UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Moisés de Farias Ramos

Análise de sobrevivência marginal aplicada a dados simulados no R

Campina Grande - PB

Fevereiro de 2018

Moisés de Farias Ramos

Análise de sobrevivência marginal aplicada a dados simulados no R

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira

Campina Grande - PB

Fevereiro de 2018

R175a Ramos, Moisés de Farias.
Análise de sobrevivência marginal aplicada a dados simulados no R [manuscrito] : / Moises de Farias Ramos. - 2018.
32 p.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2018.
"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Departamento de Estatística - CCT."

1. Análise de sobrevivência. 2. Múltiplos Eventos por Indivíduo. 3. Modelagem marginal.

21. ed. CDD 519.5

Moisés de Farias Ramos

Análise de sobrevivência marginal aplicada a dados simulados no R

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 09 de fevereiro de 2018.

BANCA EXAMINADORA

Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Prof. Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Prof. Me. Ednário Barbosa de Mendonça
Universidade Estadual da Paraíba

Dedico esse trabalho aos meus pais, aos meus irmãos, aos meus avós, à minha família e a todos que sempre estiveram ao meu lado.

Agradecimentos

Agradeço primeiramente a Deus por todas as bençãos recebidas, por me iluminar durante toda essa trajetória e por me presentear com a realização desse tão sonhado objetivo.

Aos meus pais, Inácio e Maria do Carmo, por sempre me conduzir para o caminho do bem, pelos incentivos e por acreditar na minha capacidade.

Aos meus irmãos, Izaías, Izabel, Daniel, Bernadete, Karolina, Ezequiel, Beatriz e Euflaudízia, pelo companheirismo e amizade.

Aos meus avós maternos Sebastião e Terezinha, e meus avós paternos, Manoel (in memoriam) e Adalzira (in memoriam), pelos conselhos e por passar experiência de tantos anos vividos.

A toda minha família e amigos por estar sempre presente, querendo sempre o meu bem e pelo carinho.

Ao meu professor e orientador Tiago Almeida de Oliveira pela contribuição e sugestões.

A todos meus colegas de curso, em especial a Arthur, Benedito e Roseane pela amizade e companheirismo durante todos estes anos de estudo.

A todos os professores que contribuíram para a minha graduação. Obrigado pelos ensinamentos e pelo apoio.

A todos meus amigos da "Casa dos Estudantes"em especial a Alcimar, Felipe e Thiago pelos momentos de descontração e fraternidade.

A instituição Universidade Estadual da Paraíba pela organização e por me tornar possível esse sonho.

A cidade de Campina Grande, cidade na qual tenho uma grande admiração por ser bastante hospitaleira.

"A persistência é o caminho do êxito"
(Charles Chaplin)

Resumo

O objetivo deste trabalho de conclusão de curso é utilizar técnicas de análise de sobrevivência com a finalidade de aplicar a um conjunto de dados com múltiplos eventos por indivíduo. Como o modelo de Cox não se enquadra para essa situação, optou-se por trabalhar com os seguintes modelos marginais: Modelo AG, Modelo PWP e Modelo WLW. Para selecionar o modelo com o melhor ajuste ao conjunto de dados, utilizou-se o método *backward* e os critérios AIC e BIC. Após isso, verificou-se se o pressuposto da proporcionalidade é satisfeito para o modelo selecionado via teste de proporcionalidade e gráfico dos resíduos padronizados de Schoenfeld e por último verificou-se se tinha pontos atípicos ou *outliers* através do gráfico dos resíduos de Martingale. O modelo que melhor se ajustou ao conjunto de dados foi o WLW. A simulação do conjunto de dados foi feita utilizando o pacote *survsim* do software R e as variáveis preditoras consideradas na simulação foram denominadas de x_1, x_2 e x_3. O intuito da modelagem é saber a influência que cada uma das variáveis ou fatores têm sobre o tempo de falha.

Palavras-chaves: Análise de Sobrevivência, Múltiplos Eventos por Indivíduo, Modelos Marginais.

Abstract

The objective of this course completion work is the application of survival analysis techniques in order to apply a data set with multiple events per individual. As the Cox model does not fit into this situation, we opted for working with the marginal models: Model AG, Model PWP and Model WLW. To select the model with the best fit to the dataset, we used the backward method and the AIC and BIC criteria. After which verify that the proportionality assumption is satisfied for the model selected via the proportionality test and Schoenfeld's standardized residue chart and lastly it was found to have atypical or "outliers" points across the Martingale residuos chart. The model that best fit the dataset was WLW. A simulation of the data set was done from the software R package and as predictor variables considered in the simulation were denominated x_1, x_2 e x_3. The purpose of modeling is to know the influence that each of the variables or factors has on the failure time.

Key-words: Survival analysis. Recurrence data. Marginal Models.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A análise de sobrevivência é um ramo da estatística que se originou a partir de pesquisas biomédicas, na qual o objetivo dos pesquisadores era estudar a mortalidade. Esta abordagem tem como objetivo analisar dados cuja variável resposta é o tempo até a ocorrência de um evento de interesse. A análise de sobrevivência tem uma característica que distingue de outras análises, que é a possibilidade de trabalhar com dados censurados. E esta censura acontece quando não se tem o tempo exato de falha por causa da perda de informações. A censura é dividida em três tipos: censura à direita, censura à esquerda e censura intervalar.

O modelo de Cox é uma ferramenta da análise de sobrevivência utilizada para avaliar o impacto que alguns fatores de risco têm no tempo até a ocorrência do evento de interesse, ele é adequado para situações em que ocorra apenas um evento de interesse por indivíduo e os tempos correspondentes a esses indivíduos devem ser independentes, segundo Cabete et al. (2012). Para múltiplos eventos por indivíduo o modelo de Cox não é indicado, e para resolver este problema surgiram vários métodos, dentre eles a Modelagem Marginal que também pode ser considerada a extensão do modelo proposto por Cox.

Os modelos marginais mais conhecidos são: AG (ANDERSEN; GILL, 1982), PWP (PRENTICE; WILLIAMS; PETERSON, 1981) e WLW (WEI; LIN; WEISSFELD, 1989). Em termos gerais, o modelo PWP destina-se a analisar eventos ordenados, onde o risco de ocorrência do evento seguinte é alterado pela ocorrência do evento que o antecede. O modelo AG, este apareceu na mesma linha de raciocínio do modelo anterior, mas pressupõe que os eventos têm o mesmo risco de ocorrerem, apesar de serem independentes entre si. No que diz respeito ao modelo WLW, este permite modelar separadamente o tempo até a ocorrência de cada evento, resolvendo assim a falta de robustez revelada pelos dois primeiros modelos quando os eventos têm uma estrutura de dependência não condicional, segundo Sousa-Ferreira (2013).

