Juliana Luz Passos Argenton

Árvore de Regressão para Dados Censurados e Correlacionados

CAMPINAS

2013

Universidade Estadual de Campinas

Instituto de Matemática, Estatística
e Computação Científica

Juliana Luz Passos Argenton

Árvore de Regressão para Dados Censurados e Correlacionados

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em estatística.

Orientadora: Hildete Prisco Pinheiro

Este exemplar corresponde à versão final da dissertação defendida pela aluna Juliana Luz Passos Argenton, e orientada pela Profa. Dra. Hildete Prisco Pinheiro.

Assinatura da Orientadora

Hildete Pinheiro

Campinas

2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

Ar37a

Argenton, Juliana Luz Passos, 1984-
Árvore de regressão para dados censurados e correlacionados / Juliana Luz Passos Argenton. – Campinas, SP : [s.n.], 2013.

Orientador: Hildete Prisco Pinheiro.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Árvores de decisão. 2. Análise de sobrevivência (Biometria). 3. Correlação (Estatística). I. Pinheiro, Hildete Prisco,1966-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Regression tree for censored and correlated data
Palavras-chave em inglês:
Decision tree
Survival analysis (Biometry)
Correlation (Statistics)
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Hildete Prisco Pinheiro [Orientador]
Victor Hugo Lachos Dávila
Antonio Carlos Pedroso de Lima
Data de defesa: 05-12-2013
Programa de Pós-Graduação: Estatística

Dissertação de Mestrado defendida em 05 de dezembro de 2013 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Hildete Pinheiro

Prof(a). Dr(a). HILDETE PRISCO PINHEIRO

Prof(a). Dr(a). VICTOR HUGO LACHOS DÁVILA

Prof(a). Dr(a). ANTÔNIO CARLOS PEDROSO DE LIMA

Abstract

The objective of this work is to present methods of regression trees for censored and correlated data. The dataset analyzed was obtained from a survey, in which 119 families (1712 individuals) living in Baependi village, in the Brazilian state of Minas Gerais, were interviewed. Two methodologies based on the proportional hazard model are presented. The first disregards the possible correlation among the individuals of the same family, using the first step of a full likelihood estimation procedure for splitting nodes. In the second methodology, the correlation among the individuals of the same family is incorporated in the proportional hazard model through a frailty variable with Gamma distribution. In this case, the value of the Score statistic is used for choosing the best splitting node. The main purpose of the analysis is to evaluate the variables that increase the risk of hypertension, type II diabetes and high cholesterol, which are the top three main factors that increase the risk of heart conditions. The response variables are the age-of-onset of these risk factors. Censoring is defined by observing the individual's age-of-onset at the moment of diagnosis and also at the moment of the survey. This way, an age-of-onset higher than the age at the moment of the survey indicates censoring.

Keywords: Decision Tree, Survival Analysis (Biometry) and Correlation (Statistics).

Resumo

O objetivo deste trabalho é apresentar uma metodologia de árvore de regressão para dados censurados e correlacionados. O conjunto de dados analisado foi obtido a partir de uma pesquisa

realizada entre Dezembro de 2005 e Janeiro de 2006, que entrevistou 119 famílias (1712 indivíduos) que vivem no pequeno vilarejo de Baependi, no Estado de Minas Gerais. São apresentadas duas metodologias com base no modelo de riscos proporcionais, a primeira desconsidera a possível correlação existente entre os indivíduos de uma mesma família e usa a primeira iteração da estimativa da verossimilhança completa nas divisões dos nós. Na segunda metodologia apresentada, a correlação entre os indivíduos de uma mesma família é incorporada no modelo de riscos proporcionais através de uma variável de fragilidade com distribuição Gama, neste caso o valor da estatística Escore é usado para escolher a melhor divisão dos nós. O objetivo da análise é avaliar as variáveis que aumentam o risco de apresentar hipertensão, diabetes tipo II e colesterol alto, que são os três principais fatores que aumentam o risco de doenças no coração. As variáveis respostas são as idades de diagnóstico desses fatores de risco. A censura é definida de acordo com a observação da idade do indivíduo no momento do diagnóstico da doença e a idade do indivíduo no momento da pesquisa. Desta forma, uma idade de diagnóstico maior que a idade no momento da pesquisa caracteriza a censura.

Palavras-chave: Árvores de Decisão, Análise de Sobrevivência (Biometria) e Correlação (Estatística).

Sumário

Agradecimentos

A Deus, por ter me dado condições de conquistar mais esse objetivo.

À minha querida orientadora, Hildete, pela condução do desenvolvimento da dissertação, pela amizade e também por tornar bastante agradável nossas reuniões.

À professora Júlia por fornecer o banco de dados e por ter sido muito prestativa, respondendo todas as minhas dúvidas com bastante solicitude.

Aos meus pais, Erinaldo e Elisete, que sempre apoiaram as minhas decisões.

Ao meu esposo Gease, pela paciência com minhas muitas ausências e principalmente pela paciência com minhas lamentações de cansaço.

Ao meu amigo Rafael, que sempre confiou que eu seria capaz de alcançar mais essa meta e me ajudou muito a entender parte da teoria que encontrei dificuldades.

À minha amiga Natália, que me ajudou a entender um pouco melhor as variáveis do conjunto de dados, contribuindo bastante para a interpretação dos resultados.

À minha sogra Ireni, que com muito carinho cuidou do meu bem mais precioso, meu filho, aos fins de semana e feriados, enquanto eu escrevia essa dissertação.

À minha irmã, Josiane, pela companhia contínua e por me ajudar, com muita eficiência, a resolver alguns problemas da minha vida pessoal, me doando, assim, mais tempo para os estudos.

À Allergisa que não só me incentivou a fazer o mestrado como me deu suporte, me liberando em horário de serviço para assistir às aulas e também escrever essa dissertação.

A todos meu muito obrigada!!

Lista de Ilustrações

Capítulo 1

Introdução

As doenças cardiovasculares, como infartos e acidentes vasculares encefálicos, são a primeira causa de morte no mundo, com 7 milhões de mortos em 2011, segundo a lista das dez principais causas de mortalidade publicada em Julho de 2013 pela Organização Mundial de Saúde. No Brasil, segundo o Ministério da Saúde, aproximadamente 30% das mortes que ocorreram em 2010 tiveram como principal causa doenças do aparelho circulatório. Fatores de risco são situações que podem facilitar e tornar mais rápido o desenvolvimento dessas doenças e que em geral, precedem por muitos anos o aparecimento da doença cardiovascular. O conhecimento dos fatores de risco é de extrema importância e representa grande avanço na medicina preventiva, já que sua detecção precoce poderá reduzir de modo significativo o desenvolvimento de doenças, ou pelo menos retardar seu início. Alguns fatores de risco para doenças do aparelho circulatório são diabetes tipo II, pressão arterial elevada, colesterol elevado, fumo, obesidade, hereditariedade, sexo, idade entre outros. O sexo, idade e hereditariedade, por exemplo, são fatores não controláveis, fazendo com que o controle dos outros seja ainda mais importante, uma vez que esses fatores tem efeito aditivo.

Vários estudos de doenças complexas, ou seja, doenças que não exibem o padrão clássico de herança Mendeliana, incluem o estudo da variável idade de diagnóstico, como, por exemplo, os cânceres de mama e próstata, hipertensão, além de diabetes tipo I e mal de Alzheimer. Entre os membros de uma família, é possível observar correlação entre as idades de diagnóstico dos mesmos.

O objetivo desse trabalho é avaliar o efeito de fatores ambientais e familiares na idade de

diagnóstico de três fatores de risco das doenças cardiovasculares: hipertensão, diabetes tipo II e colesterol alto.

O conjunto de dados analisado foi obtido do Estudo do Coração das Famílias de Baependi, descrito em Oliveira et al. (2008). Este estudo obteve informações sobre 119 famílias (1712 indivíduos), que vivem no pequeno vilarejo de Baependi, no Estado de Minas Gerais. Os dados foram coletados entre Dezembro de 2005 e Janeiro de 2006, de acordo com um desenho amostral planejado. Considerando que famílias com apenas um ou dois indivíduos não fornecem muita informação para estudos de famílias, foram analisados os dados de 81 famílias, envolvendo 1673 indivíduos. Para cada participante um questionário foi usado para obter informações sobre as relações familiares, características demográficas, histórico médico e fatores de risco ambientais. Foram realizadas medidas antropométricas, exames físicos e eletrocardiograma dos participantes por estudantes de medicina treinados. Além disso, glicemia de jejum, colesterol total, frações de lipoproteínas e triglicerídeos foram obtidos por técnicas padrões em amostras de sangue.

Árvores de sobrevivência para dados correlacionados são construídas com o objetivo de encontrar os grupos de indivíduos com maior predisposição a adquirir as doenças segundo as variáveis coletadas na pesquisa. Árvores de sobrevivência que não incorporam a dependência entre indivíduos de uma mesma família também são construídas com o intuito de comparar os resultados das duas metodologias. Ambas as metodologias estão apresentadas no capítulo 4. No capítulo 2 encontra-se uma introdução à análise de sobrevivência, como o método de estimação do estimador de Kaplan-Meier, o teste *Logrank* para comparação das curvas de sobrevivência, e introdução aos modelos paramétricos, semi-paramétricos de Cox e de fragilidade. No capítulo 3 é apresentado um resumo dos métodos *Bootstrap* e Validação Cruzada *V-Fold* que são usados para corrigir o viés de estimação das medidas de qualidade de divisão. Estas medidas também são apresentadas no capítulo 4. As aplicações estão no capítulo 5, onde é feita uma descrição detalhada do conjunto de dados e são aplicadas duas metodologias de árvore de sobrevivência, em que a primeira não incorpora no modelo a informação de correlação entre os indivíduos de uma mesma família, e a segunda o faz utilizando o modelo de fragilidade gama.

Capítulo 2

Análise de Sobrevivência

Análise de Sobrevivência é a expressão utilizada para designar a análise estatística de dados quando a variável em estudo representa o tempo desde um instante inicial bem definido até a ocorrência de determinado acontecimento de interesse. Assim sendo, a variável aleatória em estudo é não negativa e pode representar, por exemplo, o tempo até a falha de determinado componente elétrico, no campo da Confiabilidade Industrial, a duração de uma greve ou período de desemprego no contexto econômico, o tempo de resposta a um inquérito no âmbito da Psicologia ou o tempo até a morte de um indivíduo com determinada doença em Medicina. As observações resultantes são chamadas de tempos de vida e a ocorrência do evento de interesse é chamada falha.

A análise de sobrevivência permite estudar tempos de vida, também designados por tempos de sobrevivência, ultrapassando as dificuldades inerentes a este tipo de dados. A característica fundamental é a existência de censura, ou seja, para alguns indivíduos pode não ser possível observar o acontecimento de interesse durante todo o período em que estiveram em observação. As censuras podem ser classificadas em três grupos: censura à direita, censura à esquerda e censura intervalar.

Se o tempo de ocorrência do evento de interesse é maior, ou seja, está à direita do tempo registrado, estamos perante um caso de **censura à direita**. Como exemplo, considere um estudo clínico em que o acontecimento de interesse é a morte de um indivíduo após o diagnóstico de um determinado tumor maligno; se o indivíduo estiver vivo no final do estudo, tem-se uma observação

censurada à direita.

Outro tipo de censura é a chamada **censura à esquerda**, que ocorre quando o tempo registrado é maior que o tempo de falha (não observado), isto é, o evento de interesse já aconteceu quando o indivíduo foi observado. Um estudo para determinar a idade em que as crianças aprendem a ler em uma determinada comunidade pode ilustrar a situação de censura à esquerda. Quando os pesquisadores começaram a pesquisa algumas crianças já sabiam ler e não lembravam com que idade isto tinha acontecido, caracterizando, desta forma, observações censuradas à esquerda.

Pode-se ainda ter outro tipo de censura, que se chama **censura intervalar**. A censura intervalar ocorre, por exemplo, quando não é conhecido o instante exato de morte de um indivíduo, mas sabe-se que esta ocorreu dentro de um determinado intervalo de tempo.

Este trabalho foca casos de censura à direita, que pode ser classificada da seguinte maneira:

Censura de tipo I: Suponha que o tempo de duração de um determinado estudo é definido previamente, isto é, a data final é pré-determinada. Neste caso, só se sabe o tempo de vida de um indivíduo se o evento ocorrer antes do instante pré-definido, isto é, estando os n indivíduos sujeitos a períodos de observação $C_1, C_2, \cdots, C_n$ fixados pelo investigador, o tempo de vida t_i é observado se, e somente se, $t_i < C_i$. Neste caso, o número de eventos observados é aleatório.

Censura de tipo II: São colocados em estudo n indivíduos, mas o estudo termina quando se der a r-ésima falha, sendo r um número pré-definido ($1 \leq r \leq n$). A amostra obtida consiste das r primeiras estatísticas ordinais, isto é, $t_{(1)} \leq \cdots \leq t_{(r)}$, e os restantes $n - r$ indivíduos são censurados no instante $t_{(r)}$. Neste caso, o tempo de duração do estudo é uma variável aleatória.

Censura de tipo III: Na maioria dos estudos clínicos, o período do estudo é fixado e os pacientes entram no estudo em diferentes tempos durante aquele período. Alguns podem falecer antes do fim do estudo; seus tempos exatos de sobrevivência são conhecidos, outros podem ser perdidos durante o acompanhamento, outros podem sobreviver até o fim do estudo. Para os pacientes perdidos, os tempos de sobrevivência são, no mínimo, o período compreendido desde as suas chegadas até o último contato no estudo. Para aqueles que sobreviverem até o fim do estudo, os tempos de sobrevivência são, pelo menos, o período entre suas chegadas e o fim do estudo. Esses dois últimos tipos de observações são censuradas. Uma vez que os tempos de entrada no estudo não são simultâneos, os tempos de censura são também diferentes.

Este trabalho lidará basicamente com Censuras do Tipo I, visto que o tempo de duração do estudo foi determinado pelo tempo de duração das entrevistas, e os indivíduos que não foram diagnosticados com a doença de interesse até o momento da entrevista foram censurados.

Uma característica muito importante, amplamente empregada em trabalhos de análise de sobrevivência é a hipótese de **censura não-informativa**. Este tipo de censura ocorre quando a distribuição do tempo de sobrevivência (T) não fornece nenhuma informação sobre a distribuição do tempo de censura (C).

Censura não-informativa ocorre, por exemplo, quando o instante de censura é definido previamente no estudo. Outro caso de censura não-informativa é quando n indivíduos são acompanhados e decide-se censurar todos os indivíduos sobreviventes no instante da m-ésima resposta ($m < n$), então, dada a hipótese de independência entre os indivíduos, este procedimento de censura é não-informativo (apesar de T não ser estritamente independente de C).

Censura informativa ocorre quando existe associação entre o mecanismo de censura e o tempo de sobrevivência. Por exemplo, quando um indivíduo abandona um estudo por razões (por exemplo doença) associadas ao tempo de sobrevivência ou quando se perde o contato com um paciente pelo fato de o mesmo sentir-se suficientemente recuperado e não mais comparecer para a continuidade do estudo.

2.1 Distribuição do Tempo de Vida

Seja T o tempo até a ocorrência de um determinado evento (falha, morte, desenvolvimento de uma doença). Então T é uma variável aleatória não negativa. Considere T como sendo uma variável aleatória absolutamente contínua, cuja distribuição pode ser caracterizada por qualquer uma das seguintes funções, que são matematicamente equivalentes: Função densidade de probabilidade, Função de sobrevivência e Função risco.

Na prática, as três funções podem ser usadas para ilustrar diferentes aspectos dos dados. Um problema básico na análise de dados de sobrevivência é estimar pelo menos uma dessas três funções e inferir padrões de sobrevivência dos dados.

1. **Função densidade de probabilidade:** é definida como o limite da probabilidade de um

indivíduo falhar no pequeno intervalo $(t,\, t + \Delta t)$ por unidade de tempo Δt:

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{P(\text{um indivíduo falhar no intervalo}(t, t + \Delta t))}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(T \in (t, t + \Delta t))}{\Delta t} \end{aligned} \tag{2.1.1}$$

2. **Função de sobrevivência:** é a probabilidade de um indivíduo sobreviver mais que t unidades de tempo:

$$\begin{aligned} S(t) = P(T > t) = 1 - F(t) \\ S(0) = 1 \\ S(\infty) = 0 \end{aligned} \tag{2.1.2}$$

3. **Função Risco:** também chamada de taxa de falha é a probabilidade de falha num pequeno intervalo de tempo, dado que o indivíduo sobreviveu até o início do intervalo:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \tag{2.1.3}$$

A função risco pode, também, ser interpretada como a taxa instantânea de ocorrência de falha, dado que o indivíduo estava em risco até o tempo t, e também pode ser vista como potencial instantâneo de falha ou velocidade de falha.

2.1.1 Equivalência entre as funções

De (2.1.2) e (2.1.3), tem-se que:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{2.1.4}$$

Usando a relação entre densidade e distribuição, obtém-se:

$$f(t) = -S'(t) \tag{2.1.5}$$

Empregando (2.1.5) em (2.1.4), observa-se:

$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} log(S(t)) \tag{2.1.6}$$

Resolvendo a equação diferencial acima,

$$S(t) = exp\left(-\int_0^t \lambda(x)dx\right) \tag{2.1.7}$$

De (2.1.7) e (2.1.4)

$$f(t) = \lambda(t)exp\left(-\int_0^t \lambda(x)dx\right) \tag{2.1.8}$$

Define-se o risco acumulado como:

$$\Lambda(t) = \int_0^t \lambda(x)dx. \tag{2.1.9}$$

O risco acumulado mede o risco de ocorrência do evento até t. Maiores detalhes a respeito de conceitos e aplicações de análise de sobrevivência podem ser obtidos em Colosimo e Giolo (2006).

2.2 Estimador de Kaplan-Meier

É o estimador da função de sobrevivência mais utilizado em estudos clínicos, foi proposto por Kaplan e Meier (1958). Este estimador é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\hat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo t}}{\text{nº total de observações no estudo}}.$$

$\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $1/n$, em que n é o tamanho da amostra. Se existirem empates em um certo tempo t, o tamanho do degrau fica multiplicado pelo número de empates. O estimador de Kaplan-Meier, na sua construção, considera tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra. Considere:

1. $t_1 < t_2 \cdots < t_k$, os k tempos distintos e ordenados de falha,

2. d_j o número de falhas em t_j, $j = 1, \cdots, k$, e

3. n_j o número 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.

O estimador de Kaplan-Meier é, então, definido como:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \tag{2.2.1}$$

e sua variância assintótica é dada por:

$$\hat{\text{Var}}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{j:t_j<t} \left(\frac{d_j}{n_j(n_j - d_j)} \right). \tag{2.2.2}$$

O Estimador de Kaplan-Meier é um estimador de máxima verossimilhança, como pode ser verificado a seguir.

Seja $t_1 < t_2 < \cdots < t_k$ os tempos de falha observados em uma amostra de tamanho $n = n_0$ de uma população homogênea com função de sobrevivência S. Suponha que d_j observações falham no tempo t_j e m_j observações são censuradas no intervalo $[t_j, t_{j+1})$, $j = 0, \cdots, k$, onde $t_0 = 0$ e $t_{k+1} = \infty$. Seja $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$ o número de observações em risco no tempo imediatamente anterior ao t_j. A probabilidade de falha em t_j é $P(T = t_j) = S(t_j^-) - S(t_j)$.

Assumindo que a contribuição de um tempo de sobrevivência censurado em t_{jl} para a verossimilhança é $P(T > t_{jl}) = S(t_{jl})$, ou seja, assumindo que o tempo censurado t_{jl} fornece informação somente que o tempo não observado de falha é maior que t_{jl}, tem-se que a verossimilhança é da forma

$$L = \prod_{j=0}^{k} \left(\left[S(t_j^-) - S(t_j) \right]_j^d \prod_{l=1}^{m_j} S(t_{jl}) \right), \tag{2.2.3}$$

O estimador (não-paramétrico) de máxima verossimilhança é a função de sobrevivência $\hat{S}(t)$ que maximiza L.

$\hat{S}(t)$ é descontínua nos tempos observados de falha (i.e., dá alguma probabilidade positiva em cada t_j), caso contrário L seria igual a 0. Como $t_{jl} \geq t_j$, $S(t_{jl})$ é maximizada fazendo $S(t_{jl}) = S(t_j)$ $(j = 1, \cdots, k; l = 1, \cdots, m_j)$. Então o estimador de máxima verossimilhança, $\hat{S}(t)$, é uma função de sobrevivência discreta com componentes de risco $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$ em $t_1, \cdots, t_k$, respectivamente. Assim,

$$\hat{S}(t_j) = \prod_{l=1}^{j} (1 - \hat{\lambda}_l) \tag{2.2.4}$$

e

$$\hat{S}(t_j^-) = \prod_{l=1}^{j-1}(1-\hat{\lambda}_l), \tag{2.2.5}$$

em que os $\hat{\lambda}_l$'s são escolhidos de tal forma que maximizem a função

$$\prod_{j=1}^{k}\left[\lambda_j^{d_j}\prod_{l=1}^{j-1}(1-\lambda_j)^{d_j}\prod_{l=1}^{j}(1-\lambda_l)^{m_j}\right] = \prod_{j=1}^{k}\lambda_j^{d_j}(1-\lambda_j)^{(n_j-d_j)}, \tag{2.2.6}$$

obtida substituindo (2.2.4) e (2.2.5) em (2.2.3). O valor de $\hat{\lambda}_j$ que maximiza (2.2.6) é $\hat{\lambda}_j = \frac{d_j}{n_j}$ $(j = 1, \cdots, k)$, e portanto o EKM é dado por (2.2.1).

2.3 Teste *logrank*

O teste *logrank* (Mantel, 1966) é o mais usado para comparação de curvas de sobrevivência e é particularmente apropriado quando a razão das funções risco dos grupos a serem comparados é aproximadamente constante. A estatística deste teste é a diferença entre o número observado de falhas em cada grupo e uma quantidade que, para muitos propósitos, pode ser pensada como o correspondente número esperado de falhas sob a hipótese nula. A seguir é apresentado o teste de igualdade de duas funções de sobrevivência.

Considere as funções de sobrevivência $S_1(t)$ e $S_2(t)$. Sejam $t_1 < t_2 < \cdots < t_k$ os tempos de falhas distintos da amostra formada pela combinação das duas amostras individuais. Suponha que no tempo t_j aconteçam d_j falhas e que n_j indivíduos estejam sob risco em um tempo imediatamente inferior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra $i; i = 1, 2$ e $j = 1, \cdots, k$. Em cada tempo de falha t_j, os dados podem ser dispostos em forma de uma tabela de contingência 2 x 2 com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i (ver Tabela 2.1).

Condicional à experiência de falha e censura até o tempo t_j (fixando as marginais de coluna) e ao número de falhas no tempo t_j (fixando as marginais de linha), a distribuição de d_{2j} é, então,

Tabela 2.1: Tabela de contingência gerada no tempo t_j.

	Grupo 1	**Grupo 2**	
Falha	d_{1j}	d_{2j}	d_j
Não-Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

uma hipergeométrica:

$$D_{2j}/D_{1j} + D_{2j} = d_j \sim \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}}, \tag{2.3.1}$$

como mostrado a seguir.

Seja D_{ij} a variável aleatória que representa o número de falhas que ocorre na amostra i no tempo t_j, então $D_{ij} \sim Binomial(n_{ij}, p_i)$. Sob $H_0 : p_1 = p_2 = p$,

$$P[D_{2j}/D_{1j} + D_{2j} = d_j] = \frac{\binom{n_{1j}}{d_{1j}}p^{d_{1j}}(1-p)^{n_{1j}-d_{1j}}\binom{n_{2j}}{d_{2j}}p^{d_{2j}}(1-p)^{n_{2j}-d_{2j}}}{\binom{n_{1j}+n_{2j}}{d_{1j}+d_{2j}}p^{d_{1j}+d_{2j}}(1-p)^{n_{1j}+n_{d2j}-(d_{1j}+d_{2j})}}. \tag{2.3.2}$$

Cancelando os termos em relação a p e a $(1-p)$ obtem-se (2.3.1). Lembre-se que $n_{1j}+n_{2j} = n_j$ e $d_{1j} + d_{2j} = d_j$.

A média de D_{2j} é $m_{2j} = \frac{n_{2j}d_j}{n_j}$, o que equivale a dizer que, se não houver diferença entre as duas populações no tempo t_j, o número total de falhas (d_j) pode ser dividido entre as duas amostras de acordo com a razão entre o número de indivíduos sob risco em cada amostra e o número total sob risco. A variância de D_{2j} obtida a partir da distribuição hipergeométrica é:

$$(V_j)_2 = \frac{n_{2j}(nj - n_{2j})d_j(n_j - d_j)}{n_j^2(n_j - 1)}.$$

Então, a estatística $D_{2j} - m_{2j}$ tem média zero e variância $(V_j)_2$. Se as k tabelas de contingência forem independentes, um teste aproximado para a igualdade das duas funções de sobrevivência pode ser baseado na estatística:

$$T_L = \frac{[\sum_{j=1}^{k}(d_{2j} - m_{2j})]^2}{\sum_{j=1}^{k}(V_j)_2},$$

que, sob a hipótese nula $H_0 : S_1(t) = S_2(t)$ para todo t no período de acompanhamento, tem distribuição aproximada qui-quadrado com 1 grau de liberdade.

