both')
summary(result3)
```

Modelos selecionados utilizando os métodos de seleção: backward,forward e stepwise considerando nível de significância de 5%:

```
ag< -coxph(Surv(início,fim,estado)~PAS+Calcio+Fosforo+MEC.MCC+
cluster(id),data=dat1,method="breslow")
summary(ag)
aicag< -extractAIC(ag)
bicag< -extractAIC(ag,k= log (160))
aiccag< -extractAIC(ag,k= 2*160/(160-4-1))
pwp< - coxph(Surv(início,fim,estado)~Calcio+Fosforo+MEC.MCC+
cluster(id)+strata(estrato),data=dat1,method="breslow")
summary(pwp)
aicpwp< -extractAIC(pwp)
bicpwp<-extractAIC(pwp,k= log (160))
aiccpwp< -extractAIC(pwp,k=2*160/(160-3-1))
```

```
wlw <- coxph(Surv(fim,estado)~Fosforo+MEC_MCC+cluster(id)+
strata(estrato),data=dat2,method="breslow")
summary(wlw)
aicwlc <- extractAIC(wlw)
bicwlc <- extractAIC(wlw,k=log(160))
aiccwlc <- extractAIC(wlw,k=2*160/(160-2-1))
```

Análise dos modelos condicionais – Modelos semiparamétricos com distribuições gama e lognormal para a variável de fragilidade

```
dados <- read.csv2("dados1.csv")
require(survival)
require(MASS)
dat1 <- na.omit(dados)
Seleção de variáveis
```

Modelos nulos

```
modngm <- coxph(Surv(inicio,fim,estado)~1+frailty(id,dist="gamma"),data=dat1)
summary(modngm)
modngaus <- coxph(Surv(inicio,fim,estado)~1+frailty(id,dist="gauss"),data=dat1)
summary(modngaus)
```

Modelos Completos

```
modcgm <- coxph(Surv(inicio,fim,estado)~sexo+metodo+tempo_dial+
idade+DM+TG+PAS+PAD+CT+HDL+RR_fram_inicial+PTH+Calcio+Fosforo+
Creatinina+Kt_V+PCR+albumi+IMC+MEC_MCC+
frailty(id,dist="gamma"),data=dat1)
```

```
summary(modcgm)
```

```
modcgaus< -coxph(Surv(início,fim,estado)~sexo+metodo+tempo_dial+
idade+DM+TG+PAS+PAD+CT+HDL+RR_fram_inicial+PTH+Calcio+Fosforo+
Creatinina+Kt_V+PCR+albumi+IMC+MEC_MCC+
frailty(id,dist="gauss"),data=dat1)
summary(modcgaus)
```

Seleção modelo Gamma

```
result1< -stepAIC(modcgm,method='backward')
summary(result1)
result2< -stepAIC(modngm,scope=list(upper=modcgm,lower=modngm),
method='forward')
summary(result2)
result3< -stepAIC(modngm,scope=list(upper=modcgm,lower=modngm),
method='both')
summary(result3)
```

Seleção modelo lognormal

```
result1< -stepAIC(modcgaus,method='backward')
summary(result1)
result2< -stepAIC(modngauss,scope=list(upper=modcgaus,lower=modngauss),
method='forward')
summary(result2)
result3< -stepAIC(modngauss,scope=list(upper=modcgaus,lower=modngauss),
method='both')
summary(result3)
```

Modelos de fragilidade selecionados ao nível de 5

Modelo Lognormal

```
mod3< -coxph(Surv(início,fim,estado)~MEC_MCC+Fosforo+
```

```

frailty(id,dist="gauss"),data=dat1,method="breslow")
summary(mod3)
extractAIC(mod3)
extractAIC(mod3,k= log(160))
mod4< -coxph(Surv(início,fim,estado)~MEC_MCC+Fosforo+
frailty(id,dist="gamma"),data=dat1,method="breslow")
summary(mod4)
extractAIC(mod4)
extractAIC(mod4,k= log(160))

```

#### Função de risco basal acumulado em relação ao modelo PWP

```

dfitAG< -coxph(Surv(início,fim, estado)~1,id,data=dat1)
dfitPWP< -coxph(Surv(início,fim, estado)~strata(estrato),id, data=dat1)
surv1< -survfit(dfitAG)
surv2< -survfit(dfitPWP)
par(mfrow=c(1,2))
plot(surv1,fun='cumhaz',col=1:3,lty=1:3,c(100, 0.25),xlim=c(0,40),ylim=c(0,1.2),
ylab="Risco basal acumulado",xlab="Tempo",mark.time=F,main="Modelo AG")
plot(surv2,fun='cumhaz',col=1:3,lty=1:3,c(100, 0.25),xlim=c(0,40),ylim=c(0,1.2),
ylab="Risco basal acumulado",xlab="Tempo",
mark.time=F,main= "Modelo PWP")
legend(2,1.2,col=1:3,lty=1:3,legend=c("1 evento", "2 evento",
"3 evento"),cex=0.8,bty="n")

```

#### Análise de resíduos do modelo PWP

Resíduos padronizados de Shoenfeld:

```
cox.zph(pwp,transform="identity")
```

```
par(mfrow=c(2,2))
plot(cox.zph(pwp))
Resíduos de Martingale:
res.mart< -residuals(pwp,type="martingale",collapse=dat1$id)
plot(res.mart,ylim=c(-3,3),xlim=c(0,145),xlab="individuos",
ylab="Resíduos de Martingale")
abline(0,0,lty=2)
res.mart< -residuals(pwp,type="martingale")
```