A simulação de dados consiste na utilização de ferramentes computacionais somado com o conhecimento matemático, probabilístico, etc, nas quais permite imitar o funcionamento de um conjunto de dados real. A simulação não visa apenas imitar a construção do modelo, ela também objetiva utilizar todo o método experimental através de teorias e hipóteses para que se possa ter um conjunto de dados que se aproxime da realidade.

Os objetivos deste trabalho são: analisar dados de sobrevivência que tenham como característica eventos recorrentes, utilizar os principais modelos marginais pra dados recorrentes em análise de sobrevivência, interpretar os coeficientes destes modelos marginais e realizar estatísticas diagnósticas nestes modelos marginais.

2 Fundamentação Teórica

2.1 Marco Histórico

O que modernamente se conhece por Análise de Sobrevivência teve seus primórdios no século XVII, mais precisamente em janeiro de 1662 quando o inglês John Graunt publicou em Londres o livro "Natural and Political Observations upon the Bill of Mortality" e o evento de interesse a ser estudado era a morte de indivíduos, posteriormente em 1693 o matemático e astrônomo Edmund Halley desenvolveu a primeira tábua de vida que analisava os dados de óbitos ocorridos em Breslaw entre 1691 e 1693, de acordo com Bastos e Rocha (2006).

Podemos dizer que o grande impulso que acelerou o desenvolvimento das técnicas de análise de sobrevivência foi a Segunda Guerra Mundial, com a finalidade de aplicá-las à indústria militar. E esse motivo somado com a modernização e a velocidade de computadores revolucionaram essa área da estatística.

Segundo Aalen et al. (2009), o artigo proposto por Kaplan e Meier (1958) foi basicamente a inauguração desta nova fase da análise de sobrevida pós-Segunda Guerra Mundial, onde eles apresentam seu famoso estimador não-paramétrico para a função de sobrevivência. Esse artigo é um dos mais citados em toda a história da estatística.

Outra grande revolução que impactou o cenário estatístico aconteceu em 1972, quando Sir David Cox desenvolveu um modelo de regressão semi-paramétrico. O modelo de Cox (1972) fornece as estimativas das razões de risco dos fatores estudados e a partir dessas estimativas observamos quais fatores tem mais influência no tempo até a ocorrência do evento de interesse. E é por essa inovação que o artigo proposto por Cox é um dos mais influentes tanto na análise de sobrevivência, como propriamente dito em toda literatura estatística.

No início dos anos 80 surge a ideia de que um evento de interesse pode ocorrer múltiplas vezes num mesmo indivíduo. Prentice, Williams e Peterson (1981) propuseram dois modelos que podem ser considerados como extensões do modelo de riscos proporcionais de Cox (1972).

Posteriormente, Andersen e Gill (1982) apresentaram um modelo de incrementos independentes. Dentre os três modelos referidos no trabalho para eventos recorrentes, este é o mais simples e o que possui os pressupostos mais fortes.

Wei, Lin e Weissfeld (1989) propuseram um modelo semi-paramétrico para analisar os tempos de falha. Este modelo trabalha com eventos ordenados e eventos não ordenados.

2.2 Análise de Sobrevivência

A análise de sobrevivência, eventualmente chamada de análise de sobrevida, é uma área da estatística que reúne um conjunto de técnicas e métodos estatísticos fundamentais para analisar o tempo de vida de indivíduos. Em análise de sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse, e este tempo é chamado de tempo de falha, segundo Colosimo e Giolo (2006). Essa técnica tem uma vasta aplicabilidade, por exemplo, na engenharia, medicina, ciências biológicas e financeiras. De acordo com Kleinbaum e Klein (2010), uma característica de dados de análise de sobrevivência é a presença de censura, que é a observação parcial da resposta, ou seja, por alguma razão não se pôde ter a observação completa do tempo de falha.

Uma das diferenças de análise de sobrevivência para outras técnicas estatísticas é sua capacidade de extrair informações de dados censurados. Já para dados sem censura, o que não é o caso deste trabalho, poderíamos trabalhar com outras técnicas como análise de regressão e planejamento de experimentos. Na tabela 1 tem-se as peculiaridades da análise de sobrevivência.

Tabela 1 – Características de Análises de Sobrevivência

Item	Descrição
Variável Dependente	Tempo até a ocorrência de um evento
Evento de Interesse	Falha
Tempo Inicial	Momento de início do estudo
Escala de Medida	Tempo real (dias, semanas, meses, anos)
Dados	Censurados

Tempo de falha é o tempo a partir do momento que o indivíduo começou a ser estudado até a ocorrência do evento de interesse. A falha pode ainda ocorrer devido a uma única causa ou devido a duas ou mais.

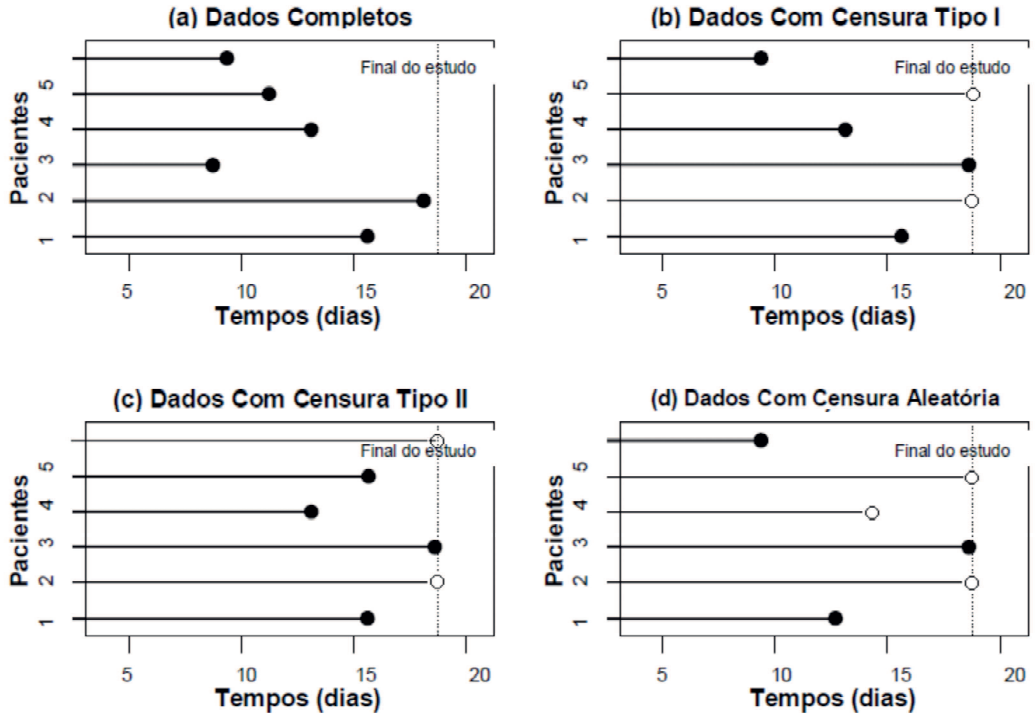
2.2.1 Censura

Segundo Colosimo e Giolo (2006), a principal característica de dados de sobrevivência é a censura, que é a observação parcial da resposta, ou seja, houve uma perda de informação que impossibilitou o conhecimento do tempo de falha do indivíduo. Há várias situações que podem acarretar em censura, por exemplo: a perda de acompanhamento do paciente ou indivíduo em estudo e a não ocorrência da falha durante todo o tempo estudado. Durante o estudo não podemos omitir nenhuma das observações censuradas, pois elas fornecem informações sobre o tempo de vida dos pacientes. Caso contrário, o cálculo das estatísticas de interesse pode gerar conclusões viciadas.