2.4 Modelos Paramétricos

Suponha que o interesse seja utilizar um modelo de regressão para estudar a relação entre a variável resposta e uma covariável. No entanto, o tipo de resposta (tempo até a ocorrência de um evento) e a presença de censura não permitem, em geral, a utilização do modelo de regressão linear. Junte a isso o fato de que a distribuição da resposta tende também, em geral, a ser assimétrica na direção dos maiores tempos de sobrevivência, o que torna inapropriado o uso da distribuição normal para o componente estocástico do modelo. Existem duas formas de enfrentar o problema da modelagem estatística em análise de sobrevivência. São elas:

1. transformar a resposta para tentar retornar ao modelo linear normal; ou

2. utilizar um componente determinístico não-linear nos parâmetros e uma distribuição assimétrica para o componente estocástico.

Na verdade, as duas formas podem ser equivalentes. Utilizar um modelo linear para a transformação logarítmica da resposta é equivalente a usar o componente determinístico: $exp\{\beta_0 + \beta_1 x\}$ e distribuição log-normal para o erro. Existem, no entanto, outras distribuições assimétricas possíveis para o erro, que não possibilitam o retorno para o modelo linear. A seguir são descritos alguns modelos paramétricos usuais que apresentam distribuições assimétricas para o erro.

Na Tabela 2.2 é apresentada a função distribuição de probabilidade e a função de sobrevivência para algumas distribuições.

Nos modelos paramétricos a função de sobrevivência e taxa de falhas dependem de um vetor de parâmetros $\boldsymbol{\theta}$ que pode ser estimado via máxima verossimilhança, desta forma os estimadores de máxima verossimilhança para $\boldsymbol{\theta}$ são obtidos maximizando o logaritmo da verossimilhança dado a seguir:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i log(f_{\boldsymbol{\theta}}(t_i)) + (1 - \delta_i) log(S_{\boldsymbol{\theta}}(t_i))$$

Tabela 2.2: Função densidade de probabilidade e função de sobrevivência para algumas distribuições

Distribuição	**f.d.p.**	**Sobrevivência**	$\boldsymbol{\theta}$
Exponencial	$\lambda exp\{-\lambda t\}$	$(-\lambda t)$	(λ)
Weibull	$\lambda\rho t^{\rho-1}exp\{-\lambda t^{\rho}\}$	$exp\{-\lambda t^{\rho}\}$	(λ,ρ)
Valor Extremo	$\frac{1}{\rho}exp\{\frac{t-\lambda}{\rho} - exp\{\frac{t-\lambda}{\rho}\}\}$	$exp\{-exp\{\frac{t-\lambda}{\rho}\}\}$	(λ,ρ)
Log Normal	$\frac{1}{t}\phi(\frac{log(t)-\lambda}{\rho})$	$1-\Phi(\frac{log(t)-\lambda}{\rho})$	(λ,ρ)

em que $\delta_i = 1$ se falha e $\delta_i = 0$ se censura.

As derivadas parciais em relação à $\boldsymbol{\theta}$ são

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \delta_i v(t_i,\boldsymbol{\theta})) + (1-\delta_i)a(t_i,\boldsymbol{\theta})f'_{\boldsymbol{\theta}}(t_i), \tag{2.4.1}$$

sendo

$$v(t_i,\boldsymbol{\theta}) = \frac{f'_{\boldsymbol{\theta}}(t_i)}{f_{\boldsymbol{\theta}}(t_i)},\ f'_{\boldsymbol{\theta}}(t_i) = \frac{\partial f_{\boldsymbol{\theta}}(t_i)}{\partial \boldsymbol{\theta}} \text{ e } a(t_i,\boldsymbol{\theta}) = \frac{f_{\boldsymbol{\theta}}(t_i)}{S_{\boldsymbol{\theta}}(t_i)}.$$

A matriz de informação observada é dada por:

$$\Sigma(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = -\sum_{i=1}^{n} \delta_i V(t_i,\boldsymbol{\theta}) + (1-\delta_i)M(t_i,\boldsymbol{\theta}), \tag{2.4.2}$$

em que $V(t_i,\boldsymbol{\theta}) = \frac{\partial v(t_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ e $M(t_i,\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}[a(t_i,\boldsymbol{\theta})f'_{\boldsymbol{\theta}}(t_i)]$. Desta forma os EMV são obtidos usando métodos numéricos, como o de Newton-Raphson:

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \Sigma(\hat{\boldsymbol{\theta}}_k)^{-1}U(\hat{\boldsymbol{\theta}}_k),$$

em que $U(\boldsymbol{\theta})$ e $\Sigma(\boldsymbol{\theta})$ são dados por (2.4.1) e (2.4.2), respectivamente. Seja $diag(x)$ uma matriz de zeros fora da diagonal principal e com os elementos de x na diagonal principal. Assim, a convergência é obtida quando $\epsilon_k = max(diag(\boldsymbol{\theta}_{k-1})^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}))$ é suficientemente pequeno.

A estimação da função de sobrevivência e da função taxa de falha é feita utilizando as propriedades de invariância dos estimadores de máxima verossimilhança. Desta forma $\hat{S}_{\boldsymbol{\theta}}(t) = S_{\hat{\boldsymbol{\theta}}}(t)$ e $\hat{a}(t,\boldsymbol{\theta}) = a(t,\hat{\boldsymbol{\theta}})$.

Os testes de hipótese do tipo $H_0 : \boldsymbol{C\theta} = \boldsymbol{d}$ contra $H_1 : \boldsymbol{C\theta} \neq \boldsymbol{d}$ podem ser feitos via estatística de Wald:

$$T_W = (\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d})'\Sigma(\hat{\boldsymbol{\theta}})(\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{d}) \xrightarrow{D} \chi^2_c;$$

em que c é o posto de $\boldsymbol{C}$.

2.5 O Modelo de Cox

Modelos que utilizam a função risco têm sido amplamente utilizados para modelar dados de sobrevivência. Um dos mais empregados é o modelo de Cox que foi o primeiro método semi-paramétrico proposto para modelar dados de sobrevivência na presença de covariáveis. O modelo de riscos proporcionais proposto por Cox (1972) é dado por:

$$\lambda(t|x) = g(\boldsymbol{\beta}'\boldsymbol{x})\lambda_0(t), \tag{2.5.1}$$

em que $g(\cdot)$ é uma função positiva que assume o valor 1 quando seu argumento é igual a zero, $\lambda_0(\cdot)$ representa uma função risco não-negativa para uma observação quando $\boldsymbol{x} = \mathbf{0}$ e $\boldsymbol{\beta}'$ é o vetor de coeficientes a serem estimados. Várias formas funcionais podem ser empregadas para $g(\cdot)$, entretanto a candidata natural, e que será tratada aqui, é a função $exp\{\cdot\}$.

Este modelo assume que o vetor de covariáveis $\boldsymbol{x}$ tem um efeito multiplicativo na função risco. Isto implica que a sua estrutura impõe proporcionalidade entre funções de risco de diferentes níveis de covariáveis, não permitindo que elas se cruzem e dependam do tempo t.

Este modelo é composto pelo produto de dois componentes, um não paramétrico e outro paramétrico. O componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não negativa do tempo. É usualmente chamado de função de risco basal, pois $\lambda(t) = \lambda_0(t)$ quando $\boldsymbol{x} = \mathbf{0}$. Utilizar a função exponencial como o componente paramétrico garante que $\lambda(t)$ seja não negativa.

O modelo de Cox é também denominado de modelo de riscos proporcionais, pois a razão das taxas de falha de dois indivíduos/ grupos diferentes é constante no tempo, isto é, a razão das funções de taxa de falha para os indivíduos/ grupos i e j,

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\lambda_0(t)exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = exp\{\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_j'\boldsymbol{\beta}\},$$

não depende do tempo. A presença do componente não-paramétrico $\lambda_0(t)$ na função de verossimilhança torna o método de máxima verossimilhança inapropriado. Por isso Cox propôs o método de máxima verossimilhança parcial, apresentado a seguir:

Considere que em uma amostra de n indivíduos existam $k \leq n$ falhas distintas nos tempos $t_1 < t_2 \cdots < t_k$. Uma forma simples de entender a verossimilhança parcial considera o seguinte

argumento condicional: a probabilidade condicional da i-ésima observação vir a falhar no tempo t_i conhecendo quais observações estão sob risco em t_i é:

$$P[\text{indivíduo falhar em } t_i|\text{uma falha em } t_i \text{ e história até } t_i] =$$

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

$$= \frac{\lambda_i(t|\boldsymbol{x}_i)}{\sum_{j\in R(t_i)}\lambda_j(t|\boldsymbol{x}_j)} = \frac{\lambda_0(t)exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\lambda_0(t)exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = \frac{exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} \tag{2.5.2}$$

em que $R(t_i)$ é o conjunto dos índices das observações sob risco no tempo t_i. Observe que condicional à história de falhas e censuras até o tempo t_i, o componente não paramétrico $\lambda_0(t)$ desaparece de (2.5.2). A função de verossimilhança a ser utilizada para se fazer inferências acerca dos parâmetros do modelo é, então, formada pelo produto de todos os termos representados por (2.5.2) associados aos tempos distintos de falha, isto é,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k}\frac{exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n}\left(\frac{exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}}\right)^{\delta_i}, \tag{2.5.3}$$

em que δ_i é o indicador de falha. Os valores de $\boldsymbol{\beta}$ que maximizam a função de verossimilhança parcial, $L(\boldsymbol{\beta})$, são obtidos resolvendo-se o sistema de equações definido por $U(\boldsymbol{\beta}) = \mathbf{0}$, em que $U(\boldsymbol{\beta})$ é o vetor Escore de derivadas de primeira ordem da função $l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta}))$. Isto é,

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

A função de verossimilhança parcial (2.5.3) assume que os tempos de sobrevivência são contínuos e, consequentemente, não pressupõe a possibilidade de empates nos valores observados. Na prática, empates podem ocorrer nos tempos de falha ou de censura devido à escala de medida. Então a função de verossimilhança parcial (2.5.3) deve ser modificada para incorporar as observações empatadas quando estas estão presentes. A aproximação para (2.5.3) proposta por Breslow (1972) é simples e frequentemente usada nos pacotes estatísticos comerciais. Considere $\boldsymbol{s}_i$ o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo

tempo t_i $(i = 1, \cdots, k)$ e d_i o número de falhas neste mesmo tempo. A aproximação mencionada anteriormente considera a seguinte função de verossimilhança parcial:

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

2.6 Modelo de Fragilidade Gama

Modelo de fragilidade é um modelo de efeitos aleatórios para variáveis de tempo, em que o efeito aleatório (fragilidade) tem um efeito multiplicativo no risco basal. Os modelos de fragilidade são extensões do modelo de riscos proporcionais de Cox apresentado na Seção 2.5. O modelo de Cox assume homogeneidade na população de estudo, porém em muitas aplicações essa suposição não é satisfeita, como por exemplo, em estudos de famílias e aplicações de genética quantitativa ou em casos de medidas repetidas. Nessas situações é preciso incorporar no modelo a correlação entre os membros da mesma família ou correlações entre medidas repetidas para o mesmo indivíduo. A seguir é apresentado o modelo de fragilidade gama, que é um modelo semiparamétrico:

$$\lambda_{ik}(t|\boldsymbol{x}) = exp\{\boldsymbol{\beta'x}\}\lambda_0(t)w_i, \tag{2.6.1}$$

em que $\lambda_0(\cdot)$ representa uma função risco arbitrária para uma observação quando $\boldsymbol{x} = \mathbf{0}$, $\boldsymbol{\beta'}$ é o vetor de coeficientes de efeitos fixos a serem estimados e W é o efeito aleatório ou a variável de fragilidade. A fragilidade W é uma variável aleatória que varia na população reduzindo (w <1) ou aumentando (w >1) o risco individual. O ponto mais importante é que W é não-observável. As fragilidades w_i $(i = 1, \cdots, n)$ são assumidas serem uma amostra independente de variáveis aleatórias W_i com distribuição Gama de média igual a 1 e variância desconhecida ϑ, isto é, $W_i \sim Gama(1/\vartheta, 1/\vartheta)$. A variância ϑ, pode ser vista neste modelo como uma escolha natural para medir o quanto de heterogeneidade está presente. Valores grandes de ϑ refletem um alto grau de heterogeneidade entre as famílias e uma forte associação dentro das famílias.

Uma contribuição com relação à razão de riscos neste modelo foi dada por Klein (1992). Três situações distintas relacionadas à interpretação do vetor ocorrem quando $\vartheta \neq 0$. Estas são apresentadas a seguir.

1. Se forem comparados dois indivíduos, j e l, de uma mesma família, ou seja, com variáveis de fragilidade iguais, tem-se a proporcionalidade dos riscos e, consequentemente, a mesma interpretação do vetor $\boldsymbol{\beta}$ do caso tratado para dados independentes (seção 2.5). De fato, neste caso, tem-se para a razão dos riscos, $H(t)$, que :

$$H(t) = \frac{w_i \lambda_0(t) exp\{\boldsymbol{x}_j' \boldsymbol{\beta}\}}{w_i \lambda_0(t) exp\{\boldsymbol{x}_l' \boldsymbol{\beta}\}} = exp\{(\boldsymbol{x}_j - \boldsymbol{x}_l)' \boldsymbol{\beta}\}.$$

2. Se, no entanto, forem comparados dois indivíduos com os mesmos valores das covariáveis, suponha $\boldsymbol{x}_1$, mas pertencentes a famílias distintas, por exemplo, famílias 1 e 2, a razão dos riscos de falha não será 1, mas sim a razão entre as variáveis de fragilidade, isto é:

$$H(t) = \frac{w_1 \lambda_0(t) exp\{\boldsymbol{x}_1' \boldsymbol{\beta}\}}{w_2 \lambda_0(t) exp\{\boldsymbol{x}_1' \boldsymbol{\beta}\}} = \frac{w_1}{w_2}.$$

3. Se, finalmente, forem comparados dois indivíduos com covariáveis diferentes, $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$, pertencentes a famílias distintas, famílias 1 e 2, tem-se, de acordo com Klein (1992), que:

$$H(t) = exp\{(\boldsymbol{x}_1 - \boldsymbol{x}_2)' \boldsymbol{\beta}\} \left[\frac{1 + \vartheta \hat{\Lambda}_0(t) exp\{\boldsymbol{x}_2' \boldsymbol{\beta}\}}{1 + \vartheta \hat{\Lambda}_0(t) exp\{\boldsymbol{x}_1' \boldsymbol{\beta}\}} \right]. \tag{2.6.2}$$

A razão dos riscos depende do tempo t. A primeira parte de (2.6.2) tende a 1 quando $t \to \infty$, independentemente de quais sejam os valores das covariáveis, ou seja, Ainda, conforme ϑ cresce, a razão converge para 1 mais rapidamente.

2.6.1 Estimação

Procedimentos de estimação têm sido baseados na construção de uma função de verossimilhança e sua otimização. Em particular, o algoritmo EM (Dempster et al., 1977) tem sido usado, considerando para isto que os valores da fragilidade são dados perdidos (*missing*). Outra abordagem considerada, que apresenta similaridade com o algoritmo EM, é a que considera o modelo de fragilidade gama como um modelo penalizado, otimizando, assim, no processo de estimação, a função de verossimilhança parcial penalizada. Procedimentos bayesianos que fazem uso de métodos computacionalmente intensivos, como o *Monte Carlo Markov chain* (MCMC) (Hastings 1970), têm sido também sugeridos para estimação dos parâmetros desse modelo. A seguir é apresentado

o método de estimação via verossimilhança penalizada mais detalhadamente.
Considere a seguinte formulação de (2.6.1)

$$\lambda_{ik}(t) = \lambda_0(t)exp\{\boldsymbol{x}'_{ik}\boldsymbol{\beta} + \zeta_i\}. \tag{2.6.3}$$

A alternativa proposta é considerar (2.6.3) como um modelo de Cox penalizado usando, assim, no processo de estimação, a função de verossimilhança penalizada (Hougaard, 2000, Therneau e Grambsch, 2000). Essa abordagem é baseada em uma modificação da função de verossimilhança parcial de Cox, apresentada em (2.5.3), de modo que tanto os coeficientes de regressão, quanto as fragilidades, são incluídas e otimizadas sobre $\boldsymbol{\beta}$ e ζ .

Formalmente, a função de verossimilhança é descrita como um produto em que o primeiro termo é a função de verossimilhança parcial, incluindo as fragilidades como parâmetros, e o segundo termo é uma penalidade introduzida para evitar diferenças grandes entre fragilidades para os diferentes grupos. O logaritmo da função de verossimilhança parcial penalizada é, desse modo, expresso por:

$$PPL(\boldsymbol{\beta}, \zeta, \theta) = log(L(\boldsymbol{\beta}, \zeta)) - g(\zeta, \theta),$$

sendo

$$log(L(\boldsymbol{\beta}, \zeta)) = \sum_{i=1}^{n} \delta_i[(\boldsymbol{x}'_i j\boldsymbol{\beta} + \zeta_j) - log(\sum_{k \in R(t_i)} exp\{x'_{kj}\boldsymbol{\beta} + \zeta_{kj}\})],$$

e $g(\zeta, \theta)$ a função penalidade. É frequente o uso do logaritmo de uma densidade como função de penalidade. Se a fragilidade tem, por exemplo, distribuição gama com média 1 e variância $\theta = \vartheta$, o logaritmo da função densidade de $w = exp\{\zeta\}$ pode ser escrito como

$$log(f(w)) = log[(1/\vartheta) - 1]log(w) - (1/\vartheta)w + (1/\vartheta)log(1/\vartheta) - log(\Gamma(1/\vartheta))$$

e, sendo assim, o logaritmo da densidade de ζ é $\frac{(\zeta - exp\{\zeta\})}{\theta}$ mais uma função de θ, o que resulta no logaritmo da função de verossimilhança parcial penalizada:

$$PPL(\boldsymbol{\beta}, \zeta, \theta) = log(L(\boldsymbol{\beta}, \zeta)) - \frac{1}{\theta}\sum_{j=1}^{m}(\zeta_j - exp\{\zeta_j\}).$$

Na prática, o procedimento começa tomando valores iniciais iguais a 1 para as fragilidades. Um procedimento iterativo é, então, inicializado tratando as fragilidades como parâmetros fixos

e conhecidos no primeiro passo do processo de otimização da função de verossimilhança parcial. No segundo passo as fragilidades são avaliadas como médias condicionais, dado suas observações. Este procedimento é repetido até convergência ser obtida. No pacote R o modelo de fragilidade compartilhado (2.6.3) é ajustado por meio desse procedimento de estimação.

2.7 Modelo Exponencial Por Partes

O modelo exponencial por partes (MEP) é bastante utilizado em análise de sobrevivência. Segundo Ibrahim (2001), grande parte desta popularidade se deve ao fato deste modelo ser capaz de acomodar funções de risco com diversas formas, tornando o modelo bastante flexível. Outra vantagem do MEP é a possibilidade de se trabalhar com este modelo tanto na versão paramétrica quanto na versão não-paramétrica. O MEP é caracterizado pela aproximação da função risco por segmentos de retas cujos comprimentos são determinados por uma partição do eixo do tempo em intervalos, dentro dos quais a função taxa de falha é considerada constante. Nesta seção é apresentada uma introdução ao MEP.

Seja T uma variável aleatória não-negativa representando o tempo até a ocorrência da falha. Considere uma partição finita e arbitrária $\{\tau_1, ..., \tau_k\}$ de $\mathbb{R}^+$, tal que $0 = \tau_0 < \tau_1 < \cdots < \tau_k < \infty$, com $\tau_k > t$, para algum t observado, com $t > 0$, e admita que tal partição divida o eixo do tempo $\mathbb{R}^+$ em k intervalos disjuntos, denotados por $I_1 = (\tau_0, \tau_1]$, $I_2 = (\tau_1, \tau_2]$, $\cdots$, $I_k = (\tau_{k-1}, \tau_k]$.

Como apresentado no primeiro parágrafo desta seção, o MEP é caracterizado pela aproximação da função risco, $\lambda(t)$, através de segmentos de retas definidos pelos intervalos determinados pela partição $\{\tau_1, ..., \tau_k\}$, isto é, assume-se que, em cada intervalo $I_j = (\tau_{k-1}, \tau_j]$, $j = 1, \cdots, k$, a função risco é constante, e denotada por $\lambda(t) = \theta_j$, $\theta_j > 0$, $\forall t \in I_j$. Consequentemente, a função taxa de falha acumulada, $\Lambda(t)$, associada ao $j-$*ésimo* intervalo, $I_j = (\tau_{j-1}, \tau_j]$, é dada pela soma das áreas dos retângulos, cujas bases são determinadas pelos intervalos definidos pela partição $\{\tau_1, ..., \tau_k\}$, e com alturas dadas pela função taxa de falha, $\lambda(t)$, ou seja,

$$\Lambda(t) = \sum_{r=0}^{j-1} \theta_r(\tau_r - \tau_{r-1}) + \theta_j(t - \tau_{j-1}), \tag{2.7.1}$$

para $t \in I_j$. Logo, a função de sobrevivência é dada por:

$$S(t|\theta_1,\cdots,\theta_k)=\begin{cases} exp\{-\theta_1(t)\}, \text{se } t\in I_1; \\ exp\left\{-\left[\sum_{r=1}^{j-1}\theta_r(\tau_r-\tau_{r-1})+\theta_j(t-\tau_{j-1})\right]\right\}, \text{se } t\in I_j, j>1, \end{cases}$$

com $\theta_j>0, \forall j=1,\cdots,k$.

Consequentemente, a função densidade de probabilidade da distribuição exponencial por partes é dada por:

$$f(t|\theta_1,\cdots,\theta_k)=\begin{cases} \theta_1 exp\{-\theta_1(t)\}, \text{se } t\in I_1; \\ \theta_j exp\left\{-\left[\sum_{r=1}^{j-1}\theta_r(\tau_r-\tau_{r-1})+\theta_j(t-\tau_{j-1})\right]\right\}, \text{se } t\in I_j, j>1, \end{cases}$$

com $\theta_j>0, \forall j=1,\cdots,k$.

Considere que uma observação seja distribuída com o vetor $(T,\delta,\boldsymbol{X})$ em que T é uma variável absolutamente contínua e representa o tempo de acompanhamento, δ é um indicador de falha ou censura, e $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)$ é um vetor com p covariáveis. Assuma que os tempos de falha não dependam do mecanismo de censura e que as observações são independentes. Considere uma partição arbitrária do tempo, $\{\tau_1,...,\tau_k\}$, que divida o tempo em k intervalos, de tal forma que haja pelo menos um tempo de falha em cada intervalo. Denote por n_j o número de observações associadas ao j-ésimo intervalo e assuma que a distribuição dos tempos de falha possa ser aproximada pela distribuição exponencial por partes definida pela partição $\{\tau_1,...,\tau_k\}$. Então, a função de verossimilhança associada ao $j-$ é*simo* intervalo, I_j, é dada por:

$$L(\theta_1,\cdots,\theta_k,I_j;T)=\prod_{l=1}^{n_j}\lambda(t|\theta_j)^{\delta_l}S(t|\theta_1,\cdots,\theta_k)=\prod_{l=1}^{n_j}\theta_j^{\delta_l}exp\left\{-\left[\sum_{r=1}^{j-1}\theta_r(\tau_r-\tau_{r-1})+\theta_j(t-\tau_{j-1})\right]\right\}.$$

Logo, a função de verossimilhança associada ao vetor de observações $(T,\delta,\boldsymbol{X})$, é dada pelo produto de verossimilhanças sobre os diferentes intervalos determinados pela partição $\{\tau_1,...,\tau_k\}$, isto é,

$$L(\theta_1,\cdots,\theta_k,I_j;T)=\prod_{j=1}^{k}\prod_{l=1}^{n_j}\theta_j^{\delta_l}exp\left\{-\sum_{r=1}^{j-1}\theta_r(\tau_r-\tau_{r-1})-\theta_j(t_l-\tau_{j-1})\right\}. \qquad (2.7.2)$$

Capítulo 3

Bootstrap e Validação Cruzada

Neste capítulo é apresentada uma introdução aos métodos de *Bootstrap* e Validação Cruzada, pois eles são usados na construção das árvores de regressão para dados censurados apresentadas no capítulo 4.

3.1 Reamostragem *Bootsrap*

Reamostragem consiste em sortear com reposição dados pertencentes a uma amostra retirada anteriormente, de modo a formar uma nova amostra. Técnicas de reamostragem são úteis em especial quando o cálculo de estimadores por métodos analíticos é complicado.