A censura se divide em três tipos. Censura à direita, à esquerda e intervalar.

- Censura à direita acontece quando o evento de interesse ocorre após o término do estudo. A censura à direita se divide em três:
  - Censura tipo I: Ocorre quando o estudo é encerrado após um período preestabelecido de tempo
  - Censura tipo II: Ocorre quando o estudo é encerrado após ter ocorrido o evento de interesse em um número preestabelecido de indivíduos
  - Censura aleatória: Ocorre se a observação for retirada no decorrer do estudo sem ter ocorrido o evento de interesse ou se o evento de interesse ocorrer por uma razão diferente da estudada.

Figura 1 – Ilustração de alguns mecanismos de censura em que (•) representa a falha e (○) a censura.



Fonte: Battistella et al. (2008)

- Censura à esquerda acontece quando não conhecemos o momento da ocorrência do evento de interesse, mas sabemos que ele ocorreu antes do tempo observado.
- Censura Intervalar ocorre quando não se sabe o tempo exato da ocorrência do evento de interesse, sabe-se que ele ocorreu dentro de um intervalo especificado, observa-se em estudos onde as visitas são periódicas aos pacientes.

2.2.2 Função de Sobrevivência

A função de sobrevivência é uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência. Ela é definida como a probabilidade de uma observação não falhar até um certo tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t. Em termos probabilísticos, isto é escrito como:

$$S(t) = P(T > t),$$

e de acordo com Lawless (2011), a função $S(t)$ é uma função monótona decrescente e contínua com $S(t) = 1$, quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$.

Já a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver ao tempo t, isto é:

$$F(t) = P(T \leq t) = 1 - S(t).$$

Portanto a partir dessas definições, percebe-se que $S(t)$ é simplesmente o complemento da função de distribuição acumulada:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t).$$

2.2.3 Função de Taxa de Falha ou de Risco

A função de risco representa o risco instantâneo de um indivíduo sofrer o evento entre o tempo t e $t + \Delta t$, dado que ele sobreviveu até o tempo t. Denota-se essa função por:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

A função de taxa de falha é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente (COLOSIMO; GIOLO, 2006).

2.2.4 Estimador de Kaplan-Meier

O Estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), é uma técnica não-paramétrica muito utilizada nas ciências da saúde, biológicas e engenharias. Ele estima uma curva de sobrevivência incorporando a informação da censura e é por conta dessa característica em especial que ele é muito usado no ramo da estatística. Ele também é chamado de estimador limite-produto.

O estimador de Kaplan-Meier é definido da seguinte forma,

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right),$$

em que:

- $t_1 < t_2 \cdots < t_k$, são os k tempos distintos e ordenados de falha,
- d_j é o nº de falhas no tempo t_j, $j = 1, \cdots, k$ e
- n_j é o nº de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

De acordo com Colosimo e Giolo (2006), as principais propriedades do estimador de Kaplan-Meier são as seguintes:

- é não-viciado para amostras grandes;
- é fracamente consistente;
- converge assintoticamente para um processo gaussiano
- é um estimador de máxima verossimilhança de $S(t)$.

O Estimador de Kaplan-Meier é muito conveniente para avaliação individual dos fatores de risco sobre o tempo de falha. Todavia, para avaliar conjuntamente múltiplos fatores, é preciso usar técnicas mais sofisticadas, como os modelos de regressão para sobrevivência.

2.2.5 Modelo de Cox

Em 1972, Cox desenvolveu um modelo de regressão semi-paramétrico, também conhecido como modelo de riscos proporcionais de Cox, modelo de Cox, ou regressão de Cox (Cox, 1972). Essa técnica é indicada quando se deseja estudar sobrevivência sob o prisma de causalidade ou da predição, pois fornece as estimativas das razões de risco dos fatores estudados, podendo-se avaliar o impacto que alguns fatores de risco ou fatores prognósticos têm no tempo até a ocorrência do evento de interesse (BUSTAMANTE-TEIXEIRA; FAERSTEIN; LATORRE, 2002).

O Modelo de Cox (1972) e semi-paramétrico, ele tem essa denominação porque possui um componente paramétrico e outro não paramétrico. A expressão geral do modelo de regressão de Cox para o *i-ésimo* indivíduo é a seguinte:

$$\lambda(t|x_i) = \lambda_0(t) exp\{x_i'\beta\},$$

em que $\lambda(t|x_i)$ é a função taxa de risco, $\lambda_0(t)$ é o componente não paramétrico conhecido como função de base, $\beta = (\beta_1, \ldots, \beta_p)$ é o vetor de dimensão 1xp de parâmetros associados às covariáveis e x_i' é o vetor de dimensão px1 de covariáveis observadas para o *i-ésimo* indivíduo.

O modelo proposto por Cox é também denominado modelo de riscos proporcionais, pois a razão das taxas de falha de dois indivíduos diferentes é constante no tempo. Portanto essa é a suposição básica para o uso deste modelo, ou de forma equivalente, as taxas de falha acumulada devem também ser proporcionais.

2.2.5.1 Estimação dos Parâmetros do Modelo

De acordo com Colosimo e Giolo (2006), os coeficientes β's medem os efeitos das covariáveis sobre a função de taxa de falha e são estimados a partir das observações amostrais para que o modelo fique determinado. Como o método da Máxima verossimilhança é inapropriado para estimar esses coeficientes por possuir um componente não-paramétrico, utiliza-se o Método da Máxima Verossimilhança Parcial que foi uma solução proposta por Cox em 1975.