Existem diversas técnicas de reamostragem que visam estimar parâmetros de uma distribuição de interesse. Uma vantagem em utilizar *Bootstrap* (Efron (1993)) é a generalidade com que pode ser aplicada, pois requer que menos suposições sejam feitas. Outras vantagens são que geralmente fornece respostas mais precisas, além de favorecer o entendimento. Muitas vezes a distribuição de probabilidade da estatística de interesse é desconhecida. Nesse caso o *Bootstrap* é muito útil, pois é uma técnica que não exige diferentes fórmulas para cada problema e pode ser utilizada em casos gerais, não dependendo da distribuição original da estatística do parâmetro estudado. Duas das desvantagens desta metodologia são a dependência de uma amostra representativa e a variabilidade devida a replicações finitas (Monte Carlo).

A técnica *Bootstrap* não paramétrica utiliza reamostragem baseada nos dados da amostra mes-

tre, uma vez que a distribuição de probabilidade da estatística do parâmetro a ser estimado é desconhecida. Através desta técnica é possível obter a distribuição amostral de uma estatística a partir da amostra original.

A técnica *Bootstrap* consiste em colher uma amostra de tamanho n, a amostra mestre. Essa amostra deve ser coletada de maneira planejada, uma vez que se for mal selecionada e não representar bem a população, a técnica de *Bootstrap* não levará a resultados confiáveis. Para que a aplicação da técnica resulte em valores confiáveis devem ser feitas, a partir da amostra mestre, centenas de reamostras do mesmo tamanho n. É importante que a reamostragem seja realizada com reposição sempre selecionando os valores de forma aleatória.

Para a geração destas reamostras as técnicas computacionais são de grande utilidade, pois sem estas, o tempo para que fossem feitas todas as reamostras desejadas de forma manual seria excessivamente grande. Uma vez geradas as reamostras, deve-se calcular para cada uma delas a estatística solicitada no problema. Essa técnica não altera nenhum valor da amostra mestre, ela apenas trabalha na análise da combinação dos valores iniciais com a finalidade de se obter as conclusões desejadas.

3.2 Validação Cruzada

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta técnica é amplamente empregada em problemas onde o objetivo da modelagem é a predição. Busca-se então estimar o quão acurado é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

O conceito central das técnicas de validação cruzada é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, utilizam-se alguns destes subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) é empregado na validação do modelo.

Diversas formas de realizar o particionamento dos dados foram sugeridas, sendo as três mais utilizadas: o método *holdout*, o *V-fold* e o *leave-one-out*.

Para todos os métodos de particionamento citados, a acurácia final do modelo estimado é obtida

por:

$$Ac_f = \frac{1}{v}\sum_{i=1}^{v}(y_i - \hat{y}_i) \tag{3.2.1}$$

em que v é o número de dados de validação e $(y_i - \hat{y}_i)$ é o resíduo dado pela diferença entre o valor real da saída i e o valor predito. Com isso, é possível inferir de forma quantitativa a capacidade de generalização do modelo.

Método *V-fold*

O método de validação cruzada denominado *v-fold* consiste em dividir o conjunto total de dados em V subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste, os V-1 restantes são utilizados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é realizado V vezes alternando de forma circular o subconjunto de teste. A Figura 3.1 mostra o esquema realizado pelo 3-*fold*. Suponha que o conjunto de dados tenha duas variáveis, X e Y. Esse conjunto de dados é divido em três outros conjuntos mutuamente exclusivos, de tamanhos mais ou menos iguais. Em cada iteração do processo, dois dos conjuntos são usados para estimação e o terceiro conjunto é usado para validação. Após a validação dos três modelos é possível calcular a acurácia sobre os erros encontrados através da equação (3.2.1).

A vantagem da validação cruzada *v-fold* é que a ordem de divisão dos dados não é importante. Cada observação fica na amostra de treinamento exatamente uma vez e na amostra teste $V - 1$ vezes. A variância da estimativa diminui com o aumento de V. A desvantagem desse método é que o algoritmo de treinamento precisa ser rodado V vezes, ou seja, demora V vezes o número de iterações para fazer uma avaliação.

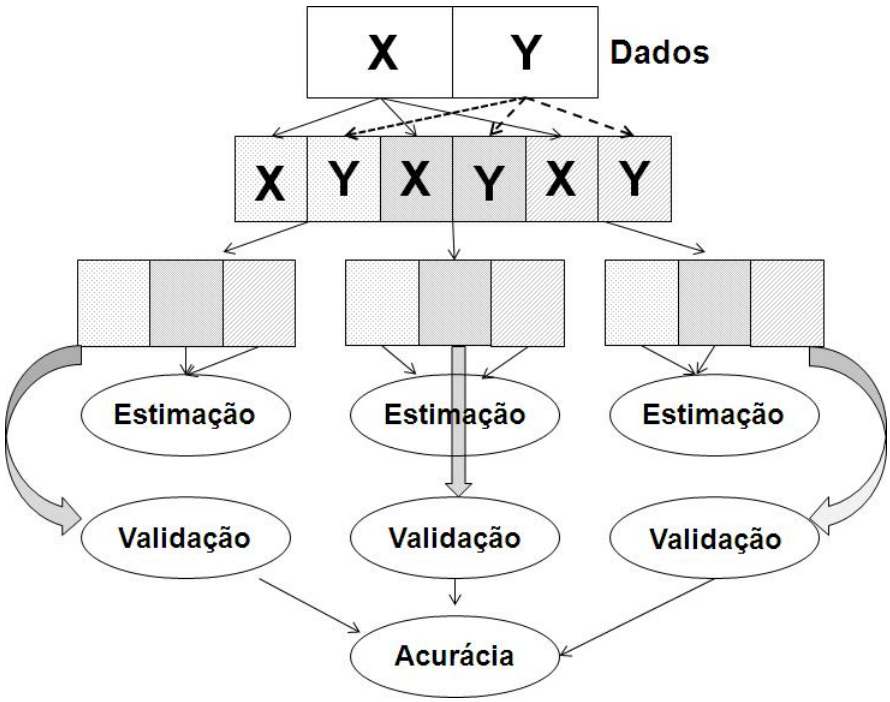


Figura 3.1: Exemplo do esquema de particionamento e execução do método *V-fold* com V=3 (reproduzido de Wikipédia)

Capítulo 4

Árvores de Classificação e Regressão

A utilização de técnicas de segmentação ou de aproximação com recurso de árvores foi motivada pela necessidade de lidar com problemas complexos (envolvendo, por exemplo, dados de dimensão elevada). Esta técnica teve início nas ciências sociais no trabalho de Morgan e Sonquist (1963). Mais tarde Hunt e Stone (1966), Messenger e Mandell (1972) e Morgan e Messenger (1973) desenvolveram este método para problemas de classificação. Contudo, foram as modificações introduzidas por Quinlan em 1979, 1983 e 1986 (Quinlan (1986)) e os trabalhos de Breiman et al. (1984) que mais contribuíram para a grande popularidade da utilização de árvores binárias em problemas de classificação. A utilização de árvores em problemas de regressão iniciou-se nos trabalhos de Morgan e Sonquist (1963) com o seu programa AID (*Automatic Interaction Detection*). Generalizações do mesmo são descritas em Sonquist (1970), Sonquist, Baker e Morgan (1973) e em Van Eck (1980). Breiman et al. (1984) estenderam ainda mais estas técnicas dando origem ao programa CART (*Classification and Regression Trees*) hoje implementado em vários programas estatísticos, como o R, por exemplo.

As árvores apresentam, além do seu poder preditivo, um forte poder descritivo, o qual permite compreender quais as variáveis que originam o fenômeno em estudo, e o modo como estão relacionadas nesse fenômeno. A utilização e interpretação simples das árvores são outros dos atrativos da utilização das mesmas.

As árvores binárias são construídas de acordo com regras de divisão baseadas nas variáveis

preditivas do domínio em estudo. O domínio é particionado recursivamente de forma binária, com o objetivo de aumentar a homogeneidade dentro dos nós, a qual é determinada pela variável resposta do problema. Quando o processo de partição termina, a cada um dos nós terminais é associada uma classe nos problemas de classificação, ou um valor constante real nos problemas de regressão. Assim, os ingredientes principais da construção de uma árvore resumem-se nos seguintes pontos:

1. Determinação de todas as divisões possíveis de um nó para cada variável do espaço de predição (usualmente as divisões são determinadas por questões binárias);

2. Seleção da melhor divisão de todas;

3. Determinação de quando se deve considerar um nó como terminal;

4. Atribuição de um valor resposta a cada nó terminal.

Após a construção da árvore inicial, uma árvore grande e bastante complexa, é realizado o procedimento de poda dessa árvore. Com a ajuda de uma medida de poda, que depende do método usado na construção da árvore, nós que, segundo a medida de poda, não são importantes, são retirados da árvore.

4.1 Árvores de Decisão

A árvore de classificação surgiu como uma técnica alternativa à regressão logística múltipla, para o desenvolvimento de modelos de predição de enfermidades e seus desfechos. Árvore de Classificação e Regressão (CART) é um método de classificação que usa dados históricos para construir a árvore de decisão. As árvores de decisão são então usadas para classificar um novo dado. É uma ferramenta útil em várias áreas do conhecimento, como por exemplo, em medicina para facilitar diagnósticos e prognósticos em contextos clínicos e na área financeira para classificação de bons e maus pagadores segundo perfil dos clientes.

Árvores de decisão binárias são representadas por um conjunto de questões que divide a amostra em partes menores. As perguntas são do tipo sim/não. Uma questão possível pode ser: “A idade

é maior que 50?" ou "O sexo é masculino?". O algoritmo CART irá avaliar todas as variáveis possíveis e todos os valores possíveis a fim de encontrar a melhor divisão: a questão que divide os dados em duas partes com máxima homogeneidade dentro das partes e heterogeneidade entre as partes. O processo é então repetido para cada um dos fragmentos resultantes.

Seja $\boldsymbol{x}$ o vetor de p covariáveis observadas e $\boldsymbol{y}$ o vetor de resposta para n indivíduos. Para o i-ésimo indivíduo temos: $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ e y_i com $i = 1, \cdots, n$. O objetivo é modelar a distribuição de probabilidade, $P(\boldsymbol{y}|\boldsymbol{x})$, ou alguma função dessa distribuição condicional. O vetor $\boldsymbol{x}$ pode ser de variáveis categorizadas, contínuas ou uma mistura dos dois tipos de variáveis. É admitido que contenha valores faltantes também. A variável resposta, $\boldsymbol{y}$, pode ser contínua (com ou sem censura) ou categórica. A Figura 4.1 é um exemplo simples de árvore de decisão, usada pelo Centro Médico de San Diego para classificação de seus pacientes em diferentes níveis de risco.

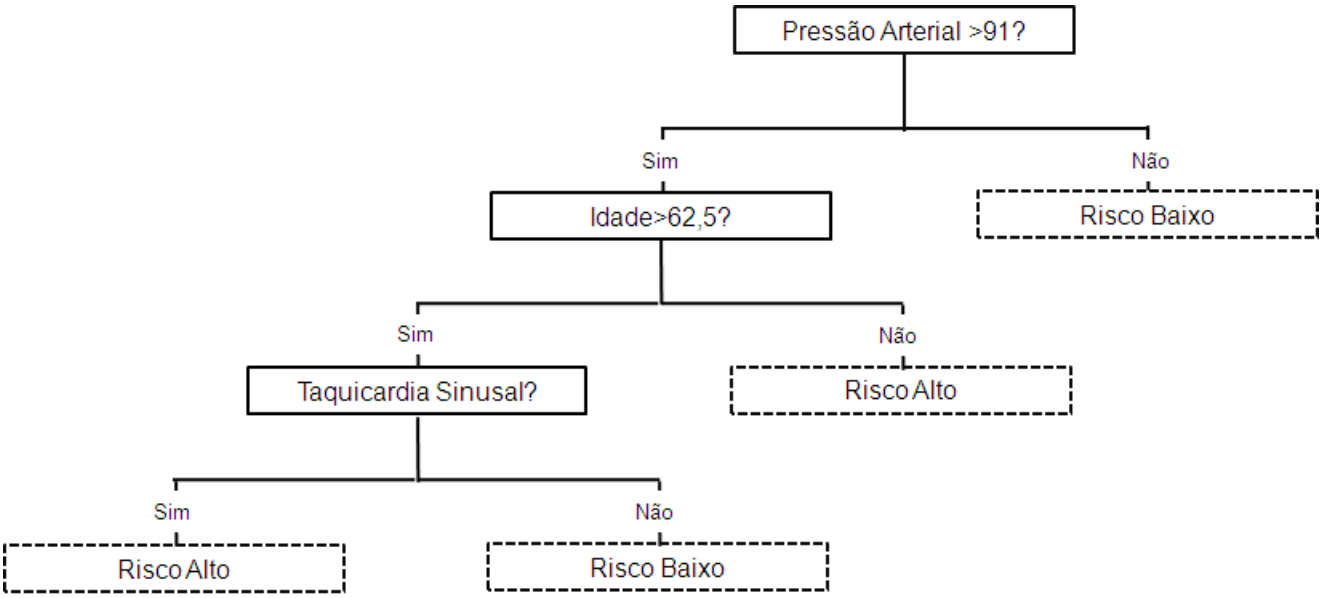


Figura 4.1: Árvore de Classificação do Centro Médico de San Diego (reproduzida de Breiman et al. (1984)).

A árvore apresentada na Figura 4.1 tem 4 carreiras de nós. Em geral, o número de carreiras varia de caso para caso. A primeira carreira sempre contém apenas um nó, que é chamado de nó raiz. Nesta figura há 2 nós internos (linhas contínuas) e 4 nós terminais (linhas pontilhadas). O nó raiz e os nós internos são conectados a outros dois nós na próxima carreira, que são chamados

de nó-filho da esquerda e nó-filho da direita. Nós terminais não tem filhos.

Uma das vantagens da metodologia CART é sua robustez a *outliers*. Usualmente o algoritmo de divisão isolará *outliers* em um nó individual, que quase certamente será podado posteriormente. Uma propriedade prática importante da metodologia CART é que a estrutura de suas árvores de classificação e regressão é invariante com respeito a transformações monótonas das variáveis independentes. Qualquer variável pode ser substituída pelo seu logaritmo ou raiz quadrada, por exemplo, que a estrutura da árvore não será modificada.

Como desvantagem do uso das árvores de decisão pode-se citar que essas árvores podem ser instáveis. Pequenas modificações na amostra de aprendizagem podem levar a grandes mudanças na árvore, como aumento ou redução da complexidade da mesma e mudanças nas variáveis e valores escolhidos para divisão.

As árvores de regressão abordadas nesse trabalho são focadas em variável resposta que consiste em contagem de tempo com censura.

O custo-complexidade de uma árvore A é definido em Breiman et al. (1984) por

$$R_\alpha(A) = \sum_{h \in \tilde{A}} R(h) + \alpha|\tilde{A}|, \tag{4.1.1}$$

para um parâmetro de complexidade não negativo α, em que $R(h)$ é a heterogeneidade dentro do nó h e sua fórmula depende do critério de divisão escolhido, considerando o modelo de risco relativo, e $|\tilde{A}|$ é o número de nós terminais de A, ou seja a complexidade de A.

A medida de custo-complexidade controla o tamanho ou complexidade da árvore, e como a árvore se ajusta aos dados. Se o parâmetro de complexidade α é grande, a árvore que minimiza o custo-complexidade é pequena, e conforme α decresce, a árvore que minimiza o custo-complexidade aumenta em tamanho.

4.2 Árvores de Regressão para Dados Censurados

Como dito anteriormente, a construção da árvore tem dois passos principais, o primeiro é definir um critério de divisão, que divide cada nó em outros dois, o outro é escolher o tamanho

certo da árvore para uso subsequente. Muitos critérios foram propostos na literatura para dados censurados, eles diferem principalmente na forma de declarar quais nós-filhos são desejáveis.

A seguir são apresentadas três diferentes metodologias para construção de árvore de decisão para dados censurados.

4.2.1 Regra de Gordon e Olshen

A primeira proposta de árvore de regressão para dados censurados foi feita por Gordon e Olshen (1985). A ideia é que quando um nó é dividido em dois, pode-se calcular as curvas de sobrevivência de Kaplan-Meier separadamente para cada nó resultante da divisão. Uma divisão desejável pode ser aquela que resulta em duas funções de sobrevivência muito diferentes entre os nós-filhos. Nessa proposta é usada a chamada métrica Lp de Wasserstein, $d_p(\cdot,\cdot)$ como medida de discrepância entre as duas funções de sobrevivência. Especificamente, para $p = 1$, a distância de Wasserstein, $d_1(S_L, S_R)$ entre as duas curvas de Kaplan-Meier, S_L e S_R é a área sombreada na Figura 4.2.

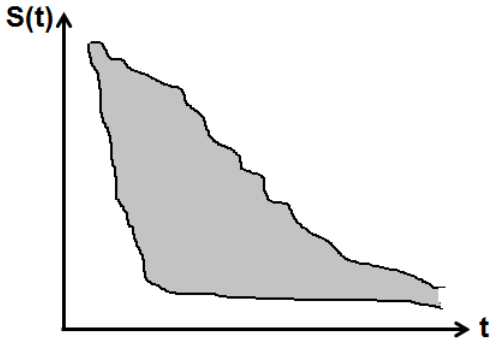


Figura 4.2: Distância L^1 de Wassertein entre duas curvas de Kaplan Meier

Sejam F_L e F_R as funções de distribuição impróprias das curvas de sobrevivência de Kaplan-Meier S_L e S_R, onde $F_L(y) = 1 - S_L(y)$, $F_R(y) = 1 - S_R(y)$ e $lim_{y\to\infty}F_L(y) = m_1 \leq 1$ e $lim_{y\to\infty}F_R(y) = m_2 \leq 1$. Seja m o $min(m_1, m_2)$. Então, d_1 é definido como

$$d_1 = \int_0^m |F_L^{-1}(u) - F_R^{-1}(u)|du,$$

onde $F_i^{-1}(u) = min(y : F_i(y) \geq u)$, $i = 1, 2$.

Uma divisão ótima é escolhida de tal forma que maximize a distância $d_1(S_L, S_R)$. Aqui, S_L e S_R são, respectivamente, as curvas de Kaplan-Meier para os nós-filhos da esquerda e da direita.

Para podar uma árvore inicial, A, Gordon e Olshen (1985) sugeriram um custo-complexidade da árvore como segue. Considere um nó terminal, h pertencente a A^*. Deve-se então:

1. Estimar a curva de sobrevivência por K-M denotada por S_h.

2. Encontrar o ξ_h mais próximo de S_h em termos de $d_1(S_h, \xi_h)$; aqui ξ_h precisa ser escolhido a partir de emenda de funções de sobrevivência constantes que tenha no mínimo um ponto de descontinuidade, ou seja, ξ_h, tem pelo menos duas partes constantes. Então, defina o custo dentro do nó, $R(h)$, como $d_1(S_h, \xi_h)$. Isso pode ser visualizado como o desvio dos tempos de sobrevivência sobre sua mediana. Finalmente, aplicando a fórmula (4.1.1) , tem-se o custo-complexidade da árvore.

4.2.2 Uso do teste *Logrank*

Em análise de sobrevivência, o teste *logrank* é uma abordagem popular para testar a significância das diferenças entre os tempos de sobrevivência de dois grupos. Motivados por este fato, Ciampi et al. (1986) e Segal (1988) sugeriram um método de divisão dos nós que resulta no maior valor da estatística do teste *logrank*, definido na Seção 1.3.

A estatística *logrank* (ou qualquer teste similar para duas amostras) é uma medida da diferença entre nós. Portanto, com esta abordagem, uma medida de custo para cada nó não é prontamente avaliável para usar na poda. Segal (1988) recomendou um procedimento prático: para cada nó interno (incluindo o nó raiz) de uma árvore inicial, valores são atribuídos ao nó que é igual ao máximo da estatística *logrank* sobre todas as divisões partindo do nó interno de interesse. Então, os valores da estatística para todos os nós internos são colocados em um gráfico em ordem decrescente e decide-se uma linha de corte a partir do gráfico. Se um nó interno corresponde a um valor menor que o ponto de corte, então poda-se todos os seus descendentes.

Neste caso, a qualidade-de-divisão é usada como um substituto para o custo-complexidade na

poda da árvore. Seja $G(h)$ o valor da estatística *logrank* do nó h. Então a medida de qualidade-de-divisão é

$$G_\alpha(A) = \sum_{h \notin \tilde{A}} G(h) - \alpha|A - \tilde{A}|, \tag{4.2.1}$$

em que $\tilde{A}$ é o conjunto de nós terminais de A e $|\tilde{A}|$ é o número de nós terminais. Note que o somatório em (4.2.1) é sobre o conjunto de nós internos. O sinal negativo é um reflexo do fato que G deve ser maximizada considerando que o custo-complexidade $R_\alpha(A)$, dado em (4.1.1) , é minimizado. É recomendado escolher um α entre 2 e 4 e usar técnicas de *bootstrap* para reduzir o valor de $G_\alpha(A)$.

4.2.3 Risco Relativo

Métodos baseados em árvore podem ser uma alternativa útil ao modelo clássico de Cox (1972) para exploração de dados de sobrevivência. O método de árvore de regressão desta seção, apresentado por LeBlanc, M. e Crowley, J. (1992), adota o modelo de riscos proporcionais que especifica a função de riscos a seguir, no tempo t, para um indivíduo com vetor de covariáveis $\boldsymbol{x}$:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)g(\boldsymbol{x}),$$

em que $g(\boldsymbol{x}) \geq 0$ e $\lambda_0(t)$ é a função risco basal. Tradicionalmente $g(\boldsymbol{x}) \geq 0$ é uma função log-linear de um vetor de parâmetros. Aqui, será apresentado um método para obter estrutura de árvores que representam a função de risco relativo, $g(\boldsymbol{x}) \geq 0$.

Considerando que métodos de partição recursiva envolvem avaliação de um grande número de divisões, cálculos iterativos em cada ponto de divisão tipicamente não são computacionalmente viáveis. Portanto, o modelo aqui apresentado constrói e poda a árvore usando somente o primeiro passo do procedimento de estimação da verossimilhança completa para modelo de riscos proporcionais. Depois que a árvore é escolhida, a estimativa da verossimilhança completa é obtida iterativamente. O processo de construção da árvore utiliza validação cruzada para estimar o erro de predição para uma sequência de modelos. A árvore que minimiza ou chega perto de minimizar o erro de predição estimado é então escolhida.

A verossimilhança: Assume-se que os dados incluem medidas de tempo e covariáveis que podem ser associadas com o tempo de falha. Uma observação será distribuída como o vetor $(T, \delta, \boldsymbol{X})$ em que T é uma variável absolutamente contínua e representa o tempo de acompanhamento, δ é um indicador de falha ou censura, e $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ é um vetor com p covariáveis. Suponha que J seja o verdadeiro tempo de falha com distribuição acumulada H e C seja o verdadeiro tempo de censura com distribuição acumulada L. Seja $\delta = \mathbb{I}(J \leq C)$ onde $\mathbb{I}(\cdot)$ é a função indicadora do conjunto $(\cdot)$, e $T = min(J, C)$ é o tempo observado. Assuma também que J e C são independentes dado $\boldsymbol{X}$. A amostra de aprendizagem consiste do conjunto de vetores independentes e identicamente distribuídos $(t_i, \delta_i, \boldsymbol{x}_i) : i = 1, 2, \cdots, n$.

Tipicamente inferências para o modelo de riscos proporcionais são baseadas na verossimilhança parcial. No entanto, se assumir-se que o risco acumulado basal é conhecido, estimação e modelagem baseados na verossimilhança completa são desejáveis. A verossimilhança completa da amostra de aprendizagem para a árvore A pode ser expressa como

$$L = \prod_{h \in \tilde{A}} \prod_{i \in O_h} \lambda_h(t_i)^{\delta_i} exp\{-\Lambda_h(t_i)\},$$

em que $\tilde{A}$ é o conjunto de nós terminais; O_h é o conjunto das observações que estão no nó h; (t_i, δ_i) é o vetor do tempo observado e indicador de falha para o indivíduo i; e $\lambda_h(t_i)$ e $\Lambda_h(t_i)$ são as funções risco e risco acumulado para o nó h, respectivamente. Assuma que o modelo de riscos proporcionais

$$\lambda_h(t) = \lambda_0(t)\theta_h$$

seja adequado, em que θ_h é um parâmetro não-negativo e $\lambda_0(t)$ é a função risco basal. Segue que a verossimilhança para os dados dada a árvore A é

$$L = \prod_{h \in \tilde{A}} \prod_{i \in O_h} (\lambda_0(t_i)\theta_h)^{\delta_i} exp\{-\Lambda_0(t_i)\theta_h\},$$

em que $\Lambda_0(t)$ é o risco basal acumulado. Dado o risco basal acumulado, o estimador de máxima verossimilhança de $\{\theta_h : h \in \tilde{A}\}$ é

$$\hat{\theta}_h = \frac{\sum_{i \in O_h} \delta_i}{\sum_{i \in O_h} \Lambda_0(t_i)}.$$

Na prática, o risco acumulado basal não é conhecido. No entanto, um estimador natural do risco acumulado dada a estimativa $\hat{\theta}_h$ é,

$$\hat{\Lambda}_0(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{h \in \tilde{A}} \sum_{i:t_i \geq t, i \in O_h} \hat{\theta}_h},$$

que é o estimador de Breslow (1972). Um procedimento alternativo de estimação pode ser usado para estimar Λ_0 e $\{\theta_h : h \in \tilde{A}\}$. Primeiramente, o risco acumulado de Breslow para a iteração j,

$$\hat{\Lambda}_0^j = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{h \in \tilde{A}} \sum_{i:t_i \geq t, i \in O_h} \hat{\theta}_h^j}, \tag{4.2.2}$$

é calculado usando as estimativas atuais, $\hat{\theta}_h^j$ de θ_h. Depois, a estimativa $\hat{\theta}_h^{j+1}$ de θ_h,

$$\hat{\theta}_h^{j+1} = \frac{\sum_{i \in O_h} \delta_i}{\sum_{i \in O_h} \hat{\Lambda}_0^j(t_i)}, \tag{4.2.3}$$

é calculada usando a estimativa atual $\hat{\Lambda}_0^j(t)$. Esses dois passos são repetidos até a convergência.

Apenas a primeira iteração será utilizada no procedimento de partição recursiva para construir a árvore e selecionar o tamanho da mesma. O estimador de Breslow avaliado em $\{\hat{\theta}_h^1 = 1 : h \in \tilde{A}\}$, que é o estimador de risco acumulado de Nelson (1969), é usado. A estimativa em um passo de θ_h é

$$\hat{\theta}_h^1 = \frac{\sum_{i \in O_h} \delta_i}{\sum_{i \in O_h} \hat{\Lambda}_0^1(t_i)},$$

que pode ser interpretada como o número de falhas observadas dividido pelo número esperado de falhas no nó h. Dessa forma, mesmo o procedimento em um passo dá quantidades interpretáveis para cada nó.

A *deviance* da verossimilhança completa mede quão bem a árvore se ajusta aos dados. A *deviance* para o nó h é dada por

$$D(h) = 2[L_h(saturado) - L_h(\tilde{\theta}_h)],$$

em que $L_h(saturado)$ é a log-verossimilhança para o modelo saturado que permite um parâmetro para cada observação, e $L_h(\tilde{\theta}_h)$ é a log-verossimilhança maximizada quando $\Lambda_0(t)$ é conhecido. A *deviance* residual para uma observação i no nó h é

$$d_i = 2\left[\delta_i log\left(\frac{\delta_i}{\Lambda_0(t_i)\hat{\theta}_h}\right) - (\delta_i - \Lambda_0(t_i)\hat{\theta}_h)\right],$$

que é equivalente à *deviance* residual baseada no modelo de Poisson com resposta δ_i e média $\tilde{\mu_i} = \Lambda_0(t_i)\tilde{\theta}_h$. A ligação entre a verossimilhança completa do modelo de riscos proporcionais e a verossimilhança do modelo de Poisson foi usada por vários autores, incluindo Clayton (1983) e Clayton e Cuzick (1985) e será apresentada na seção 4.2.4.

O procedimento de partição recursiva calcula a *deviance* residual utilizando $\hat{\Lambda}_0^1$ e $\hat{\theta}_h^1$.

O Algoritmo de Construção da Árvore: Uma árvore grande é construída para evitar a falta de estruturas importantes. O algoritmo de poda custo-complexidade do CART obtém uma sequência de sub-árvores ótimas (sub-árvores são obtidas removendo os galhos da árvore). Finalmente, uma estimativa da *deviance*-um-passo esperada é calculada para cada uma das sub-árvores podadas por validação cruzada e a árvore que minimiza a *deviance* estimada é escolhida como sendo a melhor.

Algoritmos de partição recursiva dividem o espaço das covariáveis baseados em uma regra que minimiza algumas medidas de ganho. Esse algoritmo irá dividir os dados e o espaço das covariáveis em regiões que maximizam a redução da *deviance* realizada na divisão. Muitos tipos de partições podem ser considerados; no entanto, serão consideradas somente divisões sobre uma única covariável. Todas as divisões possíveis para cada uma das covariáveis são avaliadas e a variável e o ponto de divisão que resultam na maior redução da *deviance*-um-passo são escolhidos. Usualmente, existe uma regra sobre o menor tamanho do nó. Se nós com poucas observações são permitidos o algoritmo frequentemente divide a árvore em grupos de observações pequenos que não são bem validados. Seja N o número total de observações na amostra de aprendizagem. O ganho para a divisão s do nó h em dois nós filhos $l(h)$ e $r(h)$ é dado por

$$R(s,h) = R(h) - [R(l(h)) + R(r(h))],$$

em que

$$R(h) = \frac{1}{N}\sum_{i\in O_h}\left[\delta_i log\left(\frac{\delta_i}{\hat{\Lambda}_0^1(t_i)\hat{\theta}_h}\right) - (\delta_i - \hat{\Lambda}_0^1 t(i)\hat{\theta}_h)\right]. \tag{4.2.4}$$

Divisões binárias continuam até que uma árvore grande seja construída e restarem poucas observações em cada nó.

Poda e Seleção da Árvore: As árvores escolhidas são aquelas que minimizam a medida de custo-complexidade (4.1.1) assim como é feito na metodologia CART; estas árvores são chamadas

de sub-árvores otimamente podadas. Na próxima definição, o símbolo "$\leq$" significa "é sub-árvore de".

Definição 4.1

A_α é uma sub-árvore otimamente podada de A para o parâmetro de complexidade α, se $R_\alpha(A_\alpha) = min_{\{A' \leq A\}} R_\alpha(A')$, e ela é a menor sub-árvore otimamente podada se $A_\alpha \leq A''$ para todas as sub-árvores otimamente podadas A''.

Breiman et al. (1984) mostraram que para a medida de custo-complexidade existe uma única sub-árvore otimamente podada para qualquer parâmetro de complexidade α. Eles também mostraram que conforme α aumenta, a sequência ótima de sub-árvores é uma sequência aninhada de árvores e que existe um algoritmo eficiente para obter a sequência ótima de sub-árvores.

A *deviance*-um-passo esperada para árvores podadas é estimada por validação cruzada *V-fold*. Os dados Ω são divididos em V conjuntos Ω_v ($\Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_V = \Omega$) e subamostras $\Omega_{(v)} = \Omega - \Omega_v$ de tamanhos mais ou menos iguais e árvores A_v são construídas a partir dos subconjuntos $\Omega_{(v)}$. Para qualquer α, uma sub-árvore otimamente podada, $A_v(\alpha)$ e estimativas $\hat{\theta}_h^1(v) : h \in \tilde{A}_v(\alpha)$ são obtidas. Para cada árvore A_v aproximadamente $1/V$ dos dados são excluídos para construir a árvore. A performance da *deviance* residual do modelo gerado com $\Omega_{(v)}$ é avaliada com a amostra Ω_v. A *deviance* residual pós validação cruzada para o indivíduo i que não está na amostra usada para construir a árvore é

$$d_{(-i)}(\delta_i, \hat{\theta}_h^1(v)) = 2\left[\delta_i log\left(\frac{\delta_i}{\hat{\Lambda}_0^1(t_i)\hat{\theta}_h^1(v)}\right) - (\delta_i - \hat{\Lambda}_0^1(t_i)\hat{\theta}_h^1(v))\right],$$

em que $\hat{\Lambda}_0^1(t_i)$ é baseado em Ω. Seja α^* o valor do parâmetro de complexidade α que minimiza a *deviance* média residual pós validação cruzada para árvores $A_v(\alpha)$ sobre V sub-amostras.

Um problema surge no cálculo dos estimadores de validação cruzada para dados censurados. Se todas as observações no nó h são censuradas na sub-amostra usada para construir a árvore, a estimativa de $\hat{\theta}_h$ é 0. Agora, se algumas das observações na amostra de validação for tal que existam nós sem nenhuma censura, a estimativa de validação cruzada da *deviance* esperada é infinita. Considerando que estimativas de riscos iguais a 0 e *deviances* infinitas são não realísticas, um ajuste é necessário. Sugere-se então uma simples solução *ad hoc*: substituir nós com 0 falhas

observadas por 0,5, similar ao que foi sugerido por Davis e Anderson (1989). Então, o estimador de θ_h para validação cruzada é dado por

$$\hat{\theta}_h = \frac{1}{2\sum_{i \in O_h} \hat{\Lambda}_0^1(t_i)}$$

para um nó h que não têm falhas. Depois de escolher uma árvore $A(\alpha^*)$ minimizando a estimativa da *deviance*-um-passo esperada pós validação cruzada, estimadores de máxima verossimilhança do risco relativo entre os nós são obtidos interagindo as equações (4.2.2) e (4.2.3), que não envolvem quaisquer inversão de matrizes. A convergência é rápida.

4.2.4 Ligação entre a verossimilhança completa do modelo de riscos proporcionais e a verossimilhança do modelo de Poisson

O modelo de risco relativo apresentado na Seção 4.2.3 será utilizado no capítulo 5 para ajustar árvores de regressão sem considerar a estrutura de correlação dentro das famílias. Para isso será utilizado o pacote *rpart* do R. Nesse pacote é usado o modelo de Poisson para ajustar árvores de regressão para dados censurados. Aqui será mostrada a ligação entre o modelo de Poisson e o modelo de Risco Relativo apresentado na Seção 4.2.3.

Holford (1980) e Laird e Oliver (1981), mostraram que o modelo de riscos proporcionais por partes, apresentado na seção 2.7, é equivalente ao modelo de regressão de Poisson.

Seja T_i o tempo até a ocorrência do evento de interesse do i-ésimo indivíduo, e δ_i o indicador de falha ou censura que recebe o valor 1 se o indivíduo experimentou o evento e 0 caso contrário. Definem-se, então, medidas análogas para cada intervalo que o indivíduo i atravessa, ou seja, criam-se várias pseudo-observações, uma para cada combinação de indivíduo e intervalo.

Seja T_{ij} o tempo até a ocorrência do i-ésimo indivíduo no j-ésimo intervalo, ou seja, entre τ_{j-1} e τ_j. Se o tempo de ocorrência do evento for maior que o final do intervalo, ou seja, $T_i > \tau_j$, então o tempo de ocorrência no intervalo é igual à largura do intervalo e $T_{ij} = \tau_j - \tau_{j-1}$. Se o indivíduo experimentou o evento ou foi censurado no intervalo, ou seja, $\tau_{j-1} < T_i < \tau_j$, então o tempo do evento no intervalo é $T_{ij} = T_i - \tau_{j-1}$, a diferença entre o tempo total de sobrevivência e o limite inferior do intervalo. Só serão considerados intervalos que foram visitados por pelo menos

um indivíduo, mas obviamente o tempo de sobrevivência em um intervalo poderá ser zero se o indivíduo tiver morrido antes do início do intervalo e $T_i < \tau_{j-1}$.

Para a criação dos indicadores de falha considere que δ_{ij} é igual a 1 se o i-ésimo indivíduo falha no intervalo j e 0 caso contrário. Seja $j_{(i)}$ o indicador do intervalo no qual o indivíduo experimenta o evento ou é censurado. É importante enfatizar que este intervalo irá variar de um indivíduo para outro. Digamos que o indivíduo experimenta o evento ou é censurado no intervalo $j_{(i)}$, então δ_{ij} será zero para todo $j < j_{(i)}$ (isto é, todos os intervalos anteriores) e será igual a δ_i para $j = j_{(i)}$, (isto é, o intervalo no qual o i-ésimo indivíduo foi observado pela última vez).

Então o modelo exponencial por partes pode ser ajustado aos dados tratando os indicadores de falha δ_{ij} como se eles fossem observações independentes de uma Poisson com média $\mu_{ij} = T_{ij}\lambda_{ij}$, onde T_{ij} é o tempo de exposição como definido anteriormente e λ_{ij} é a função risco do i-ésimo indivíduo no intervalo j. Tirando o logaritmo dessa expressão e lembrando que as taxas de falha satisfazem o modelo de riscos proporcionais, obtemos $log(\mu_{ij}) = log(T_{ij}) + log(\lambda_j) + \boldsymbol{x}_i\boldsymbol{\beta}$.

Assim, o modelo exponencial de riscos proporcionais por partes é equivalente ao modelo log-linear de Poisson para as pseudo-observações, uma para cada combinação de indivíduo e intervalo, onde o indicador de falha é a resposta e o log do tempo de exposição entra como um *offset*.

É importante ressaltar que não foi assumido que os δ_{ij} seguem distribuição de Poisson independentes, porque claramente isso não acontece. Se o indivíduo i falha no intervalo $j_{(i)}$, então é claro que ele não falhou nos intervalos anteriores $j < j_{(i)}$, portanto os indicadores não são independentes. Além disso, cada indicador só pode assumir os valores 0 ou 1, então não é possível eles seguirem distribuição de Poisson, que dá probabilidade positiva para valores maiores que 1. O resultado é mais sutil. São as funções de verossimilhança que coincidem.

Teorema 4.1 (Friedman (1982))

Dada a realização de um processo de sobrevivência exponencial por partes, é possível encontrar a realização de um conjunto de observações Poisson independentes que tem a mesma verossimilhança, e portanto levará às mesmas estimativas e testes de hipóteses.

Demonstração: A contribuição do i-ésimo indivíduo à função log-verossimilhança tem a forma $log(L_i) = \delta_i log\lambda_i(T_i) - \Lambda(T_i)$, com $\lambda_i(T)$ a função risco e $\Lambda_i(T)$ a função risco acumulado que aplica-se ao i-ésimo indivíduo no tempo T. Lembre-se que $j_{(i)}$ é o intervalo onde o $i - ésimo$ indivíduo

experimenta o evento ou é censurado.

Sob o modelo exponencial por partes, o primeiro termo da log-verossimilhança pode ser escrito como $\delta_i log\lambda_i(T_i) = \delta_{ij(i)} log\lambda_{ij(i)}$, usando o fato de que o risco é $\lambda_{ij(i)}$ quando T_i está no intervalo $j_{(i)}$ e que o indicador de falha δ_i aplica-se diretamente ao último intervalo visitado pelo indivíduo i, e portanto é igual a $\delta_{j(i)}$.

O risco acumulado no segundo termo é uma integral, e pode ser aproximada por uma soma: $\Lambda_i(T_i) = \int_0^{T_i} \lambda(T)\mathrm{d}T \cong \sum_{j=1}^{j_{(i)}} T_{ij}\lambda_{ij}$, onde T_{ij} é o tempo total do indivíduo i no intervalo j. Para ficar mais claro, note que é preciso integrar o risco de 0 a T_i. Para tanto, essa integral foi dividida em uma soma de integrais, uma para cada intervalo onde o risco é constante. Se um indivíduo sobrevive através de um intervalo, a contribuição à integral será o risco λ_{ij} multiplicado pela largura do intervalo. Se o indivíduo falha ou é censurado no intervalo, a contribuição à integral será o risco λ_{ij} multiplicado pelo tempo decorrido do início do intervalo ao tempo de falha ou censura, que é $T_i - \tau_{j-1}$. Mas essa é precisamente a definição do tempo de exposição T_{ij}.

Uma ligeira falta de simetria nesses resultados é que o risco leva a um termo sobre $\delta_{ij(i)} log\lambda_{ij(i)}$, mas o risco acumulado leva aos termos $j_{(i)}$, um para cada intervalo de $j = 1$ até $j_{(i)}$. Porém sabe-se que $\delta_{ij} = 0$ para todo $j < j_{(i)}$, então pode-se adicionar termos em $\delta_{ij} log(\lambda_{ij})$ para todos os j's anteriores; uma vez que $\delta_{ij} = 0$, eles não terão nenhuma contribuição à log-verossimilhança. Esse truque permite escrever a contribuição do i-ésimo indivíduo à log-verossimilhança como uma soma das $j_{(i)}$ contribuições, uma para cada intervalo visitado pelo indivíduo: $log(L_i) = \sum_{j=1}^{j_{(i)}} \{\delta_{ij} log(\lambda_{ij}) - T_{ij}\lambda_{ij}\}$.

Considerando que a contribuição do indivíduo à log-verossimilhança é a soma de vários termos (portanto a contribuição à verossimilhança é o produto de vários termos), pode-se tratar cada um dos termos como representativos de uma observação independente.

O passo final é identificar a contribuição de cada pseudo-observação, ou seja,

$$log(L_{ij}) = \delta_{ij} log(\lambda_{ij}) - T_{ij}\lambda_{ij}. \tag{4.2.5}$$

Note que essa equação é igual, exceto por uma constante, à verossimilhança que pode ser obtida considerando que δ_{ij} tem distribuição Poisson com média $\mu_{ij} = T_{ij}\lambda_{ij}$. Para ficar claro escreva

a log-verossimilhança da Poisson como $log(L_{ij}) = \delta_{ij}log(\mu_{ij}) - \mu_{ij} = \delta_{ij}log(T_{ij}\lambda_{ij}) - T_{ij}\lambda_{ij}$. Esta expressão é igual à log-verossimilhança em (4.2.5) exceto pelo termo $log(T_{ij})$, porém este termo é uma constante dependente dos dados e não dos parâmetros, então ele pode ser ignorado do ponto de vista de estimação. Isso completa a prova.□

4.3 Árvores de Regressão para Dados Censurados Correlacionados

Uma extensão natural das árvores de sobrevivência apresentadas em 4.2 é considerar dados correlacionados. Suponha que existem g grupos nos dados e que os índices (ik) indicam a observação da k-ésima unidade do i-ésimo grupo, $k = 1, 2, \cdots, K_i$, $i = 1, 2, \cdots, g$. É assumida independência entre os grupos, mas as observações dentro de um mesmo grupo são possivelmente correlacionadas. O objetivo é construir uma árvore levando em consideração a correlação dentro dos grupos. Os modelos Marginal e de Efeito Aleatório (Fragilidade) são as duas principais abordagens para tratar respostas de sobrevivência correlacionadas e ambos foram adaptados para construir árvores de sobrevivência. Su e Fan (2004); Gao, Manatunga e Chen (2004); Fan, Nunn e Su (2009) usaram a abordagem onde a dependência dentro dos grupos é modelada por um termo multiplicativo de efeito aleatório. Mais especificamente, a seguinte formulação da função risco é o ponto inicial de seus modelos:

$$\lambda_{ik}(t) = \lambda_0(t)exp\{\beta x_{ik}\}w_i$$

em que λ_0 é uma função risco basal não especificada e w_i é o termo de fragilidade para o grupo i que segue alguma distribuição específica. A distribuição Gama é assumida nestes artigos. Para definir o critério de divisão, Su e Fan (2004) usaram o método de Máxima Verossimilhança, Fan, Nunn e Su (2009) usaram o teste de Escore e Gao, Manatunga e Chen (2004) usaram o teste de Wald.

Fan et al. (2006) usaram a abordagem marginal em que a estrutura de dependência não é especificada. Mais precisamente, eles consideram o seguinte modelo de Cox para avaliar uma

simples variável de divisão binária $X \in \{0, 1\}$ definida por meio de uma covariável:

$$\lambda_{ik}(t|X_{ik}) = \lambda_0(t)exp\{\beta X_{ik}\}$$

em que λ_0 é uma função de risco basal não especificada. Utilizando um estimador consistente da estrutura de variância da função Escore, eles obtêm um teste Escore para a hipótese nula $H_0 : \beta_0 = 0$ que age como o critério de divisão. Este teste é uma estatística *logrank* robusta para duas amostras e sua metodologia é uma generalização do método de LeBlanc e Crowley (1993). Gao, Manatunga e Chen (2006) usaram uma abordagem similar baseada na distribuição marginal do tempo de sobrevivência, mas assumem uma estrutura de risco proporcional global para a árvore inteira. Ao contrário das abordagens usuais, o conjunto de dados inteiro é usado em cada divisão.

A seguir é apresentado o método proposto por Fan, Nunn e Su (2009).

Suponha que existem g famílias nos dados, com a i-ésima família contendo K_i tempos de falha, ou seja, indivíduos. Seja F_{ik} o tempo de falha para o k-ésimo sujeito da i-ésima família e seja C_{ik} o correspondente tempo de censura. Os dados então consistem de $(T_{ik}, \delta_{ik}, \boldsymbol{X}_{ik}) : i = 1, 2, \cdots, g$ e $k = 1, 2, \cdots, K_i$, onde $T_{ik} = min(F_{ik}, C_{ik})$ é o tempo de seguimento observado; $\delta_{ik} = \mathbb{I}(F_{ik} \leq C_{ik})$ é o indicador de falha, com 1 se $F_{ik} \leq C_{ik}$ e 0 caso contrário; e $\boldsymbol{X}_{ik} \in \mathbb{R}^p$ denota o vetor de covariáveis p-dimensional associado ao k-ésimo sujeito da i-ésima família. Os tempos de falha dentro de uma mesma família são provavelmente correlacionados. É assumido que o vetor de tempos de falha $(F_{i1}, \cdots, F_{iK_i})'$ é independente do vetor tempo de censura $(C_{i1}, \cdots, C_{iK_i})'$ condicional ao vetor de covariáveis $(X_{i1}, \cdots, X_{iK_i})'$ para $i = 1, \cdots, g$. Para construir o modelo de árvore A, foram seguidos os quatro passos do algoritmo CART (Breiman et al., (1984)) já apresentado no início deste capítulo:

1. Determinação de todas as divisões possíveis de um nó para cada variável do espaço de predição (usualmente as divisões são determinadas por questões binárias);

2. Seleção da melhor divisão de todas;

3. Determinação de quando se deve considerar um nó como terminal, ou seja, poda da árvore;

4. Atribuição de um valor resposta a cada nó terminal.

Construindo a árvore inicial: A função risco do tempo de falha F_{ik} para o k-ésimo membro da i-ésima família é formulada pelo seguinte modelo de fragilidade gama

$$\lambda_{ik}(t) = exp\{\beta_0 + \beta_1 x_{ik}\} w_i, \tag{4.3.1}$$

para $k = 1, \cdots, K_i$ e $i = 1, \cdots, g$ em que $exp\{\beta_0\}$ é o risco basal e uma constante desconhecida, β_1 é um parâmetro de regressão desconhecido, e $x_{ik} = \mathbb{I}(X_{ik_j} \leq c)$ com $\mathbb{I}(\cdot)$ a função indicadora. Para um ponto de corte constante c, $X_{ik_j} \leq c$ induz a uma partição binária de acordo com a covariável contínua X_j. Se X_j é categórica, então a forma de $X_j \in B$ é considerada, onde B pode ser qualquer subconjunto de suas categorias. Para simplificação e interpretabilidade as divisões foram restringidas em uma única covariável.

A fragilidade w_i é um termo multiplicativo de efeito aleatório que explicita a dependência dentro das famílias. Ele corresponde a alguma característica não observada compartilhada por todos os membros de uma mesma família. É assumido que, dado a fragilidade w_i, os tempos de falha dentro de uma mesma família são independentes. O termo de fragilidade w_i é assumido seguir uma distribuição Gama com média 1 e variância desconhecida ϑ . Ou seja, $W_i \sim Gama(1/\vartheta, 1/\vartheta)$ $\vartheta \geq 0$ com função densidade

$$Gama(w, \vartheta) = \frac{w^{[1/(\vartheta-1)]} exp\{-w/\vartheta\}}{(1/\vartheta)\vartheta^{1/\vartheta}}.$$

Para dividir os dados, é usado o teste de Escore para avaliar $H_0 : \beta_1 = 0$. A estimação do modelo nulo (ou seja, do modelo (4.3.1) sob H_0) é como segue:

Sob $H_0 : \beta_1 = 0$, o modelo nulo de fragilidade gama é

$$\lambda_{ik}(t) = exp\{\beta_0\} w_i.$$

Defina $E_i = 1/\vartheta + \delta_{i\cdot}$ e $B_i = 1/\vartheta + T_{i\cdot} exp\{\beta_0\}$, em que $\delta_{i\cdot} = \sum_{k=1}^{K_i} \delta_{ik}$, $T_{i\cdot} = \sum_{k=1}^{K_i} T_{ik}$ e $\delta_{\cdot\cdot} = \sum_{i=1}^{n} \delta_{i\cdot}$.