- Método da Máxima Verossimilhança Parcial

Seja uma amostra de n indivíduos onde existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 < \ldots t_k$. Diz-se que a probabilidade da *i-ésima* observação falhar no tempo t_i, conhecendo as observações que estão sob risco em t_i é tal que:

$$\frac{\mathrm{P}[\text{ indivíduo falhar em } t_i \mid \text{sobreviveu a } t_i \text{ e história até } t_i]}{\mathrm{P}[\text{ uma falha em } t_i \mid \text{história até } t_i]} =$$

$$\frac{\lambda_i(t \mid x_i)}{\sum_{j \epsilon R(t_i)} \lambda_j(t \mid x_j)} = \frac{\lambda_0(t) \exp\{x_i' \beta\}}{\sum_{j \epsilon R(t_i)} \lambda_0(t) \exp\{x_j' \beta\}} = \frac{\exp\{x_i' \beta\}}{\sum_{j \epsilon R(t_i)} \exp\{x_j' \beta\}}$$

A função de verossimilhança almejada é constituída pelo produto de todos os termos dessa equação, associados aos tempos distintos de falha, como segue:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\{x_i' \beta\}}{\sum_{j \epsilon R(t_i)} \exp\{x_j' \beta\}} = \left(\prod_{i=1}^{n} \frac{\exp\{x_i' \beta\}}{\sum_{j \epsilon R(t_i)} \exp\{x_j' \beta\}} \right)$$

Desta maneira, os valores de β que maximizam a função de verossimilhança parcial, $L(\beta)$, são obtidos resolvendo-se o sistema de equações definido por $U(\beta) = 0$, em que $U(\beta)$ é o vetor escore de derivadas de primeira ordem da função $l(\beta)$. Ou seja:

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j \epsilon R(t_i)} x_j \exp\{x_j' \hat{\beta}\}}{\sum_{j \epsilon R(t_i)} \exp\{x_j' \hat{\beta}\}} \right] = 0$$

A função de verossimilhança parcial assume que os tempos de sobrevivência são contínuos e, consequentemente, não pressupõe a possibilidade de empates nos valores observados (COLOSIMO; GIOLO, 2006). Porém se ocorrer a igualdade entre os tempos de falha e de censura devido a uma grande escala, a função de verosimilhança modifica-se e usa-se uma aproximação proposta por Breslow (1972) e Peto (1972). Assim sendo,

considera-se s_i o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo tempo t_i $(i = 1, \ldots, k)$ e d_i o número de falhas neste mesmo tempo, logo a aproximação considera a seguinte função de verossimilhança parcial:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\{s_i'\beta\}}{\left[\sum_{j \epsilon R(t_i)} \exp\{x_j'\beta\}\right]^{d_i}}.$$

Vale salientar que essa aproximação é adequada quando o número de observações empatadas não é grande.

2.3 Análise de Sobrevivência Multivariada

Estudos longitudinais onde a variável resposta é o tempo até a ocorrência do evento de interesse de um indivíduo podem ser caracterizados de dois tipos: eventos terminais e eventos não terminais (eventos recorrentes). Eventos terminais são situações em que um indivíduo pode sofrer apenas um evento. Eventos não terminais ou eventos recorrentes, são situações em que um indivíduo pode sofrer mais do que um evento de interesse. Esses múltiplos eventos intra-indivíduo chamam-se de episódios uma vez que cada ocorrência é um novo episódio do mesmo evento.

Os eventos recorrentes são divididos em duas formas: eventos ordenados (múltiplos eventos do mesmo tipo) ou eventos não ordenados (eventos de diferentes tipos). Alguns casos em que um indivíduo apresenta mais de um evento de interesse são: acidentes vasculares cerebrais, um infrator pode reincidir no crime várias vezes, o atestado médico para um paciente com uma certa doença.

A aplicabilidade da análise de sobrevivência com eventos múltiplos por indivíduo é vasta, podemos implementar nas seguintes áreas: biomédicas, criminologia, confiabilidade fabril. No contexto estatístico temos várias abordagens para tratar de múltiplos eventos por indivíduo. Entretanto utilizaremos a Modelagem Marginal.

2.3.1 Modelagem Marginal

Os Modelos Marginais são extensões do modelo de Cox que permitem trabalhar com diversas situações nas quais ocorrem eventos recorrentes. A partir dos modelos marginais é possível estimar os parâmetros sem que a estrutura de dependência das observações seja considerada, utilizando o modelo de Cox usual. Posteriormente, é determinado uma estimativa robusta da variância que permite introduzir a correção necessária, dada a presença de observações correlacionados.

Os Modelos Marginais mais conhecidos e utilizados são os seguintes: Modelo AG (ANDERSEN; GILL, 1982), Modelo PWP (PRENTICE; WILLIAMS; PETERSON, 1981) e Modelo WLW (WEI; LIN; WEISSFELD, 1989).

2.3.1.1 Modelo de Andersen e Gill (AG)

No modelo AG proposto por Andersen e Gill (1982) assume-se que eventos recorrentes não são afetados por eventos que ocorreram anteriormente no mesmo indivíduo, de modo que o risco basal é igual em todos os intervalos de tempo intra-indivíduo, evidenciando assim a independência entre os tempos de falha. O modelo para representar o *i-ésimo* indivíduo é descrito da seguinte forma:

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

em que $\lambda_i(t)$ é a função de risco para o *i-ésimo* indivíduo, $\lambda_0(t)$ é o componente não paramétrico, $\beta = (\beta_1, \ldots, \beta_p)$ é o vetor de dimensão px1 de parâmetros associado às covariáveis, $x'_{mi}(t)$ é o vetor de dimensão 1xp de covariáveis observadas para o *i-ésimo* indivíduo no tempo t e a função indicadora de risco Y_{mi} assume valor 1 se o indivíduo i estiver sob observação e em risco no tempo t, caso contrário a função assume valor zero.

A grande diferença entre o modelo de Cox e o modelo AG é a seguinte: no primeiro modelo mencionado o indivíduo deixa de estar em risco a partir do momento que ocorre o evento de interesse, já no modelo AG o indivíduo permanece em risco mesmo após ocorrer a falha.

2.3.1.2 Modelo de Prentice, Williams e Peterson (PWP)

O modelo PWP parte da seguinte pressuposição, um indivíduo não pode estar sob risco para o *m-ésimo* evento sem que tenha experimentado o evento *m-1*. A partir dessa premissa, foi certificado que existe uma dependência entre os tempos de falha de um mesmo indivíduo e é por conta dessa dependência que a função de base varia de um evento para outro. A função taxa de falha é descrita da seguinte forma:

$$\lambda_{mi}(t) = Y_{mi}\lambda_{0m}(t)\exp\{x'_{mi}(t)\beta_m\},$$

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, $\beta_m = (\beta_1, \ldots, \beta_p)$ é o vetor de dimensão px1 de parâmetros associados às covariáveis do *m-ésimo* evento, $x'_{mi}(t)$ é o vetor de dimensão 1xp 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.

2.3.1.3 Modelo de Wei, Lin e Weissfeld (WLW)

O modelo WLW trata as respostas de um conjunto de dados ordenados como se fosse um problema de riscos competitivos com respostas não ordenadas. Isto significa que o indivíduo no início do período de observação é considerado estar sob risco de sofrer m

eventos e o tempo é sempre contado a partir do zero. Desta maneira, a função de taxa de falha para o *m-ésimo* evento do *i-ésimo* indivíduo é expressa da seguinte forma:

$$\lambda_{mi}(t) = Y_{mi}\lambda_{0m}(t)\exp\{x'_{mi}(t)\beta_m\}$$

em que $\lambda_{mi}(t)$ é a função de risco do *i-ésimo* indivíduo, $\lambda_{0m}(t)$ é o componente não paramétrico no *m-ésimo* evento, $\beta = (\beta_1, \ldots, \beta_p)$ é o vetor de dimensão px1 de parâmetros associados às covariáveis no *m-ésimo* evento e $x'_{mi}(t)$ é o vetor de dimensão 1xp 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.

2.3.1.4 Estimação dos Parâmetros

Assim como no modelo de Cox, a estimação dos parâmetros dos modelos AG, PWP e WLW é feita através do método de máxima verossimilhança parcial, ignorando a correlação existente entre as observações. A função de verossimilhança parcial para o modelo AG pode ser representada como:

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

Já para os modelos PWP e WLW, a função de verossimilhança é dada por:

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

De acordo com Mota (2013), para se obter uma estimativa robusta, é necessário a correção na variância dos $\hat{\beta}$ e para isso utiliza-se o estimador $jackknife$, que consiste na exclusão de uma ou mais amostras do conjunto total observado, recalculando-se o estimador a partir dos valores restantes. Esse procedimento gera uma estimativa não viciada da variância para dados correlacionados sempre que a observação deixada de fora for independente das observações que entram.

Utiliza-se os resíduos de *jackknife* para obter uma estimativa de *jackknife* agrupado por indivíduo e os mesmos são definidos como:

$$\mathbf{J_i} = \hat{\beta} - \hat{\beta}_{(\mathbf{i})},$$

em que $\hat{\beta}_{(i)}$ é o resultado do ajuste que inclui todas as observações exceto o indivíduo i.

Therneau e Grambsch (2000) propôs uma forma de calcular os valores dos resíduos de *jackknife* pelo método de Newton-Raphson, baseado na seguinte fórmula:

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

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

Agrupada por indivíduo, essa estimativa é escrita da seguinte maneira:

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

em que $\bar{\mathbf{J}}$ é a matriz de médias das colunas de $\mathbf{J}$. A variância passa a ser escrita como $\mathbf{D}'\mathbf{D} = \mathcal{I}^{-1}(\mathbf{U}'\mathbf{U})\mathcal{I}^{-1}$, que pode ser vista como um estimador sanduíche ABA, em que $A = \mathcal{I}^{-1}$ é a estimativa usual da variância e $\mathbf{U}'\mathbf{U}$ é o termo de correção.

2.4 Métodos de Seleção das Covariáveis

2.4.1 Seleção *Backward*

O método de seleção de covariáveis *Backward* incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada.

A decisão de retirada da variável é tomada baseando-se em testes F parciais, que são calculados para cada variável como se ela fosse a última a entrar no modelo.

2.4.2 Critério de Informação de Akaike (AIC)

Pode-se utilizar o Critério de Informação de Akaike (AIC) para selecionar os modelos marginais que melhor se ajustam ao conjunto de dados. O AIC é bem representativo pois ele indica o modelo que envolve o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável dependente.

O AIC é expressado da seguinte forma:

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

em que K é o número de parâmetros estimáveis do modelo.

2.4.3 Critério de Informação de Bayesiano (BIC)

O critério BIC é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O modelo com menor BIC é considerado o de melhor ajuste. Ele é definido como:

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

2.5 Análise de Resíduos

Para modelos de sobrevivência a definição de resíduo é mais complexa do que em regressão linear. Por exemplo, na análise de sobrevivência não se usa o resíduo obtido pela diferença entre o valor observado e o valor esperado, pois este resíduo não trabalha com observações censuradas.

Durante a análise de resíduo o principal aspecto a ser verificado é a proporcionalidade dos riscos que é o pressuposto básico do modelo de Cox, outros fatores a serem considerados são a presença de pontos aberrantes (*outliers*) e pontos influentes.

2.5.1 Resíduos de Schoenfeld

Os resíduos de Schoenfeld são usados para verificar a suposição de riscos proporcionais. Se o efeito de uma covariável muda gradativamente durante o tempo de observação, diz-se que o efeito da covariável é tempo-dependente. Define-se os resíduos de Schoenfeld da seguinte maneira:

$$r_{ik} = \delta_i(x_{ik} - a_{ik}),$$

em que δ_i é o indicador de ocorrência de evento no indivíduo i, a_{ik} é uma média ponderada dos valores das covariáveis dos indivíduos em risco no tempo t_i e x_{ik} é o valor da covariável.

2.5.2 Resíduos *Martingale*

Denomina-se os resíduos *Martingale*, M_i, como sendo a diferença entre o número observado de eventos para um indivíduo e o número esperado dado o modelo ajustado. Descreve-se esses resíduos da seguinte forma:

$$M_i = N_i - E_i,$$

onde N_i é o número de eventos observados no intervalo $[0, \infty)$ e E_i é o número de eventos esperados sob o modelo ajustado no intervalo $[0, \infty)$.

Esses resíduos são usados para identificar a melhor forma funcional para uma dada covariável e auxilia na indentificação de pontos atípicos. Os gráficos desse resíduo devem apresentar um comportamento aleatório em torno do zero, caso o modelo seja adequado.

3 Metodologia

3.1 Material

O conjunto de dados foi simulado através do *software* R. Para esse trabalho foi utilizado o pacote survsim Morina e Navarro (2014), que é um pacote bastante recomendado para simular dados de sobrevivência simples e complexos. O objetivo era simular dados de sobrevivência com eventos recorrentes, e para esse processo foi utilizada a função **rec.ev.sim**, esta função armazena todos os argumentos ou objetos utilizados na rotina. Estes objetos são descritos a seguir:

- **n**: valor inteiro indicando o tamanho desejado da coorte a ser simulada.
- **foltime**: número real que indica o tempo máximo de acompanhamento da coorte simulada.
- **dist.ev**: vetor de tamanho arbitrário indicando as distribuições de sobrevivência para os tempos de sobrevivência não censurados. Os possíveis valores são, weibull para a distribuição Weibull, lnorm para a distribuição log-normal e llogistic para a distribuição log-logística. Se um indivíduo sofre mais episódios do que distribuições especificadas, a última distribuição especificada é usada para gerar tempos correspondentes a episódios posteriores.
- **anc.ev**: vetor de tamanho arbitrário de componentes reais contendo os parâmetros auxiliares das distribuições dos tempos de sobrevivência não censurados.
- **beta0.ev**: vetor de tamanho arbitrário de componentes reais contendo os parâmetros β_0 das distribuições dos tempos de sobrevivência não censurados.
- **dis.cens**: vetor de tamanho arbitrário indicando as distribuições dos tempos de sobrevivência censurados. Os possíveis valores são, weibull para a distribuição Weibull, lnorm para a distribuição log-normal, llogistic para a distribuição log-logística e unif para a distribuição uniforme. Se nenhuma distribuição for introduzida, espera-se que o tempo de censura siga uma distribuição de Weibull.
- **anc.cens**: vetor de tamanho arbitrário de componentes reais contendo os parâmetros auxiliares das distribuições dos tempos de sobrevivência censurados.
- **beta0.cens**: vetor de tamanho arbitrário de componentes reais contendo os parâmetros β_0 das distribuições dos tempos de sobrevivência censurados.