Devido à forma paramétrica do modelo, os parâmetros envolvidos (β_0, ϑ) podem ser estimados

maximizando a seguinte verossimilhança:

$$L_0 = \prod_{i=1}^{g} \int_0^{\infty} \left[Gama(w_i, \vartheta) \prod_{k=1}^{K_i} [\lambda_{ik}(T_{ik})]^{\delta_{ik}} exp\left(- \int_0^{T_{ik}} \lambda_{ik}(u) \mathrm{d}u \right) \right] \mathrm{d}w_i.$$

Após integrar sobre w_i, a função log-verossimilhança para o modelo nulo pode ser escrita como:

$$l_0 = \beta_0 \delta_{..} + \sum_{i=1}^{g} log(\Gamma(E_i)) - g\left[log\left(\Gamma\left(\frac{1}{\vartheta} \right) \right) \right] - \frac{g}{\vartheta}\left(log(\vartheta) \right) - \sum_{i=1}^{g} E_i log(B_i).$$

A otimização da log-verossimilhança marginal é realizada usando o procedimento de Newton-Rapson. As componentes do gradiente são dadas por

$$\frac{\partial l_0}{\partial \beta_0} = \delta_{..} - \sum_{i=1}^{g} \frac{E_i}{B_i} T_{i.} exp\{\beta_0\}$$

$$\frac{\partial l_0}{\partial \vartheta} = \frac{1}{\vartheta^2} \left[- \sum_{i=1}^{g} \frac{\Gamma'(E_i)}{\Gamma(E_i)} + g \frac{\Gamma'(1/\vartheta)}{\Gamma(1/\vartheta)} - g + g log(\vartheta) + \sum_{i=1}^{g} log(B_i) + \sum_{i=1}^{g} \frac{E_i}{B_i} \right].$$

As componentes da matriz de informação observada de Fisher, $\boldsymbol{I}_0$, são dadas por:

$$\boldsymbol{I}_0 = \begin{pmatrix} i_{\beta_0\beta_0} & i_{\beta_0\vartheta} \\ i_{\vartheta\beta_0} & i_{\vartheta\vartheta} \end{pmatrix},$$

em que

$$i_{\beta_0\beta_0} = -\frac{\partial^2 l_0}{\partial \beta_0^2} = \frac{exp\{\beta_0\}}{\vartheta} \sum_{i=1}^{g} \frac{E_i}{B_i^2} T_{i\cdot} \ ,$$

$$i_{\beta_0\vartheta} = -\frac{\partial^2 l_0}{\partial \beta_0 \partial \vartheta} = \frac{exp\{\beta_0\}}{\vartheta^2} \sum_{i=1}^{g} \frac{T_{i.}(E_i - B_i)}{B_i^2} = \boldsymbol{i}_{\vartheta\beta_0}$$

$$i_{\vartheta\vartheta} = -\frac{\partial^2 l_0}{\partial \vartheta^2} = \frac{2g}{\vartheta^3} \frac{\Gamma'(1/\vartheta)}{\Gamma(1/\vartheta)} - \frac{3g}{\vartheta^3} + \frac{2g log(\vartheta)}{\vartheta^3} + \frac{g}{\vartheta^4} \frac{\Gamma''(1/\vartheta)}{\Gamma'(1/\vartheta)} - \frac{g}{\vartheta^4} \left[\frac{\Gamma'(1/\vartheta)}{\Gamma(1/\vartheta)} \right]^2$$
$$+ \sum_{i=1}^{g} \left[\frac{2\Gamma'(E_i)}{\vartheta^3 \Gamma(E_i)} - \frac{2log(B_i)}{\vartheta^3} - \frac{2E_i}{\vartheta^3 B_i} + \frac{1}{\vartheta^4} \frac{\Gamma''(E_i)}{\Gamma(E_i)} - \frac{1}{\vartheta^4} \left[\frac{\Gamma'(E_i)}{\Gamma(E_i)} \right]^2 - \frac{2}{\vartheta^4 B_i} + \frac{E_i}{\vartheta^4 B_i^2} \right].$$

Seja l a função log-verossimilhança associada ao modelo (4.3.1) e $(\hat{\beta}_0, \hat{\vartheta})$ os estimadores de máxima verossimilhança de (β_0, ϑ) do modelo nulo. Também defina $\boldsymbol{\theta} = (\beta_0, \vartheta)'$ e $m_i = \sum_{k=1}^{K_i} T_{ik} X_{ik}$. Então a Estatística do teste de Escore é dada por

$$S = \frac{U^2}{i_{\beta_1\beta_1} - \boldsymbol{i}'_{\beta_1\boldsymbol{\theta}}\boldsymbol{I}_0^{-1}\boldsymbol{i}_{\beta_1\boldsymbol{\theta}}},$$

em que

$$U = \frac{\partial l_0}{\partial \beta_1} = \delta_{\cdot\cdot} - exp\{\beta_0\}\sum_{i=1}^{g}\frac{m_i E_i}{B_i} \ ,$$

$$i_{\beta_1\beta_1} = -\frac{\partial^2 l_0}{\partial \beta_1^2} = exp\{\beta_0\}\sum_{i=1}^{g}\frac{m_i E_i B_i - exp\{\beta_0\}m_i^2 E_i}{B_i^2} \ ,$$

$$\boldsymbol{i}_{\beta_1\boldsymbol{\theta}} = (i_{\beta_1\beta_0}, i_{\beta_1\vartheta})' = \left(-\frac{\partial^2 l_0}{\partial\beta_1\partial\beta_0}, -\frac{\partial^2 l_0}{\partial\beta_1\partial\vartheta}\right)',$$

$$i_{\beta_1\beta_0} = exp\{\beta_0\}\sum_{i=1}^{g}\frac{m_i E_i}{B_i^2}[B_i - exp\{\beta_0\}T_{i\cdot}] \ ,$$

$$i_{\beta_1\vartheta} = exp\{\beta_0\}\sum_{i=1}^{g}\frac{m_i}{B_i^2}(E_i - B_i) \ .$$

A melhor divisão é aquela com o maior valor da estatística Escore entre todas as divisões possíveis. De acordo com a melhor divisão, todo o conjunto de dados é dividido em dois nós filhos. O mesmo procedimento é então aplicado a cada um desses dois nós filhos. Repetindo esses passos a árvore inicial, A_0, é construída.

Poda: A árvore final pode ser qualquer sub-árvore de A_0. Porém o número de sub-árvores pode ser enorme, mesmo para uma A_0 de tamanho moderado. Para diminuir as escolhas foi seguida a ideia de poda do algoritmo CART (Breiman et al., (1984)) para podar, iterativamente, o nó mais fraco da árvore inicial A_0. Já que o valor máximo da estatística Escore é usado para construir a árvore e provê uma medida natural de qualidade de ajuste para cada nó interno, pode-se adotar a medida de complexidade da divisão de LeBlanc e Crowley (1993):

$$G_\alpha(A) = G(A) - \alpha|A - \tilde{A}|,$$

em que $|\cdot|$ indica o número de nós, $\tilde{A}$ é o conjunto de todos os nós terminais da árvore A, $|A-\tilde{A}|=|\tilde{A}|-1$ é o conjunto de todos os nós internos, α é o parâmetro de complexidade, e $G(A) = \sum_{h\notin\tilde{A}} S(h)$ soma os valores máximos das estatísticas Escore dos nós internos da árvore A, penalizada pela complexidade da árvore, i.e., $|A - \tilde{A}|$, o número total de nós internos.

Como o parâmetro de complexidade α cresce do zero, haverá um nó interno h que será o primeiro a se tornar ineficaz porque A_h, o galho o qual h é o nó raiz, é inferior comparado a h como sendo um simples nó terminal. Esse nó é então podado.

A partir da árvore inicial A_0, para qualquer nó terminal h de A_0, calcula-se o ponto de corte

$$g(h) = \frac{\sum_{i \in (A_h - \tilde{A}_h)} S(i)}{|A_h - \tilde{A}_h|},$$

em que $S(i)$ representa o valor máximo da estatística Escore associada ao nó interno i. Então o nó h^* em A_0 é o nó interno tal que $g(h^*) = min_{h \in (A_0 - \tilde{A}_0)} g(h)$. Seja A_1 a sub-árvore após podar h^*, ou seja, $A_1 = A_0 - A_{h^*}$ e seja $\alpha_1 = g(h^*)$. Então, aplica-se o mesmo procedimento para truncar A_1 e obter A_2 e α_2. Continuando esse procedimento, será obtida uma sequência decrescente de sub-árvores $A_0 \geq A_1 \geq \cdots \geq A_{M+1} \geq A_M$, em que A_M é a árvore nula contendo apenas o nó raiz, e a notação $A_0 \geq A_1$ significa que A_1 é sub-árvore de A_0, assim como a sequência de valores crescentes de α, $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_M < \infty$.

A otimização deste algoritmo de complexidade de divisão é estabelecida por Breinam et al. (1984) e LeBlanc e Crowley (1993). Pode ser mostrado que, para qualquer árvore A e qualquer $\alpha > 0$, existe uma única sub-árvore de A que maximiza a medida de complexidade de divisão. Esta única sub-árvore, denotada por $A(\alpha)$ é chamada de sub-árvore otimamente podada de A para o parâmetro de complexidade α. Foi ainda estabelecido que a sub-árvore A_m resultante do procedimento de poda acima é a sub-árvore otimamente podada de A_0 para qualquer α tal que $\alpha_m \leq \alpha \leq \alpha_{m+1}$. Em particular, este é o caso para a média geométrica de α_m e α_{m+1}, $\alpha'_m = \sqrt{\alpha_m \alpha_{m+1}}$.

Escolha do tamanho da árvore: Uma árvore com o melhor tamanho precisa ser selecionada a partir daquela sequência aninhada. Novamente, a medida de complexidade da divisão $G_\alpha(A)$ é o critério para comparar essas candidatas, ou seja, a árvore A^* é a sub-árvore com melhor tamanho se $G_\alpha(A^*) = max_{m=0,\cdots,M}[G(A_m) - \alpha|A_m - \tilde{A}_m|]$. LeBlanc e Crowley (1993) sugerem que, para selecionar a melhor sub-árvore, α deve ser fixado entre $2 \leq \alpha \leq 4$. Porém, devido à natureza adaptativa dos algoritmos gulosos, a medida de qualidade de divisão $G(A_m)$ é muito otimista e precisa ser validada para prover uma avaliação mais honesta. Dois métodos de validação são propostos, o método com amostra de aprendizagem e o método de reamostragem, dependendo do tamanho da amostra estudada. A seguir, a notação $G(\Omega_2; \Omega_1, A)$ denota a medida de qualidade de divisão validada para a árvore A, construída usando a amostra Ω_1, e validada usando a amostra Ω_2.

Quando o tamanho amostral é grande, o método de amostra de aprendizagem pode ser aplicado. Primeiro divide-se o conjunto de dados em duas partes: a amostra de aprendizagem Ω_1 e a amostra teste Ω_2. Então constrói-se e poda-se a árvore inicial A_0 usando Ω_1. No estágio da determinação do tamanho da árvore, a medida de qualidade de divisão $G(A_m)$ é recalculada ou validada como $G(\Omega_2; \Omega_1, A)$, usando a amostra teste Ω_2. A sub-árvore que maximiza o G_α validado é a árvore de melhor tamanho.

Quando o tamanho amostral é pequeno ou moderado, pode-se aplicar a técnica de reamostragem. Foi adotado o método *Bootstrap* proposto por Efron (1983) para corrigir o viés no problema de predição (veja também LeBlanc e Crowley, (1993)). Neste método, primeiro a árvore inicial A_0 é construída e podada usando o conjunto de dados completo. Depois amostras *bootstrap* Ω_b, $b = 1, \cdots, B$ são retiradas do conjunto de dados completo Ω. Recomenda-se usar um número de amostras *bootstrap* entre 25 e 100 (LeBlanc e Crowley (1993)). Baseado em cada amostra bootstrap Ω_b, uma árvore A_0^b é construída e podada. Seja $A_b(\alpha'_m)$,$m = 1, \cdots, M$ as sub-árvores otimamente podadas correspondentes aos valores α'_m, em que (α'_m),$m = 1, \cdots, M$ são médias geométricas dos α_m respectivamente, obtidos podando A_0. O estimador *bootstrap* de $G(A_m)$ é dado por

$$G^b(A_m) = G(\Omega; \Omega, A(\alpha'_m)) - \frac{1}{B}\sum_{b=1}^{B}[G(\Omega_b; \Omega_b, A_b(\alpha'_m)) - G(\Omega; \Omega_b, A_b(\alpha'_m))].$$

O raciocínio por trás deste estimador é o seguinte:

A primeira parte desta equação, $G(\Omega; \Omega, A(\alpha'_m))$, usa a mesma amostra tanto para construir, quanto para calcular a medida de qualidade de divisão, o que resulta em uma estimativa de G muito otimista.

A segunda parte tem como objetivo corrigir esse viés. Esta parte é uma estimativa da diferença entre os G's quando a mesma amostra (Ω_b) versus amostras diferentes (Ω_b e Ω) são usadas para construir e podar a árvore e para recalcular G, e é tirada a média sobre as B amostras *bootstrap*.

Capítulo 5

Aplicação

Nesta seção são analisados os dados do Estudo do Coração das Famílias de Baependi descrito em Oliveira et al. (2008). Este estudo obteve informações sobre 119 famílias (1712 indivíduos), que vivem no pequeno vilarejo de Baependi, no Estado de Minas Gerais. Os dados foram coletados entre Dezembro de 2005 e Janeiro de 2006 de acordo com um desenho amostral planejado. O número de gerações por família variou de 2 a 4 (54% das famílias tinham 3 gerações, e 45% tinham 2 gerações). Participaram do estudo apenas indivíduos com 18 anos ou mais. A idade média dos participantes é de 44 anos com variação entre 18 e 100 anos. Para cada participante um questionário foi usado para obter informações sobre as relações familiares, características demográficas, histórico médico e fatores de risco ambientais. Foram realizadas medidas antropométricas, exames físicos e eletrocardiograma dos participantes por estudantes de medicina treinados. Além disso, glicemia de jejum, colesterol total, frações de lipoproteínas e triglicerídeos foram obtidos por técnicas padrões em amostras de sangue.

Considerando que famílias com apenas um ou dois indivíduos não fornecem muita informação para estudos de famílias, foram analisados os dados de 81 famílias, envolvendo 1673 indivíduos sendo 43,1% do sexo masculino. O tamanho das famílias variou entre 3 e 157 indivíduos com média de 20,6 indivíduos por família. A média das idades médias por família foi de 45,5 anos (desvio padrão = 6,6). A idade mínima entre as famílias variou entre 18 e 66 anos (com média de 22,8 anos e d.p. = 7,5 anos), e a idade máxima entre 46 e 100 anos (com média de 74,3 anos e d.p.

= 10,6 anos). Indivíduos com nível de glicose $\geq$ 126 mg/dl ou que estavam usando medicação para diabetes no momento do estudo foram definidos como diabéticos. Similarmente, aqueles com colesterol LDL $\geq$ 160 mg/dl ou que estavam usando medicamento contra colesterol alto foram considerados com colesterol alto. Indivíduos com média da pressão sanguínea sistólica $\geq$ 140 mm Hg e/ou pressão sanguínea diastólica $\geq$ 90 mm Hg ou que estavam usando medicamentos anti-hipertensivos no momento do estudo foram definidos como Hipertensivos. A idade de diagnóstico de hipertensão para indivíduos diagnosticados previamente ao estudo foi definida como a idade declarada por eles. Indivíduos que ainda não haviam sido diagnosticados como hipertensos e que apresentaram pressão sanguínea sistólica e/ou diastólica igual ou maior que 140 e 90 mm Hg, respectivamente, no exame realizado durante o estudo, tiveram sua idade atual definida como a idade de diagnóstico de pressão alta. Indivíduos livres de hipertensão até o fim do estudo tiveram a idade de diagnóstico dessa doença censurada. Para diabetes e colesterol alto, a idade de diagnóstico foi definida da mesma maneira. Entre os 1673 indivíduos, 595 (35,6%) foram identificados com pressão alta, 157 (9,4%) com diabetes e 196 (11,8%) com colesterol alto.

Para este estudo foram avaliadas 44 covariáveis (apresentadas na Figura 5.1) obtidas através de combinações ou questões únicas do questionário aplicado na pesquisa, dos exames de sangue e do eletrocardiograma. Foi observada uma média de 2,3% de dados faltantes entre as covariáveis sendo o mínimo 0% e o máximo 18,4%.

5.1 Análise Descritiva

Observando a Figura 5.8 nota-se que a distribuição de Idade e IMC tem valores maiores para os grupos com doença (hipertensão, diabetes e colesterol alto) comparados com os grupos que não apresentam as doenças. Em relação à hipertensão, verifica-se que a distribuição das variáveis Creatinina, Triglicérides, Ácido Úrico, Sokolow Lyon e Cornell dos hipertensos possui valores um pouco maiores que a distribuição dos não hipertensos (Figuras 5.2, 5.3 e 5.5). Os valores de Frequência Cardíaca e Intervalo PR também são levemente maiores no grupo de hipertensos comparado com o grupo que não possui a doença (Figuras 5.4 e 5.7), o contrário ocorre para SAQRS (Figura 5.4).

Variável	n	% de observações faltantes	Categorias e/ou Descrição
Gênero	1673	0,0%	Masculino, Feminino
Nível de Educação	1671	0,1%	Analfabeto, Primário, 1ºGrau, 2ºGrau, Superior
Renda Mensal	1626	2,8%	Até 1 SM, De 1 a 5 SM, De 5 a 10 SM, De 10 a 20 SM, Mais de 20 SM
Cor da Pele	1666	0,4%	Branca, Preta, Parda, Amarela, Vermelha, Outros
Estado Civil	1664	0,5%	Casado ou em união consensual, Solteiro, Separado, Viúvo
Sedentarismo	1672	0,1%	Não, Sim
Tabagismo	1669	0,2%	Não, No passado, Sim
Hipertensão	1670	0,2%	Não, Sim
Diabetes tipo II	1667	0,4%	Não, Sim
Colesterol Elevado	1661	0,7%	Não, Sim
Angina	1673	0,0%	Não, Sim
Infarto	1673	0,0%	Não, Sim
Derrame	1673	0,0%	Não, Sim
Insuficiência Cardíaca	1673	0,0%	Não, Sim
Pedra no rim	1673	0,0%	Não, Sim
Doença no rim	1673	0,0%	Não, Sim
Dialise	1673	0,0%	Não, Sim
Depressão	1673	0,0%	Não, Sim
Varizes	1673	0,0%	Não, Sim
Doença no pulmão	1673	0,0%	Não, Sim
Angioplastia	1664	0,5%	Não, Sim
Diabetes_fam	1647	1,6%	Número de parentes que apresentam Diabetes tipo II (0, 1, 2, ≥3)
PA_fam	1646	1,6%	Número de parentes que apresentam Hipertensão (0, 1, 2, ≥3)
Derrame_fam	1659	0,8%	Número de parentes que já sofreram Derrame (0, 1, 2, ≥3)
Infarte_fam	1656	1,0%	Número de parentes que já sofreram Infarto (0, 1, 2, ≥3)
Doença_Rim_fam	1642	1,9%	Número de parentes que apresentam Doença no Rim (0, 1, 2, ≥3)
Pedra_Rim_fam	1644	1,7%	Número de parentes que apresentam Pedra no Rim (0, 1, 2, ≥3)
Angioplastia_fam	1656	1,0%	Número de parentes que já sofreram Angioplastia (0, 1, 2, ≥3)
Desconforto nas Perna	1669	0,2%	Não, Sim
Ureia	1646	1,6%	Marcador da função renal
Creatinina	1650	1,4%	Marcador da função renal
Triglicérides	1643	1,8%	Tipo de Colesterol
Ácido Úrico	1651	1,3%	Excretor renal
Frequência Cardíaca	1603	4,2%	Número de batimentos cardíacos por minuto
SAQRS	1542	7,8%	Parâmetro do Eletrocardiograma
Sokolow_Lyon	1525	8,8%	Índice de sobrecarga ventricular esquerda
Cornell	1365	18,4%	Índice de sobrecarga ventricular esquerda
Duração QRS	1579	5,6%	Parâmetro do Eletrocardiograma
Intervalo QT	1559	6,8%	Parâmetro do Eletrocardiograma
Intervalo PR	1467	12,3%	Parâmetro do Eletrocardiograma
Segmento PR	1470	12,1%	Parâmetro do Eletrocardiograma
Idade	1673	0,0%	Idade do indivíduo no momento da pesquisa, em anos

Figura 5.1: Descrição das covariáveis usadas na construção das árvores

Considerando os diabéticos, as Figuras 5.2 e 5.3 mostram que os valores de Ureia, Creatinina, Triglicérides e Ácido Úrico são um pouco maiores para os diabéticos comparados com os não diabéticos; a Figura 5.7 mostra que os valores de Intervalo PR e Segmento PR são levemente maiores para os diabéticos e que os valores SAQRS (Figura 5.4) e Duração QRS (Figura 5.6) são levemente menores para esse grupo.

O grupo que foi diagnosticado com colesterol alto possui valores um pouco maiores para Triglicérides e Ácido Úrico (Figura 5.3), e levemente maiores para Intervalo PR (Figura 5.7) quando comparado com o grupo que não foi diagnosticado com essa doença.

Como já dito, os métodos de árvore de regressão são robustos à outliers, por isso as várias observações aberrantes mostradas nos boxplots não foram tratadas com maior cuidado.

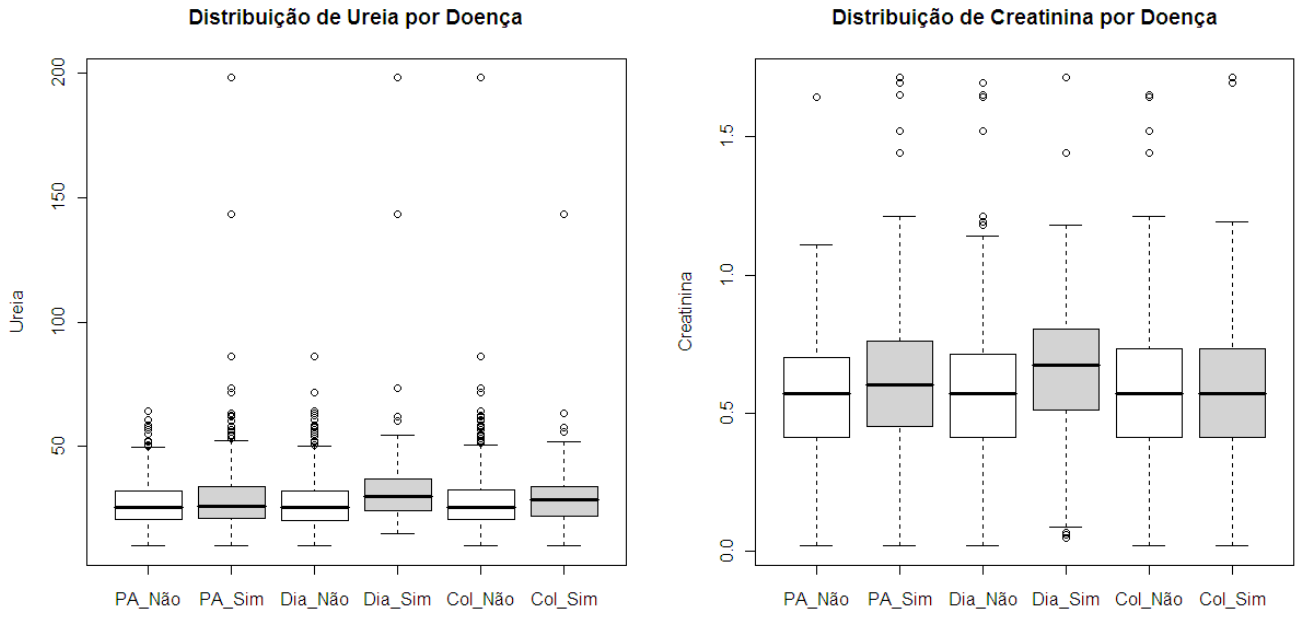


Figura 5.2: Boxplots das covariáveis Ureia e Creatinina

Os gráficos apresentados nas Figuras 5.9 a 5.17 mostram a distribuição, em porcentagem, dos indivíduos em cada uma das categorias das covariáveis categóricas do estudo, separada por presença e ausência das doenças estudadas. Pela Figura 5.9 nota-se que a diferença entre os sexos é mais