- **x**: lista de vetores que indica a distribuição e os parâmetros de qualquer covariável que o usuário precisa introduzir na coorte simulada.

- **lambda**: número real que indica a duração média de cada evento ou tempo de risco descontínuo, assumiu seguir uma distribuição de Poisson truncada em zero. Seu valor padrão é NA, correspondente ao caso em que a duração de cada evento ou tempo de risco descontínuo é uma informação desnecessária para o usuário.

- **max.ep**: valor inteiro que corresponde ao número máximo permitido de episódios por indivíduo. Seu valor padrão é Inf, ou seja, o número de episódios por indivíduos não é limitado.

- **priskb**: proporção de sujeitos em risco antes do início do acompanhamento, o padrão é 0.

- **max.old**: tempo máximo em risco antes do início do acompanhamento.

O processo de simulação é descrito a seguir: simulou-se 50 indivíduos em uma coorte de 3600 dias, com uma média de duração de 2,18; 2,33 e 2,40 dias, média para cada evento do primeiro para o terceiro, estes valores foram retirados do pacote survsim Morina e Navarro (2014) e do artigo Moriña, Navarro et al. (2014), outros parâmetros, bem como covariáveis acrescentadas não necessariamente utilizaram os mesmos valores do artigo de Moriña, Navarro et al. (2014). O parâmetro $\beta_1 = -0.3$, representando uma redução de 26% ($\exp(-0.3) = 0,74$) no fator de aceleração ou razão de tempo da variável x_1 que segue uma distribuição Bernoulli com $p = 0,5$, indicando um decréscimo em cada recorrência, o parâmetro $\beta_2 = 2\,(\exp(2) = 7,38)$, um acréscimo de 7 vezes na chance de recorrência para cada acréscimo na variável x_2 (simulada seguindo uma distribuição normal (0,1)) e $\beta_3 = -0.9\,(\exp(-0,9) = 0,40)$, um decréscimo de 60% na chance de recorrência para cada acréscimo de uma unidade na variável x_3. Um parâmetro de heterogeneidade seguindo uma distribuição uniforme no intervalo de $[0,8; 1,2]$. Foi assumido para cada evento de recorrência uma distribuição diferente (Log normal, Logística e Weibull). As variáveis podem ser fatores que podem influenciar o tempo até o evento de interesse, sendo que este pode ser o tempo até a ocorrência de ataque cardíaco, ou até a ocorrência de algum tipo de câncer. A covariável x_1, por exemplo, por se tratar de uma variável dicotômica, pode ser pensada como consumo ou não de tabaco, uso de drogas, sexo ou outra variável que pode assumir dois valores. A covariável x_2 pode ser pensada como valores de colesterol não HDL, assumindo que o valor de referência para esta covariável é de 160 mg/dl, estando a amplitude portanto entre $[60; 260]$ mg/dl. A covariável x_3 pode ser pensada como a idade do paciente, por exemplo, usando como referência a idade de 45 anos, por se tratar de uma idade de risco para o aparecimento de doenças cardiovasculares, pois, quatro entre cincos pessoas acometidas de doenças cardiovasculares estão acima dos 65 anos, deste

modo os valores para esta variável são entre [20; 70] anos. Os sintomas do colesterol alto só se manifestam quando seus valores são muito elevados. Por isso, após os 20 anos de idade recomenda-se realizar exames e apesar de raros nesta idade pode se pensar que começam os ataques cardíacos.

3.2 Métodos

A partir da simulação dos dados e do modelo marginal completo, foi feita a seleção das covariáveis para os modelos marginais AG, PWP e WLW, utilizou-se o método *backward* e os critérios AIC e BIC para essa seleção. Através do AIC e BIC selecionou-se o melhor modelo marginal dentre os três mencionados no trabalho, verificou-se se foi satisfeito o pressuposto da proporcionalidade dos riscos no modelo selecionado tanto pelo teste de proporcionalide como pelo gráfico dos resíduos padronizados de Schoenfeld, e por último, dado o modelo ajustado, verificou-se a existência de pontos atípicos ou *outliers* através do gráfico dos resíduos de Martingale.

4 Aplicação

Iniciamos o processo de modelagem dos dados com a seleção das covariáveis dos modelos AG, PWP e WLW. As covariáveis incluídas no modelo completo foram x_1, x_2 e x_3 e os métodos de seleção das covariáveis utilizados foram o *backward*, AIC e BIC. Através do Valor P testou-se a significância de cada covariável ao nível de 5%.

Na Tabela 2, temos o resultado da seleção das covariáveis para o modelo AG e o método utilizado foi o *backward* no qual foi selecionada apenas a covariável x_2 e o Valor P indicou que esta covariável é significativa ao nível de significância de 5%. Para a seleção do melhor modelo PWP, foram utilizados os critérios de informações de Akaike e Bayesiano, a covariável x_2 mostrou-se significativa ao nível de 5% de significância e os resultados se encontram na Tabela 3. Já para a seleção do melhor modelo WLW foram usados os mesmos critérios do modelo PWP e foram selecionas as covariáveis x_1 e x_2, mas como a covariável x_1 não mostrou significância ao nível de 5%, então ela foi excluída do modelo, significância essa demostrada pelo Valor P, para mais detalhes visualize a Tabela 4.

Tabela 2 – Resultado da seleção das covariáveis para o modelo AG através do método *backward*.

Covariável	Estimativas	Erro Padrão	Erro Padrão (Robusto)	Valor P
x_2	0,050532	0,011457	0,008974	$1{,}79\times10^{-08}$

A covariável x_2 pode ser entendida como o nível de colesterol não HDL, ou seja, de acordo com a Tabela 2, para cada aumento em uma unidade de colesterol não HDL, espera-se um acréscimo de $5,18\%$ ($\exp(0,050532) = 1,0518$) no número de ataques cardíacos, pressupondo-se que esta seja a doença estudada.