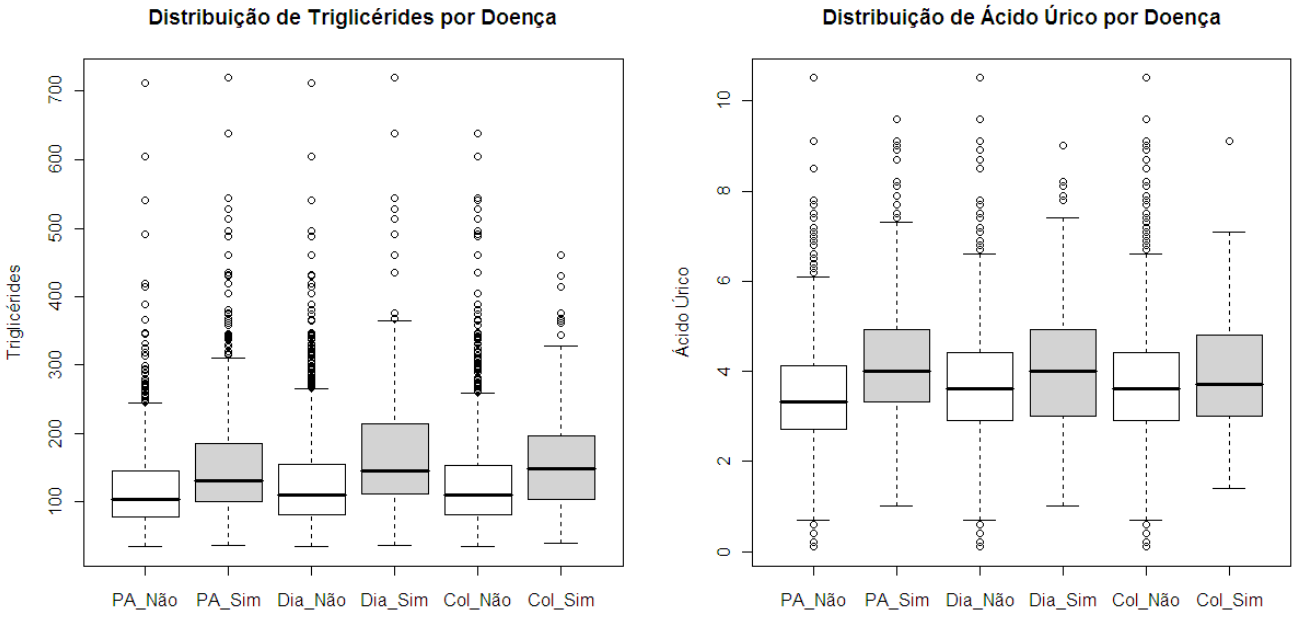


Figura 5.3: Boxplots das covariáveis Triglicérides e Ácido Úrico

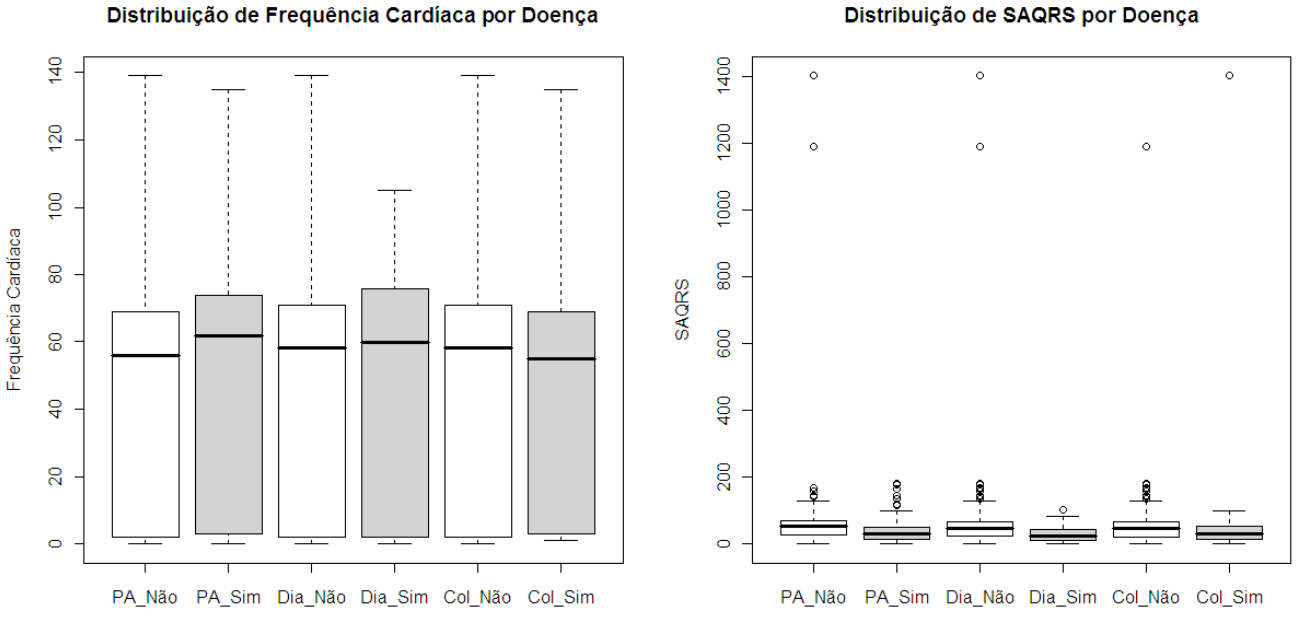


Figura 5.4: Boxplots das covariáveis Frequência Cardíaca e SAQRS

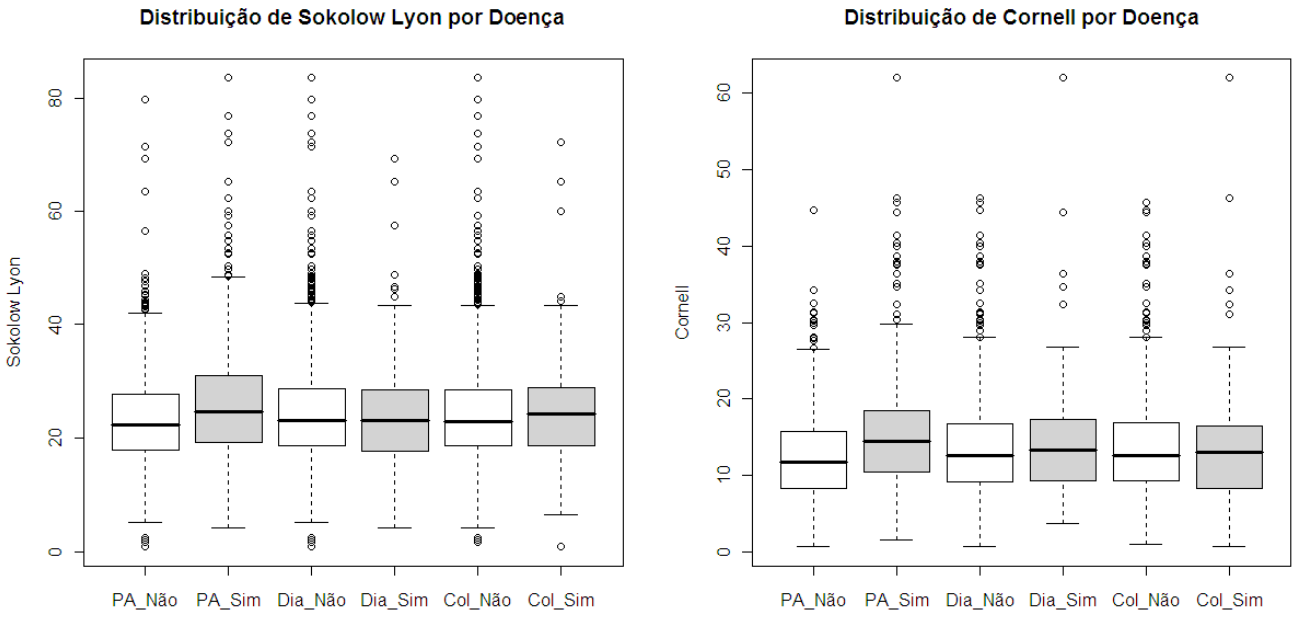


Figura 5.5: Boxplots das covariáveis Sokolow Lyon e Cornell

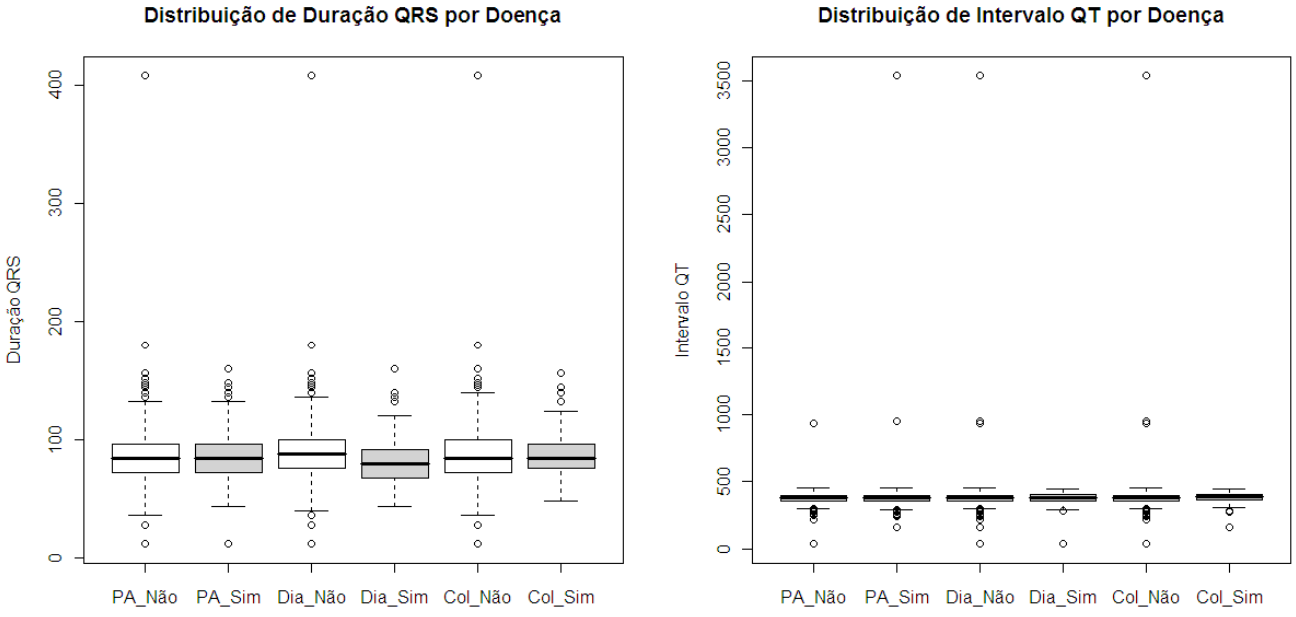


Figura 5.6: Boxplots das covariáveis Duração QRS e Intervalo QT

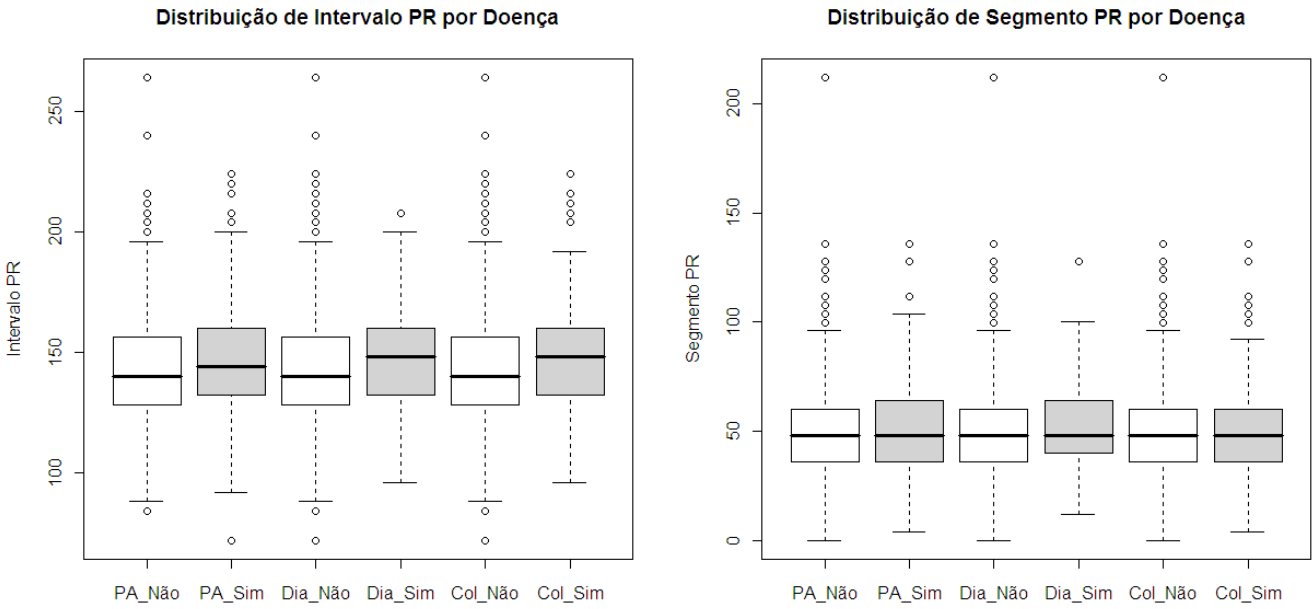


Figura 5.7: Boxplots das covariáveis Intervalo PR e Segmento PR

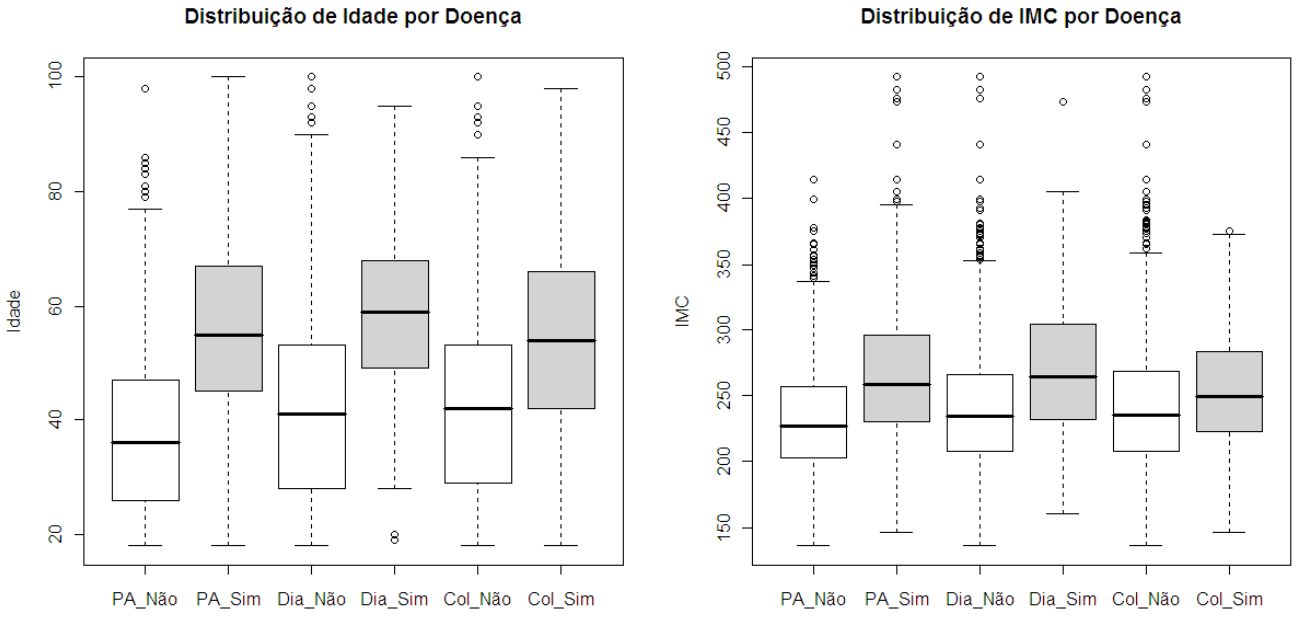


Figura 5.8: Boxplots das covariáveis Idade e IMC

acentuada nos grupos doentes, dando indícios de que há mais indivíduos do sexo feminino doente que do sexo masculino.

A Figura 5.10 mostra que a porcentagem de analfabetos doentes é um pouco maior que analfabetos não doentes. Ainda pela Figura 5.10 nota-se que a porcentagem de indivíduos com segundo grau completo é maior entre os não doentes, considerando hipertensão e diabetes, porém para colesterol alto essa diferença é menos acentuada. Já pela Figura 5.11 verifica-se que entre os hipertensos e diabéticos há maior porcentagem de indivíduos com renda mensal de até 1 salário mínimo. Verifica-se também que há mais pessoas que recebem de 1 a 5 salários mínimos por mês que não apresentam diabetes e colesterol alto se comparadas às pessoas que apresentam uma dessas duas doenças e estão na mesma faixa salarial.

Pela Figura 5.12 nota-se que há mais indivíduos da cor preta no grupo de hipertensos que no grupo de não hipertensos. Essa diferença fica ainda mais clara se a porcentagem for calculada de forma contrária, ou seja, entre os 86 indivíduos auto-declarados de cor preta, quase 57% apresentam hipertensão. A Figura 5.13 mostra que a porcentagem de casados ou em união consensual dentro do grupo com hipertensão ou colesterol alto é um pouco maior que a porcentagem dentro do grupo com pressão ou colesterol normais. A diferença entre as porcentagens de solteiros comparando os doentes e os não doentes é mais acentuada. Nota-se ainda por essa figura que entre os doentes há mais viúvos que entre os não doentes. Essa relação entre estado civil e as doenças pode estar também ligada à idade dos indivíduos, já que normalmente, solteiros são mais jovens e viúvos apresentam idade mais avançada.

Em relação ao sedentarismo (Figura 5.14) nota-se que não há diferenças entre os grupos doentes e não doentes. O mesmo é notado em relação ao tabagismo (Figura 5.15).

A Figura 5.16 apresenta a porcentagem de indivíduos que afirmaram ter cada uma das doenças durante a entrevista. Note que a porcentagem de indivíduos com Hipertensão, Diabetes e Colesterol Alto é um pouco menor que as mencionadas no início desse capítulo, isso porque alguns indivíduos foram detectados com a doença durante os exames realizados na pesquisa, mas ainda não tinham o conhecimento da mesma. Nota-se também que quase 50% dos entrevistados reclamaram de desconforto nas pernas. A segunda maior porcentagem é observada para hipertensão, as porcentagens de indivíduos com varizes, depressão e colesterol alto são parecidas e ocupam a

terceira posição.

A Figura 5.17 apresenta a porcentagem de membros da família com cada uma das doenças que constavam no questionário. Note que menos de 20% dos entrevistados afirmaram que nenhum parente possui pressão alta. Mais de 50% dos indivíduos relataram que nenhum parente possui diabetes. Para as demais cinco doenças a distribuição é semelhante.

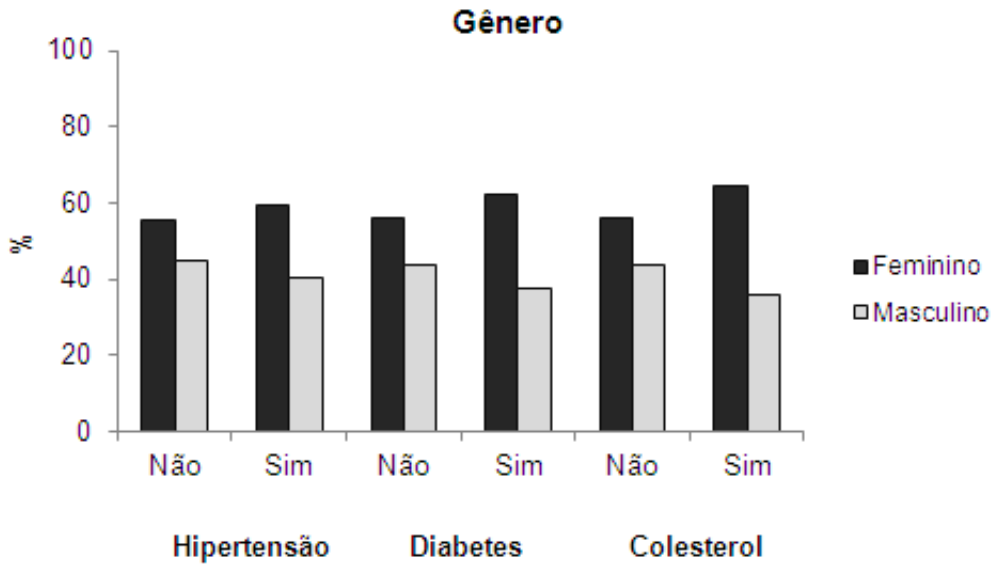


Figura 5.9: Distribuição do Gênero em relação às doenças

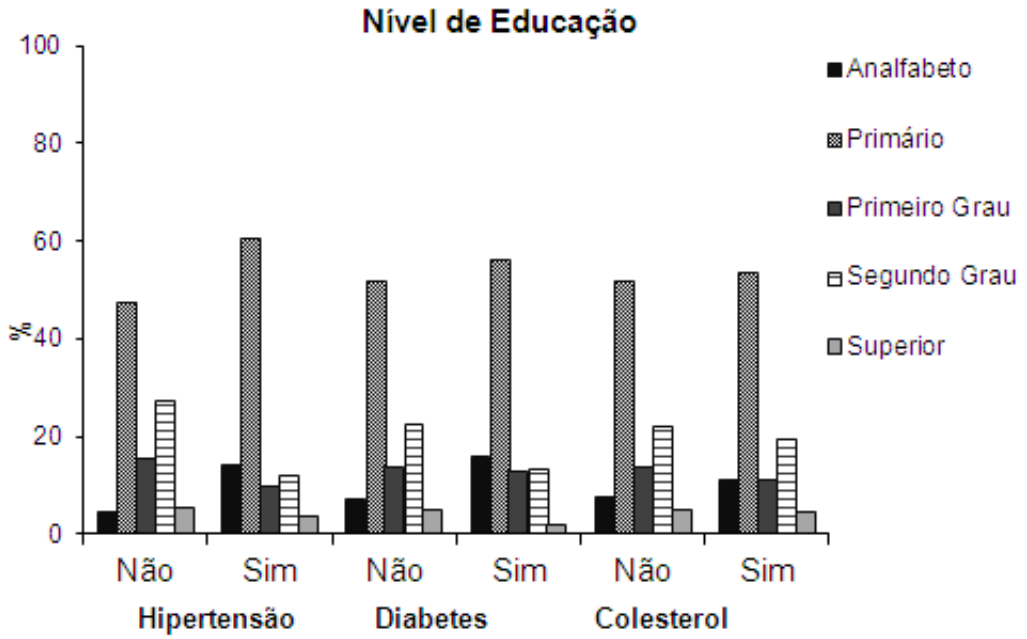


Figura 5.10: Distribuição da Educação em relação às doenças

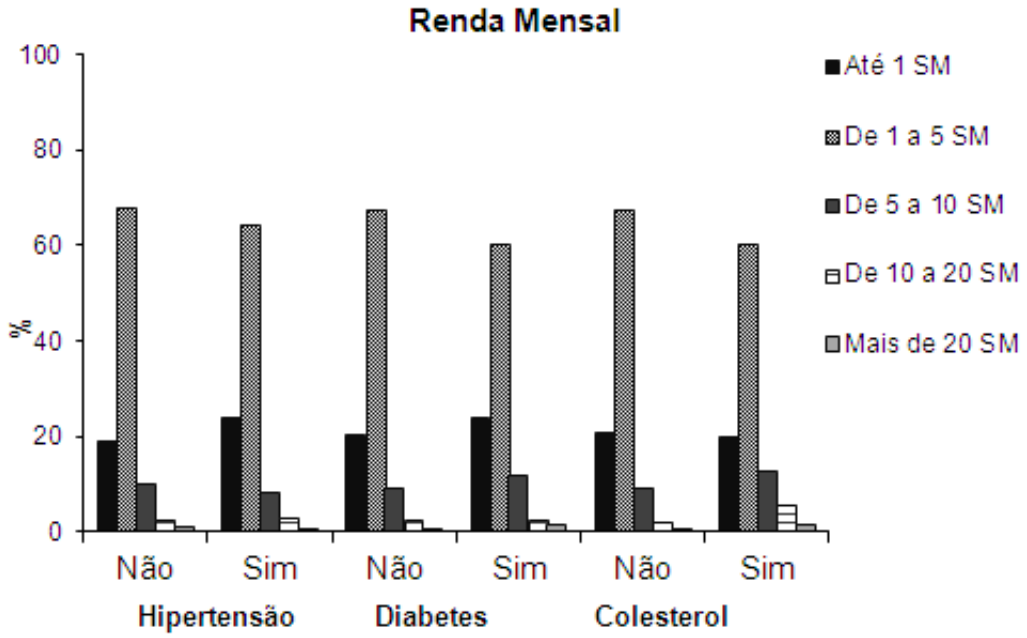


Figura 5.11: Distribuição da Renda em relação às doenças

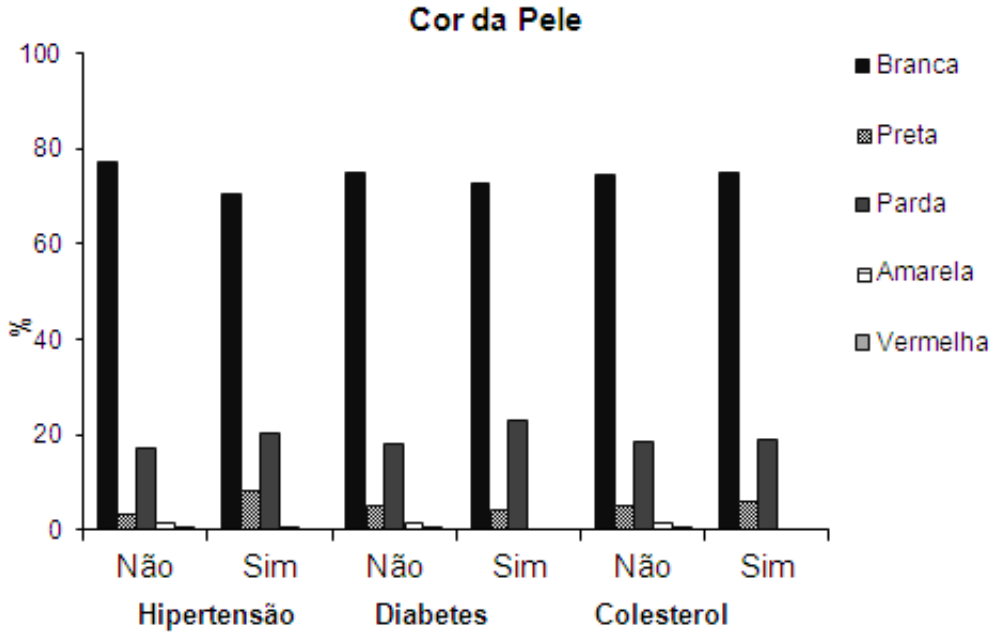


Figura 5.12: Distribuição da Cor da Pele em relação às doenças

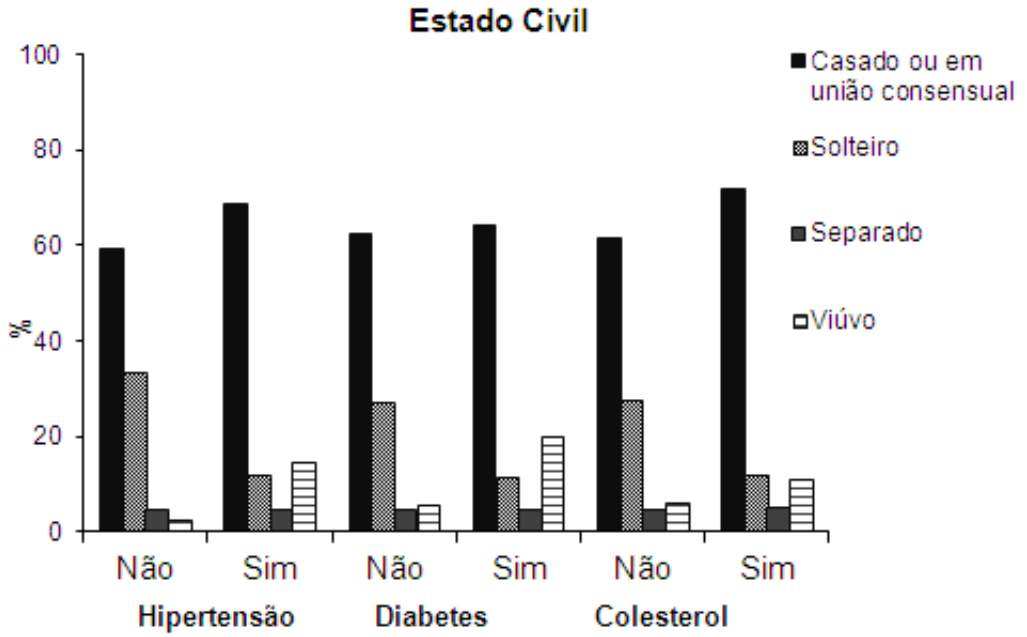


Figura 5.13: Distribuição do Estado Civil em relação às doenças

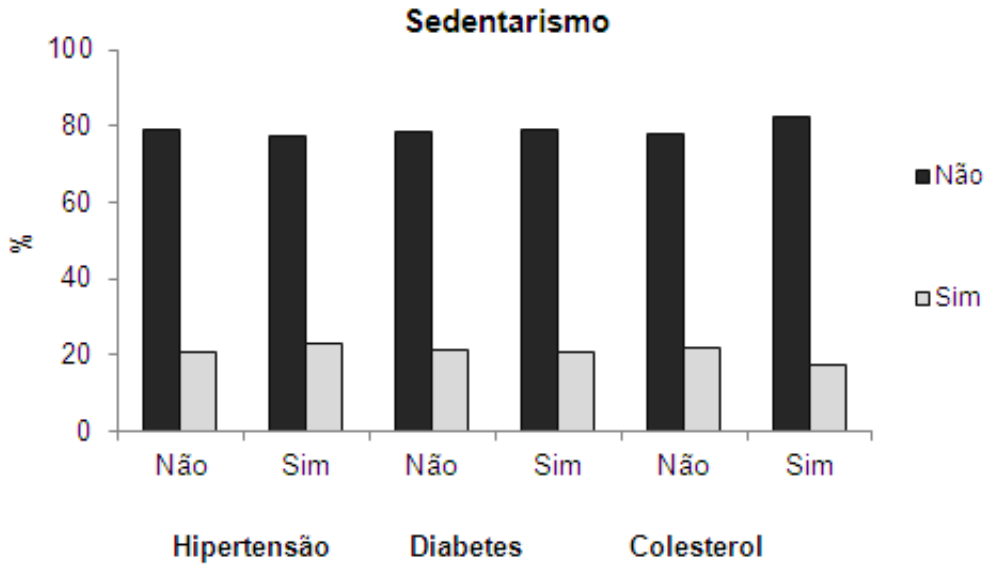


Figura 5.14: Distribuição do Sedentarismo em relação às doenças

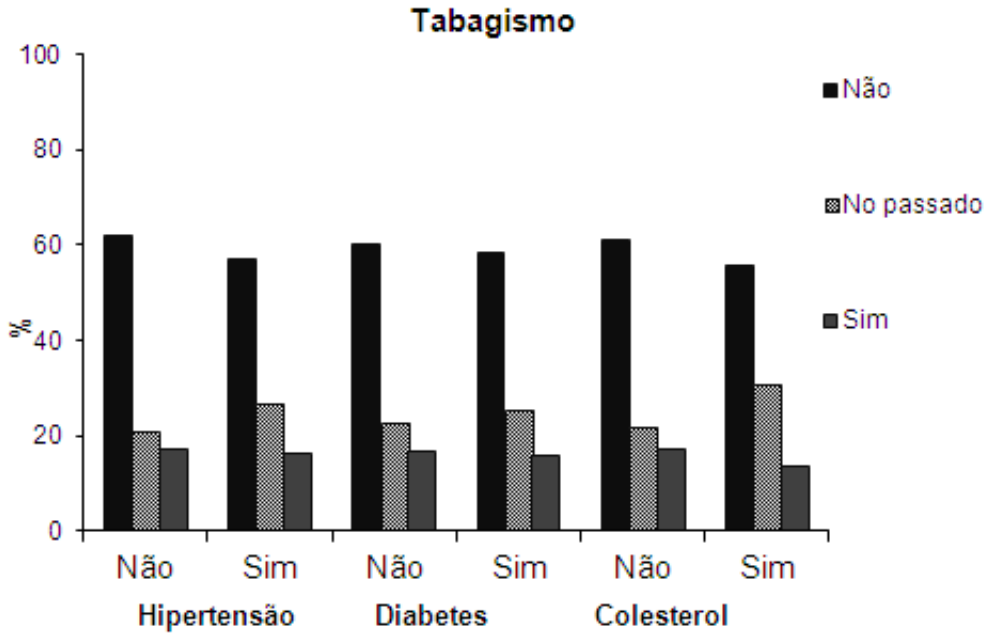


Figura 5.15: Distribuição do Tabagismo em relação às doenças

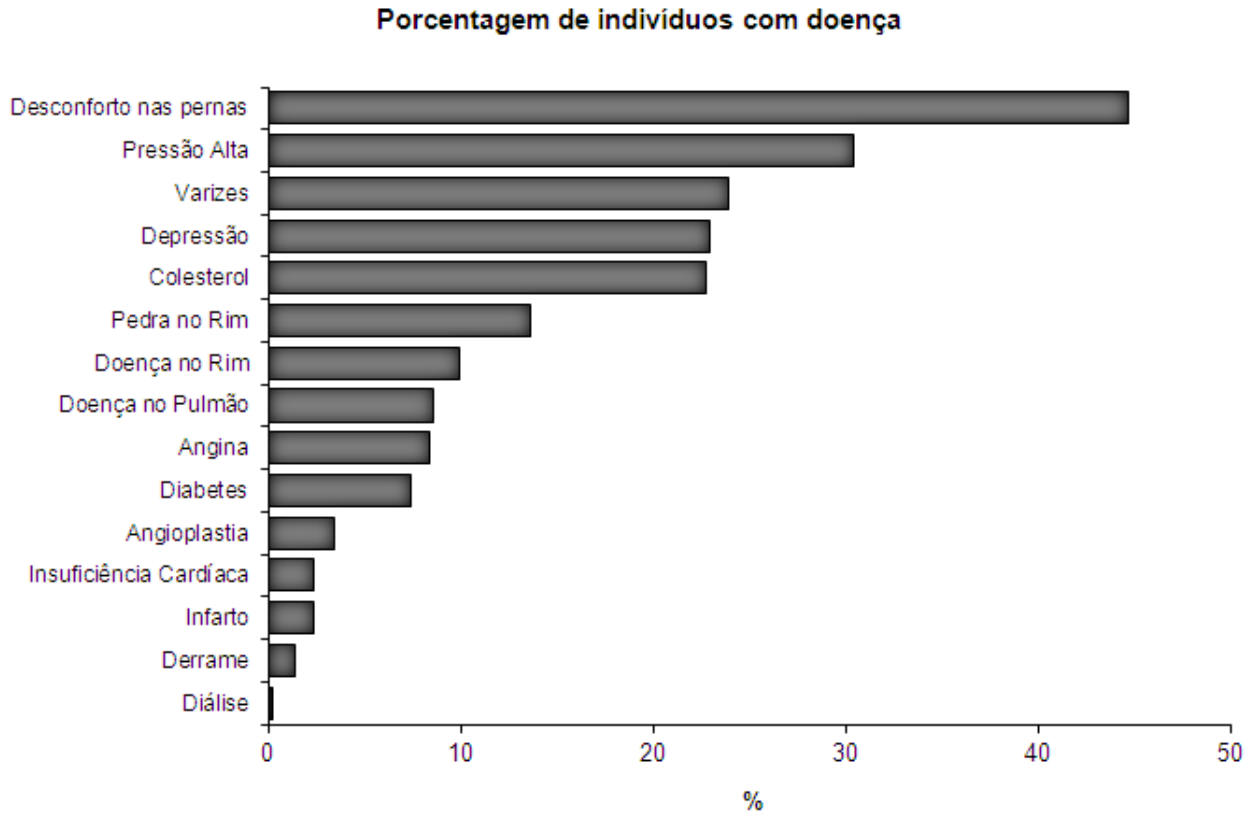


Figura 5.16: Porcentagem de indivíduos que afirmaram ter a respectiva doença durante a entrevista

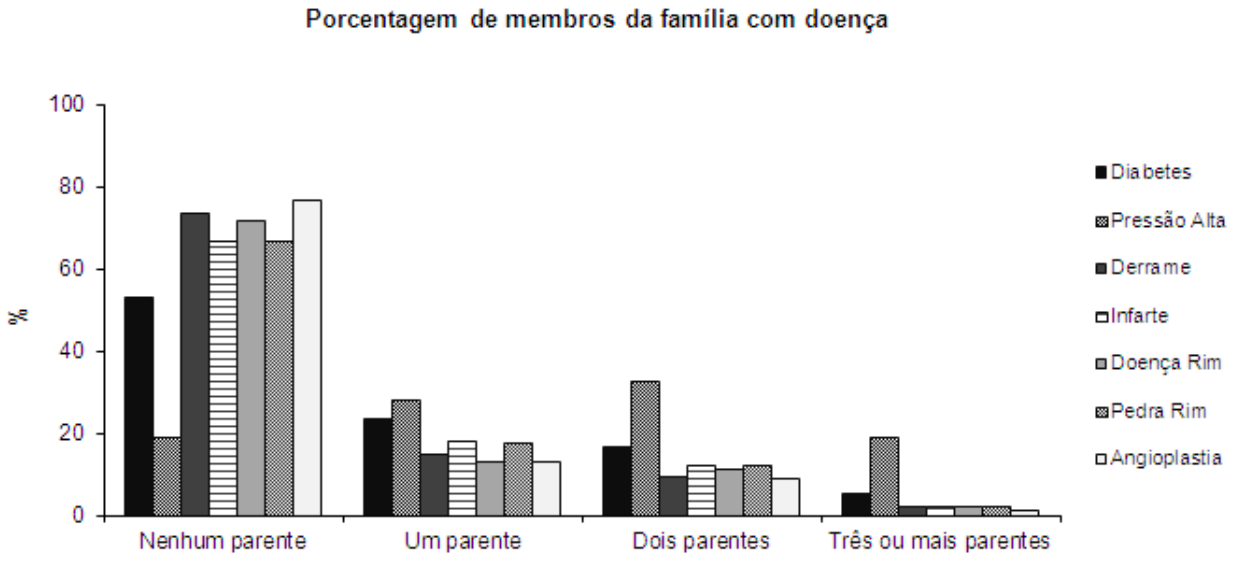


Figura 5.17: Porcentagem de membros da família com a respectiva doença

5.2 Árvore de Regressão sem considerar correlação entre as famílias

Nesta seção foi usado o método de Poisson do pacote *rpart* do R, que é equivalente ao método apresentado na Seção 4.2.3, como mostrado na Seção 4.2.4. Foram construídas três árvores de regressão, uma para cada variável resposta: idade até o diagnóstico de hipertensão, idade até o diagnóstico de diabetes e idade até o diagnóstico de colesterol alto. As árvores construídas nesta seção assumem independência entre todos os indivíduos, ou seja, não consideram correlação entre os membros de uma mesma família.

A Figura 5.18 apresenta a árvore construída para a idade de diagnóstico de hipertensão. Em cada nó da árvore é possível ler o número e porcentagem de indivíduos que foram diagnosticados com pressão alta até o término da pesquisa ou que foram censurados; um gráfico de barra dentro dos nós torna essa informação mais visual; é possível ler ainda o número total de indivíduos no nó (N) e a área sob a curva de sobrevivência (A) formada pelos indivíduos do respectivo nó. É intuitivo que quanto maior a área sob a curva menor o risco e maior a idade de diagnóstico de pressão alta. Como foi mostrado na Figura 5.1, algumas variáveis têm observações faltantes, portanto nem sempre a soma do número de indivíduos dos dois nós filhos é igual ao número de indivíduos do nó pai.

Avaliando a Figura 5.18, verifica-se que a única variável escolhida pelo algoritmo para diferenciar os indivíduos em relação à idade de diagnóstico de hipertensão foi o IMC. Inicialmente os indivíduos foram divididos em dois grupos, o primeiro com IMC menor ou igual a 27,3 e o segundo com IMC maior que 27,3. Comparando esses dois grupos entre si, nota-se que indivíduos com IMC menor ou igual a 27,3 têm menor risco de apresentar a doença (ou maior idade de diagnóstico), já que a área sob a curva para esse grupo é maior se comparada com o grupo de indivíduos com IMC maior que 27,3.

Uma medida descritiva do risco relativo entre dois nós pode ser obtida pela razão entre as áreas sob a curva desses dois nós, fazendo isso para os Nós 2 e 3, verifica-se que o risco de indivíduos com IMC $>$ 27,3 apresentar hipertensão é 16% maior se comparado com o risco dos indivíduos com IMC

menor ou igual a 27,3. Considerando o grupo de sobrepeso (Nó 3: IMC>27,3), a variável IMC foi novamente a que mais diferenciou os indivíduos em relação à idade de diagnóstico de hipertensão. Note que 70,2% dos indivíduos obesos (IMC≥32,4) foram diagnosticados com a doença (Nó 7), contra 53,7% entre os indivíduos com IMC entre 27,3 e 32,4 (Nó 6). Comparando as áreas sob a curva desses dois nós tem-se que a área do Nó 6 é 20% maior que a área do Nó 7, ou seja, o risco de indivíduos obesos apresentar hipertensão é 20% maior que o risco de pessoas com sobrepeso apresentar esta doença. Comparando as áreas sob as curvas dos Nós 2 e 7 observa-se que o risco de pessoas com IMC $\geq 32,4$ ser diagnosticadas com hipertensão é 36% maior que o risco de pessoas com IMC<27,3. A Figura 5.19 mostra de forma mais visual o quanto o IMC afeta a idade de diagnóstico de hipertensão, quanto maior o IMC maior o decaimento da curva.

Assim como para pressão alta, a primeira variável que mais diferenciou os indivíduos em relação à idade de diagnóstico de diabetes tipo II foi o IMC (Figura 5.20), e com nível bastante parecido. Note que, entre todos os grupos formados pela árvore, indivíduos com IMC $\geq$ 27,9 e que tem histórico familiar de hipertensão (Nó 7) tem o maior risco de apresentar diabetes tipo II, ou menor idade de diagnóstico. O menor risco é observado no grupo formado pelos indivíduos com IMC $<$ 27,9, Duração QRS $\geq$ 82 e Creatinina <0,345 (Nó 8). Comparando os Nós 2 e 3 (peso normal *vs* sobrepeso) tem-se que o risco de pessoas com sobrepeso serem diagnosticadas com diabetes é 8% maior que o risco de pessoas de peso normal. Entre as pessoas com sobrepeso (Nó 3), o fato de ter histórico familiar de hipertensão (Nó7) aumenta o risco de diabetes tipo II em 9% comparando com aqueles que não têm histórico familiar de pressão alta (Nó 6). Entre os indivíduos com IMC ≤27,9 a Duração QRS (um parâmetro do eletrocardiograma) dividiu os indivíduos no nível 82, sendo que aqueles com Duração QRS<82 tem maior risco ou menor idade de diagnóstico de diabetes tipo II. O Nó 4 foi dividido em outros dois nós (8 e 9) pela variável Creatinina, comparando esses dois nós, nota-se que o risco de apresentar diabetes tipo II em pessoas com resultado de creatinina maior ou igual a 0,345 é 14% maior que o risco entre aqueles com resultado de creatinina menor que 0,345, lembrando que isso entre aqueles que têm IMC menor que 27,9 e Duração QRS maior ou igual a 82. O Nó 5 também foi dividido em outros dois nós, dessa vez a variável Triglicérides foi a selecionada pelo algoritmo. Comparando as áreas sob a curva dos Nós 10 e 11 verifica-se que indivíduos com valor de triglicérides maior ou igual a 124,1 tem risco 10% maior de apresentar diabetes tipo II se

comparados com indivíduos com valor de triglicérides menor que 124,1, considerando ainda que esses indivíduos apresentam IMC menor que 27,9 e Duração QRS menor que 82.

A Figura 5.21 mostra as curvas de sobrevivência dos grupos que formam os nós terminais da árvore apresentada na Figura 5.20. Como a medida descritiva para comparar os riscos entre os nós foi a área sob a curva de sobrevivência, a informação da Figura 5.21 está inserida na 5.20, porém, a partir das curvas observa-se que todas começam a decrescer após a idade de 30 anos, mostrando que antes dessa idade o risco de ser diagnosticado com diabetes tipo II é baixo. As curvas dos Nós 7 (sobrepeso e histórico familiar de pressão alta) e 11(IMC<27,3, Duração QRS<82 e triglicérides maior que 124,1) decrescem mais rápido que as demais sendo que a curva do Nó 11 decresce ainda mais rápido que a do Nó 7. A curva do Nó 8 (MC<27,3, Duração QRS maior ou igual a 82 e creatinina menor que 0,345) tem apenas um degrau e a probabilidade de não diagnóstico de diabetes tipo II é maior que 80% para qualquer idade, considerando esse grupo de estudo.

Nível de educação foi a variável que mais distinguiu os indivíduos em relação à idade de diagnóstico de colesterol alto (Figura 5.22). Entre os indivíduos com no máximo o primeiro grau completo observou-se 12,6% com colesterol alto, entre os indivíduos com pelo menos o segundo grau incompleto essa porcentagem foi de 10,4%. Entre os indivíduos com pelo menos o segundo grau incompleto, o risco de apresentar colesterol alto é maior para aqueles com índice de Cornell menor que 8,75 (Nó 6). Comparando os Nós 6 e 7, nota-se que o risco de colesterol alto é 38% maior para aqueles que pertencem ao Nó 6. Para os indivíduos que concluíram no máximo o primeiro grau (Nó 2), o sexo apresenta-se como um fator importante, sendo que pessoas do sexo feminino tem risco 17% maior de serem diagnosticadas com colesterol alto. Entre os indivíduos que tem no máximo o primeiro grau completo e são do sexo masculino, o fato de apresentar SAQRS (um parâmetro do eletrocardiograma) menor que 59,3 aumenta o risco de ser diagnosticado com colesterol alto em 4%. Já entre os indivíduos que tem no máximo o primeiro grau completo e do sexo feminino, o fato de apresentar triglicérides $\geq$ 138,4 aumenta o risco de colesterol alto em 21%.

Analisando a Figura 5.23 nota-se que a curva do Nó 6 (pelo menos segundo grau incompleto e com índice de Cornell<8,75) começa a decrescer em idade próxima a 20 anos sendo a curva com caimento mais acentuado. A curva do Nó 11 (mulheres que no máximo concluíram o primeiro grau e com valor de triglicérides maior ou igual a 138,4) é a segunda com maior caimento, porém

nota-se que o diagnóstico de colesterol alto fica mais provável após os 50 anos.

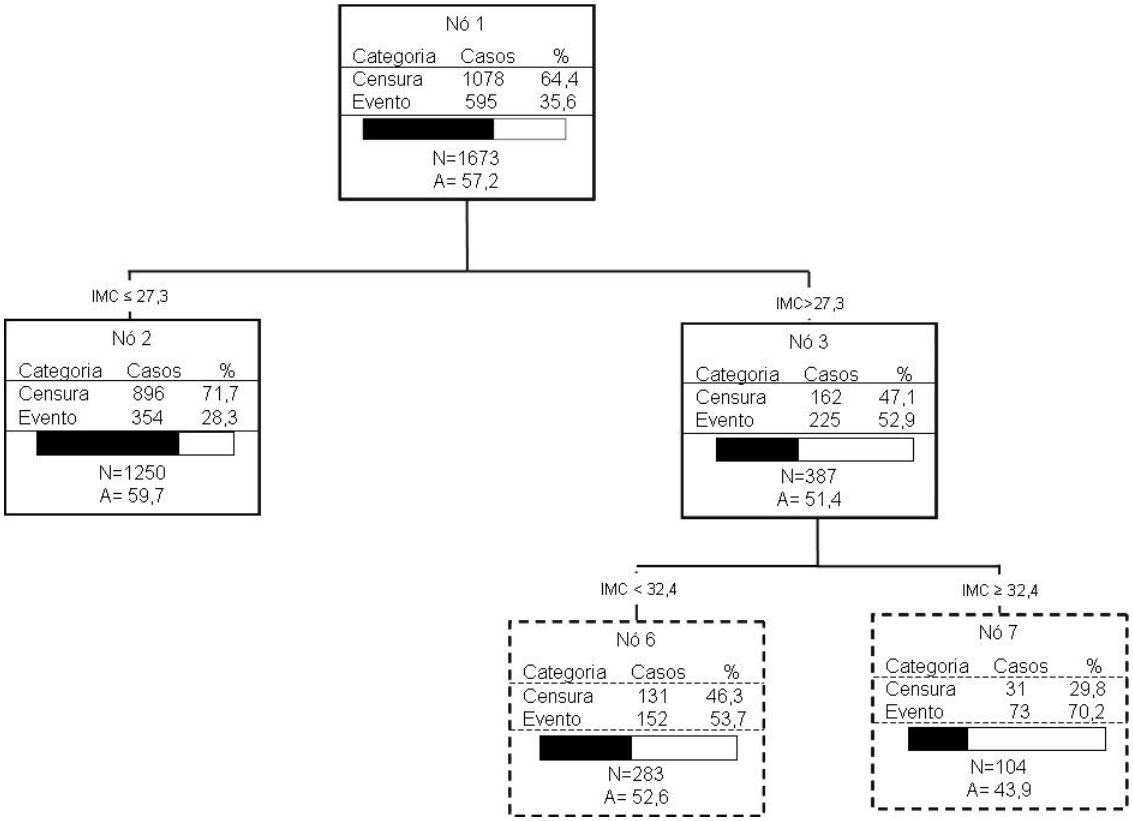


Figura 5.18: Árvore para hipertensão sem considerar a estrutura de correlação dentro das famílias. A = área sob a curva de sobrevivência e N = número de indivíduos no nó.