Tabela 3 – Resultado da seleção das covariáveis para o modelo PWP através dos critérios AIC e BIC.

Covariável	Estimativas	Erro Padrão	Erro Padrão (Robusto)	Valor P
x_2	0,16655	0,02984	0,04656	0,000347

Pelos resultados da Tabela 3, percebe-se um acréscimo de 18,12% ($\exp(0,16655) = 1,1812$) na chande de ataque cardíaco, este resultado é bastante próximo do modelo AG.

Os resultados do modelo WLW, apresentados na Tabela 4, são semelhantes aos encontrados no modelo AG, para cada uma unidade de acréscimo em x_2 espera-se o aumento de 5,22% ($\exp(0,05088) = 1,0522$) nos casos de ataque cardíaco, por se tratar

Tabela 4 – Resultado da seleção das covariáveis para o modelo WLW através dos critérios AIC e BIC.

Covariável	Estimativas	Erro Padrão	Erro Padrão (Robusto)	Valor P
x_2	0,05088	0,01075	0,01361	0,000185

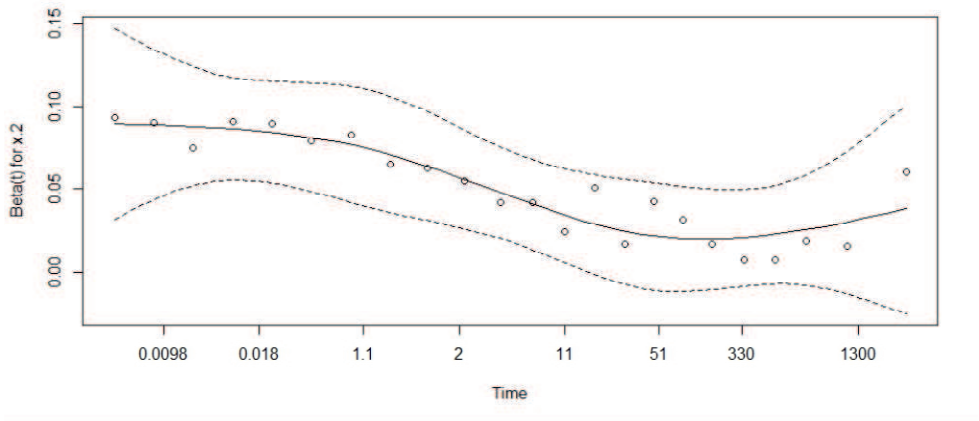
de dados simulados sem grandes pertubações é de se esperar que o desempenho dos três modelos fosse próximo, porém a literatura tem encontrado valores mais próximo entre os modelos AG e PWP. Sagara et al. (2014), utilizou a modelagem de dados recorrentes para os casos de malária, alguns dos modelos aplicados foram o AG e PWP para casos contínuos. Os resultados encontrados entre os modelos foram similares. Amorim e Cai (2015) discutiram as principais recomendações para se utilizar os modelos AG, PWP, indicando que quando houver correlação o modelo AG é usualmente indicado para análise de dados quando há dependência entre eventos subsequentes. Este modelo tem sido utilizado para doenças como câncer e hospitalizações devido a causas cardiovasculares. Eles discutiram também que o modelo PWP é preferível ao modelo AG quando o efeito das covariáveis são diferentes nos eventos subsequentes. Isto pode ser o caso de doenças tais como infecções virais.

Os valores do AIC e BIC, apesar de não serem diretamente comparáveis, dão indícios de que o melhor modelo marginal para ajustar o conjunto de dados é o PWP e esses valores são apresentados na tabela 5. Quando utilizamos esses critérios, deve-se procurar o modelo com os menores valores.

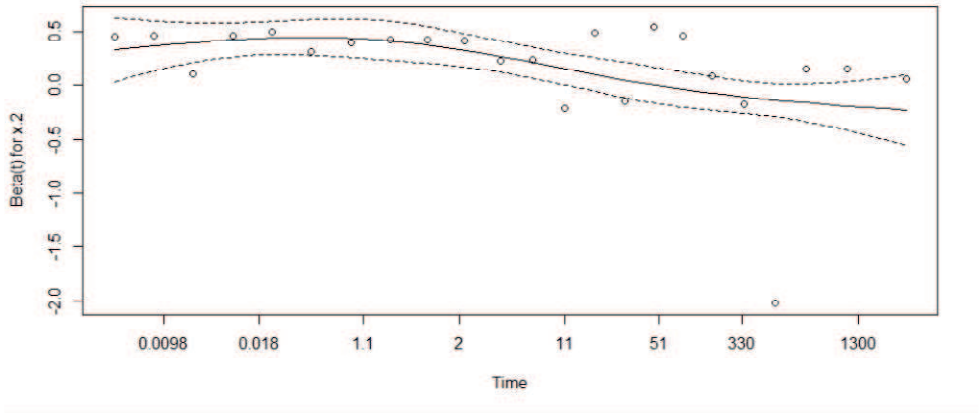
Tabela 5 – Valores do AIC e BIC para os modelos marginais

	Modelos Marginais		
Critério	AG	PWP	WLW
AIC	149,0573	94,81187	134,4653
BIC	150,1928	95,94736	136,7363

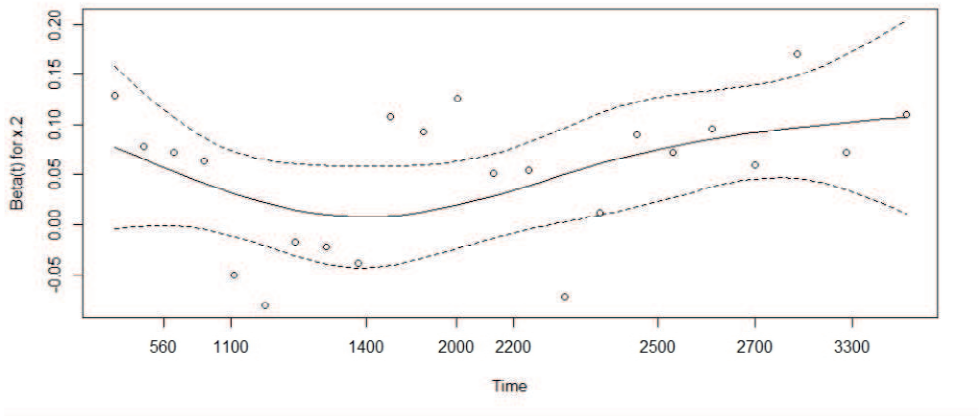
Os gráficos das Figuras 2a, 2b e 2c, são relativos aos resíduos de Schoenfeld, para os modelos selecionados AG, PWP e WLW, respectivamente. Por meio destes gráficos é possível ver que o modelo de PWP é o que apresenta uma forma sugestiva de não proporcionalidade, fato este corroborado pela Valor P do teste de proporcionalidade (Valor P = 0,0313), nos demais modelos AG e WLW, não houveram valores menores que 5%, indicando a proporcionalidade do mesmo. Esta quebra de proporcionalidade compromete as inferências feitas a partir do modelo PWP, sendo recomendado neste caso se trabalhar com a dicotomização desta variável, fato este não realizado neste trabalho.



(a) Resíduos padronizados de Schoenfeld para o modelo AG



(b) Resíduos padronizados de Schoenfeld para o modelo PWP

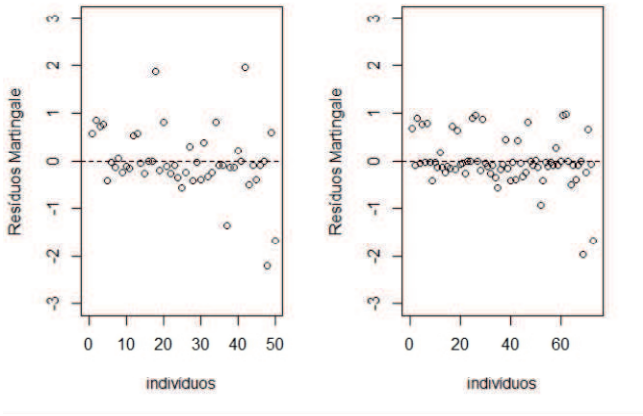


(c) Resíduos padronizados de Schoenfeld para o modelo WLW

Figura 2 – Gráficos dos resíduos padronizados de Schoenfeld vs tempo de sobrevivência para a covariável x_2 nos modelo AG, PWP, WLW, respectivamente

Por apresentar os menores valores dos critérios AIC e BIC e por satisfazer a suposição de proporcionalidade, o modelo WLW é o melhor para ajustar o conjuntos de dados.

Figura 3 – Gráficos dos resíduos padronizados de *Martingale* sem e com medidas recorrentes no modelo de WLW.



Os resíduos de martingale referentes ao modelo WLW, foram os que apresentaram um comportamento que mereça destaque neste trabalho, pois duas observações passaram os limiares de [-2,2] desvios, indicando que são observações que podem influenciar no ajuste do modelo.

5 Conclusão

O objetivo deste trabalho foi analisar dados de sobrevivência que tenham como característica eventos recorrentes.

Foram utilizados os métodos *backward* e os critérios AIC e BIC para a seleção das covariáveis. O melhor modelo que se ajustou aos dados foi o PWP, contendo a covariável x_2, porém como houve a quebra de proporcionalidade, este modelo não é recomendado. Sendo assim, o modelo selecionado foi o WLW contendo a covariável x_2. Através dos resultados da simulação e das suposições feitas, pode-se dizer que para cada unidade de acréscimo em x_2 espera-se o aumento de 5,22% nos casos de ataque cardíaco.

Ao fim deste estudo percebe-se que a Análise de Sobrevivência é uma área da estatística que deve ser trabalhada com bastante cuidado, caso contrário, pode-se ter conclusões errôneas.

Referências

AALEN, O. O. et al. History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, v. 5, n. 1, p. 1–28, 2009. Citado na página 12.

AMORIM, L. D.; CAI, J. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, Oxford University Press, v. 44, n. 1, p. 324–333, 2015. Citado na página 27.

ANDERSEN, P. K.; GILL, R. D. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, JSTOR, p. 1100–1120, 1982. Citado 3 vezes nas páginas 11, 12 e 18.

BASTOS, J.; ROCHA, C. Análise de sobrevivência: Conceitos básicos. *Arquivos de Medicina*, ArquiMed-Departamento de Edições Científicas da AEFMUP, v. 20, n. 5-6, p. 185–187, 2006. Citado na página 12.

BATTISTELLA, P. M. D. et al. Análise de sobrevivência aplicada à estimativa da vida de prateleira de salsicha. Florianópolis, SC, 2008. Citado na página 14.

BRESLOW, N. E. Contribution to discussion of papeer by dr cox. *J. Roy. Statist. Assoc., B*, v. 34, p. 216–217, 1972. Citado na página 17.

BUSTAMANTE-TEIXEIRA, M. T.; FAERSTEIN, E.; LATORRE, M. do R. Técnicas de análise de sobrevida survival analysis techniques. *Cad. Saúde Pública*, v. 18, n. 3, p. 579–594, 2002. Citado na página 16.

CABETE, A. B. d. A. et al. *Análise de sobrevivência com acontecimentos múltiplos: aplicação ao estudo do tempo até à ocorrência de enfarte do miocárdio*. Tese (Doutorado), 2012. Citado na página 11.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. 1ª edição. ed. São Paulo: Editora Edgard Blucher, 2006. Citado 4 vezes nas páginas 13, 15, 16 e 17.

COX, D. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 34, n. 2, p. 87–22, 1972. Citado 2 vezes nas páginas 12 e 16.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado na página 15.

KLEINBAUM, D. G.; KLEIN, M. *Survival analysis*. [S.l.]: Springer, 2010. v. 3. Citado na página 13.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011. v. 362. Citado na página 15.

MORINA, D.; NAVARRO, A. Survsim: simulation of simple and complex survival data. *R package version*, v. 1, n. 2, 2014. Citado 2 vezes nas páginas 23 e 24.

MORIÑA, D.; NAVARRO, A. et al. The r package survsim for the simulation of simple and complex survival data. *Journal of Statistical Software*, v. 59, n. 2, p. 1–20, 2014. Citado na página 24.

MOTA, T. S. Modelagem em análise de sobrevivencia com eventos recorrentes aplicada a dados da área médica. Universidade Estadual Paulista (UNESP), 2013. Citado na página 20.

PETO, R. Contribution to the discussion of paper by dr cox. *J. Royal stat. Soc.*, v. 34, p. 205–207, 1972. Citado na página 17.

PRENTICE, R. L.; WILLIAMS, B. J.; PETERSON, A. V. On the regression analysis of multivariate failure time data. *Biometrika*, Oxford University Press, v. 68, n. 2, p. 373–379, 1981. Citado 3 vezes nas páginas 11, 12 e 18.

SAGARA, I. et al. Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. *Malaria journal*, BioMed Central, v. 13, n. 1, p. 293, 2014. Citado na página 27.

SOUSA-FERREIRA, I. Modelos de sobrevivência aplicados à análise de acontecimentos múltiplos. 2013. Citado na página 11.

THERNEAU, T. M.; GRAMBSCH, P. M. Modeling survival data: extending the cox model. Springer, 2000. Citado na página 20.

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*, Taylor & Francis, v. 84, n. 408, p. 1065–1073, 1989. Citado 3 vezes nas páginas 11, 12 e 18.