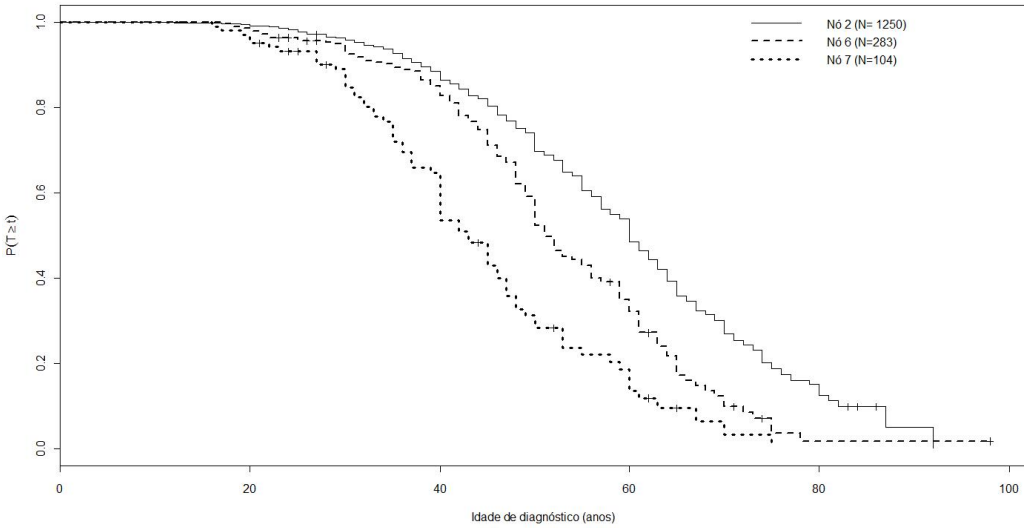


Figura 5.19: Curva de sobrevivência estimada para cada um dos nós terminais da Figura (5.18). **Nó2**: IMC≤27,3. **Nó6**: 27,3<IMC < 32,4. **Nó7**: IMC≥ 32,4.

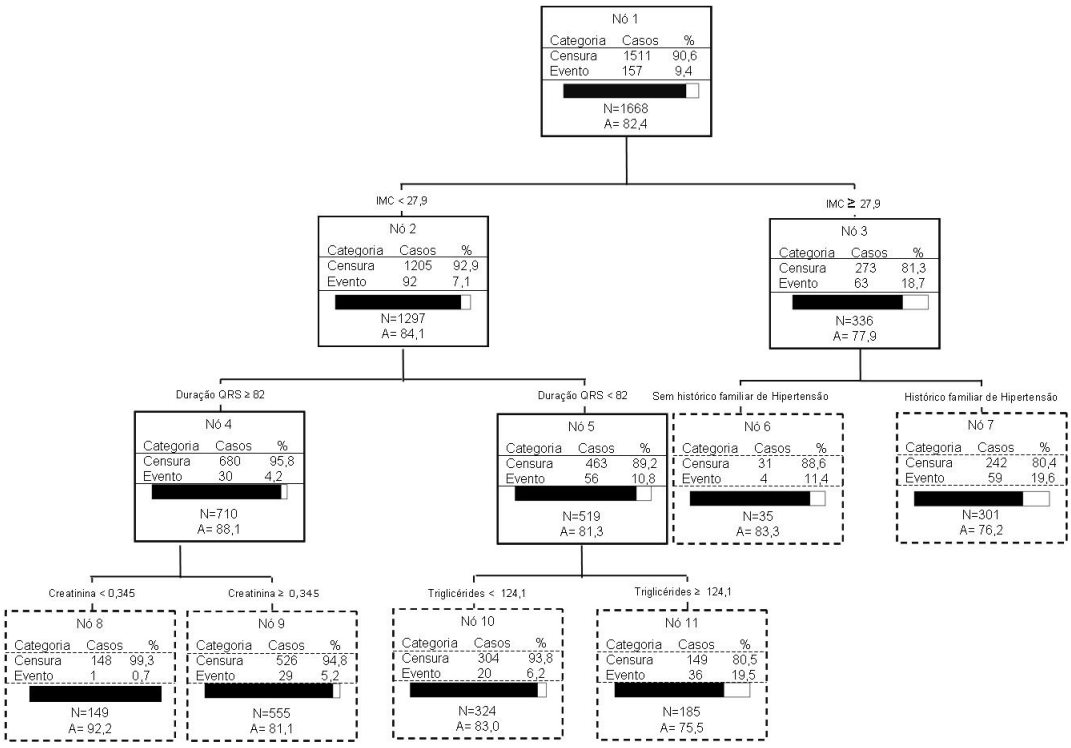


Figura 5.20: Árvore para diabetes sem considerar a estrutura de correlação dentro das famílias. A = área sob a curva de sobrevivência e N = número de indivíduos no nó.

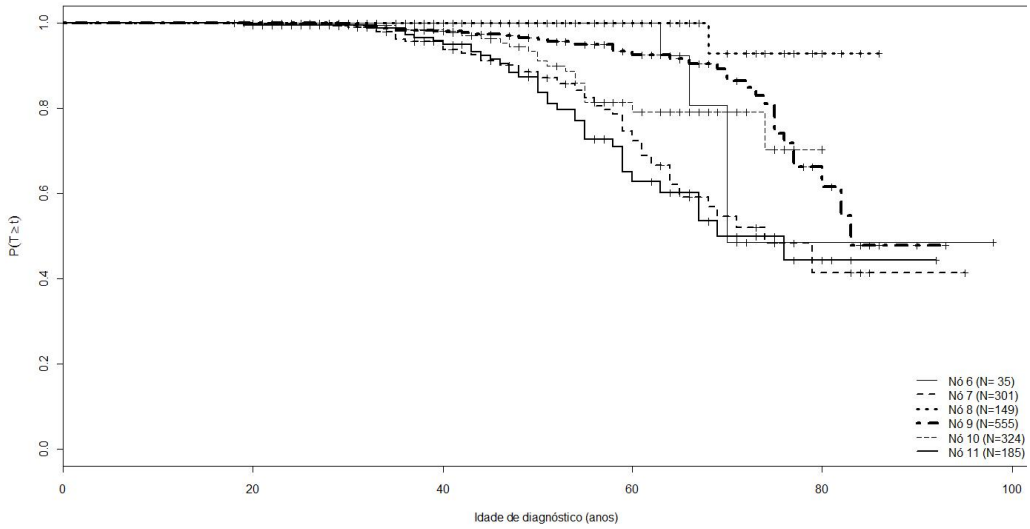


Figura 5.21: Curva de sobrevivência estimada para cada um dos nós terminais da Figura (5.20). **Nó6**: IMC≥27,9; sem histórico familiar de hipertensão. **Nó7**: IMC≥27,9; histórico familiar de hipertensão. **Nó8**: IMC<27,9; Duração QRS≥82; Creatinina<0,345.**Nó9**: IMC<27,9; Duração QRS≥82; Creatinina≥0,345. **Nó10**: IMC<27,9; Duração QRS<82; Triglicérides<124,1. **Nó11**: IMC<27,9; Duração QRS<82; Triglicérides≥124,1.

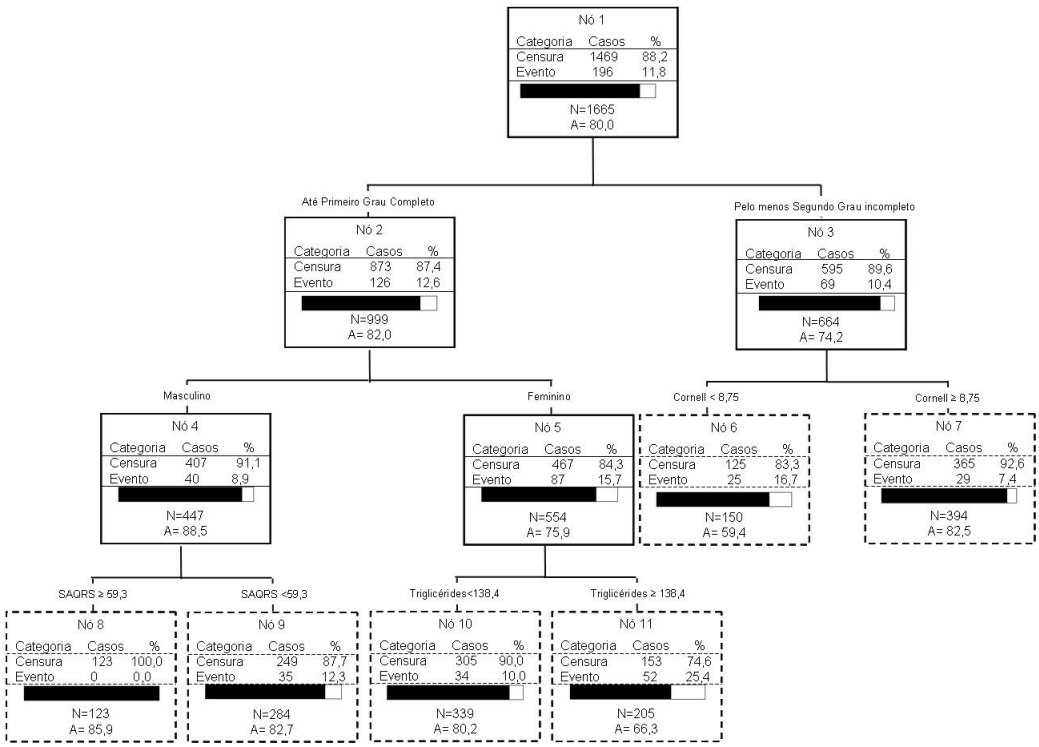


Figura 5.22: Árvore para colesterol alto sem considerar a estrutura de correlação dentro das famílias. A = área sob a curva de sobrevivência e N = número de indivíduos no nó.

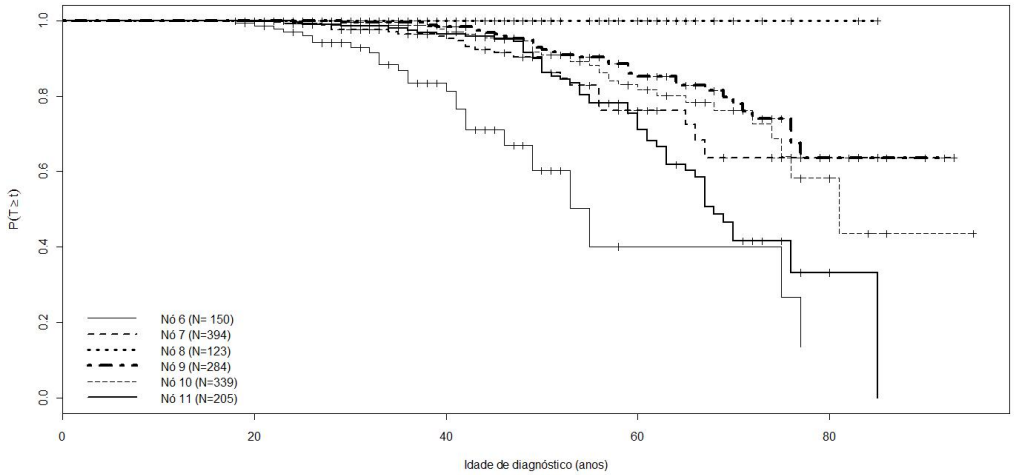


Figura 5.23: Curva de sobrevivência estimada para cada um dos nós terminais da Figura (5.22). **Nó6**: Pelo menos segundo grau incompleto; Cornel<8,75. **Nó7**: Pelo menos segundo grau incompleto; Cornel≥8,75. **Nó8**: Pelo menos primeiro grau completo; Masculino; SAQRS≥59,3.**Nó9**: Pelo menos primeiro grau completo; Masculino; SAQRS<59,3. **Nó10**: Pelo menos primeiro grau completo; Feminino; Triglicérides<138,4. **Nó11**: Pelo menos primeiro grau completo; Feminino; Triglicérides≥138,4.

5.3 Árvore de Regressão considerando correlação entre as famílias

Nesta seção são apresentadas as árvores construídas segundo o método apresentado na seção 4.3. Assim como em 5.2, é construída uma árvore para cada variável resposta, porém aqui é considerada a possível correlação existente entre os indivíduos de uma mesma família.

Na Figura 5.24 está apresentada a árvore construída considerando idade até o diagnóstico de hipertensão como variável resposta. A mesma medida descritiva usada em 5.2 é usada aqui para avaliar o risco de apresentar pressão alta, ou seja, para cada nó é construída a curva de sobrevivência e então a área sob a curva é calculada e a razão entre áreas é usada para medir o risco relativo entre dois nós. Sendo assim, o risco de pessoas com colesterol alto (Nó 2) apresentar hipertensão é 7% maior que o risco de pessoas que não apresentam colesterol alto (Nó 3). Entre os indivíduos que não foram diagnosticados com colesterol alto, apresentar diabetes tipo II aumenta o risco de ser diagnosticado com pressão alta em 14% (Nó 6 *vs* Nó 7). Para os indivíduos que não tem colesterol alto e nem diabetes tipo II, o fato de ter cor da pele preta aumenta o risco em 27% de ser diagnosticado com hipertensão (Nó 12 *vs* Nó 13).

A Figura 5.25 mostra as curvas de sobrevivência construídas para cada um dos nós terminais da árvore que está na Figura 5.24. Nota-se que o Nó 13 é o grupo com melhor prognóstico, já que apresenta maior sobrevivência em quase todos os tempos, ou seja, indivíduos neste grupo tem maior idade de diagnóstico se comparado com os demais. O Nó 12 é o com pior prognóstico já que sua curva de sobrevivência está abaixo das outras três. Note que a curva do Nó 2 é sempre maior que a curva do Nó 7 com exceção nas idades de diagnóstico entre 50 e 60 anos, aproximadamente.

Apenas a variável colesterol alto foi escolhida pelo algoritmo para a construção da árvore com idade de diagnóstico de diabetes tipo II (Figura 5.26). Note que comparando o Nó 2 com o Nó 3 (Figura 5.27) a curva do Nó 3 é sempre maior a que curva do Nó 2, a partir de 50 anos. A razão entre as áreas das duas curvas é 1,067, indicando que o risco de diabetes tipo II para pessoas com colesterol alto é 7% maior que o risco de pessoas que não apresentam colesterol alto.

Angioplastia foi a única variável escolhida pelo algoritmo para a construção da árvore de idade

de diagnóstico de colesterol alto. Porém é válido ressaltar que a relação causa-efeito aqui é contrária, pois níveis elevados de colesterol podem levar à doenças obstrutivas coronárias fazendo com que o indivíduo precise ser submetido à angioplastia. A Figura 5.29 mostra as curvas de sobrevivência para os dois nós terminais.

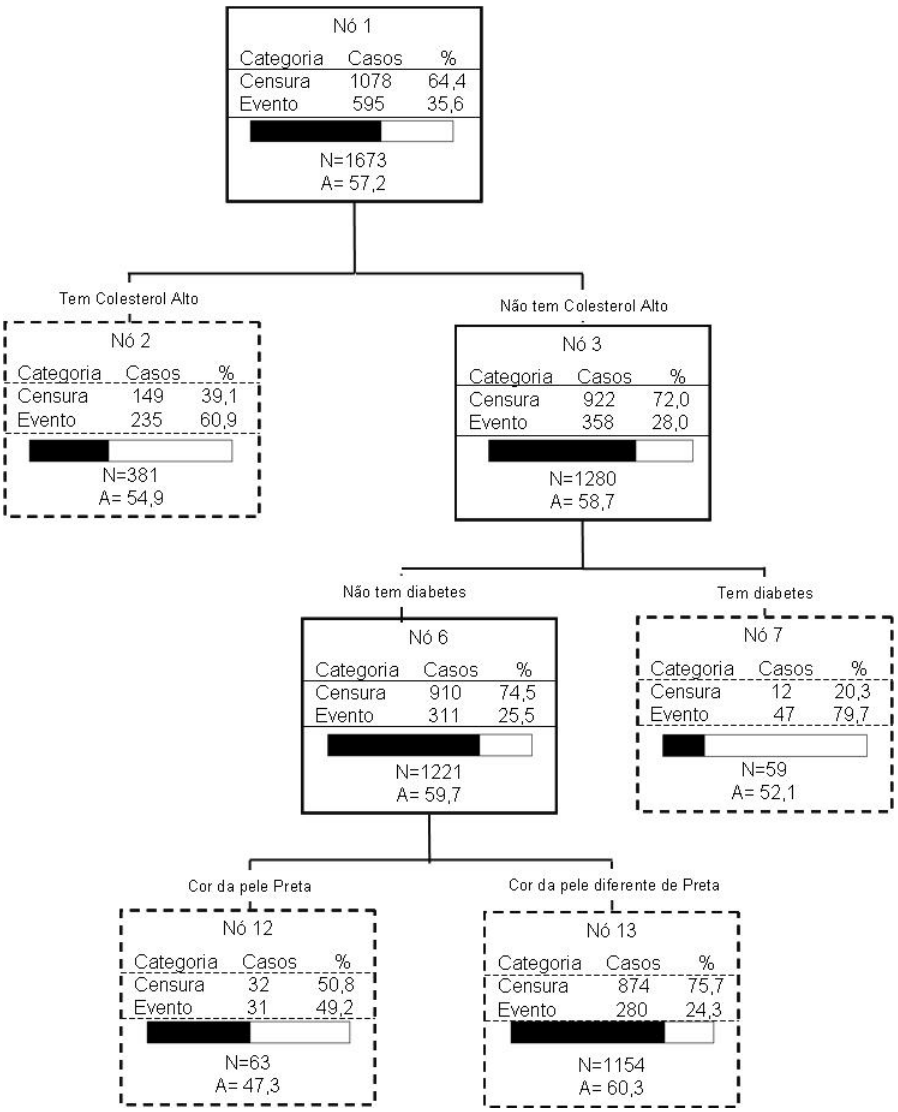


Figura 5.24: Árvore para hipertensão considerando a estrutura de correlação dentro das famílias. A = área sob a curva de sobrevivência e N = número de indivíduos no nó.

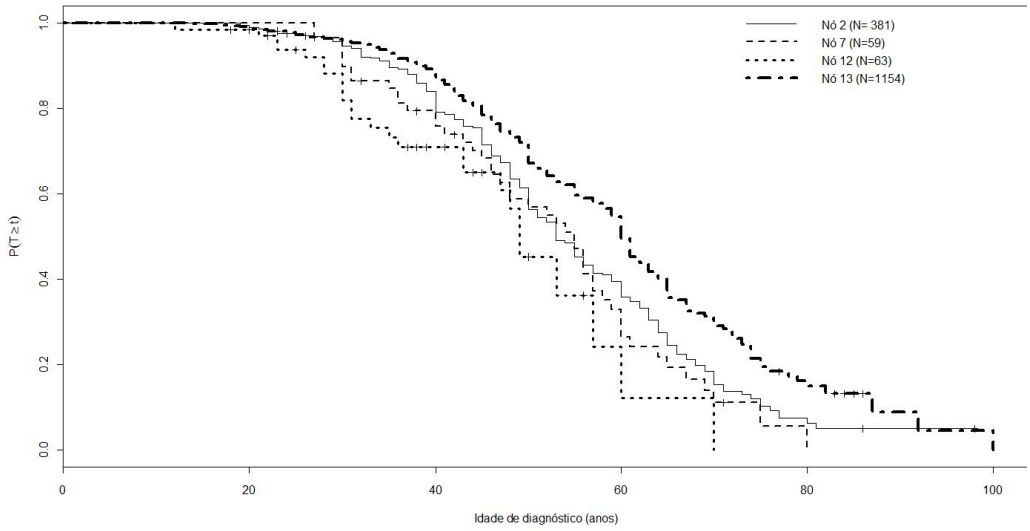


Figura 5.25: Curva de sobrevivência estimada para cada um dos nós terminais da Figura (5.24). **Nó2**: Apresenta colesterol alto. **Nó7**: Não apresenta colesterol alto; apresenta diabetes tipo II. **Nó12**: Não apresenta colesterol alto; não apresenta diabetes tipo II; cor da pele preta.**Nó13**: Não apresenta colesterol alto; não apresenta diabetes tipo II; cor da pele diferente de preta.

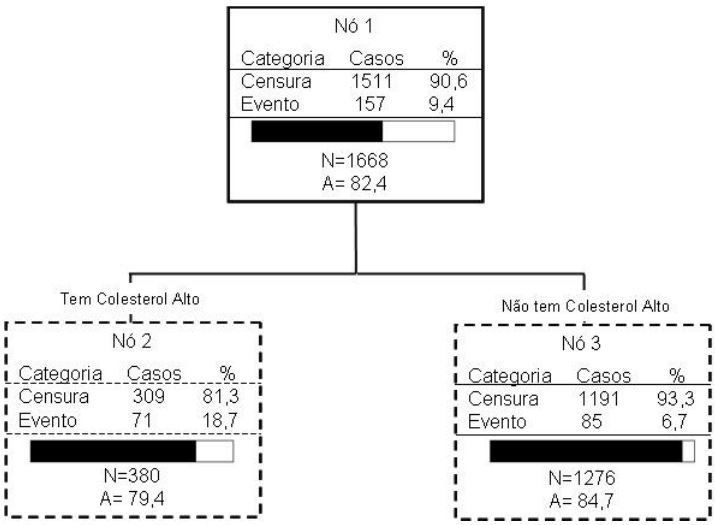


Figura 5.26: Árvore para diabetes considerando a estrutura de correlação dentro das famílias. A = área sob a curva de sobrevivência e N = número de indivíduos no nó.

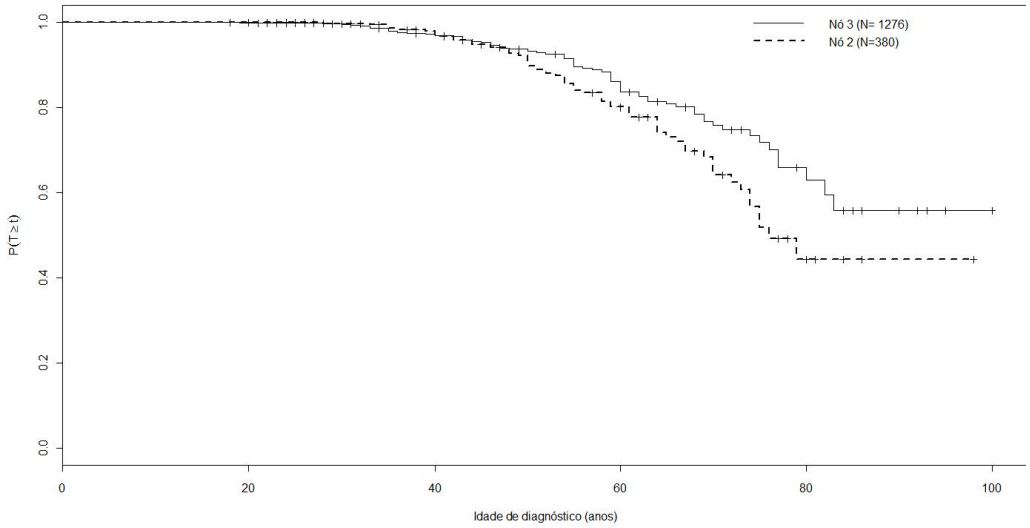


Figura 5.27: Curva de sobrevivência estimada para cada um dos nós terminais da Figura (5.26). **Nó2**: Apresenta colesterol alto. **Nó3**: Não apresenta colesterol alto.

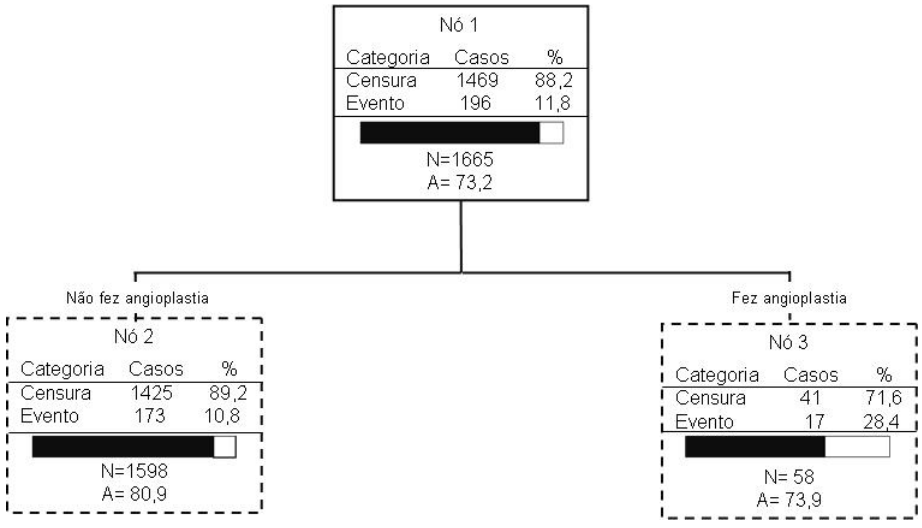


Figura 5.28: Árvore para colesterol alto considerando a estrutura de correlação dentro das famílias. A = área sob a curva de sobrevivência e N = número de indivíduos no nó.

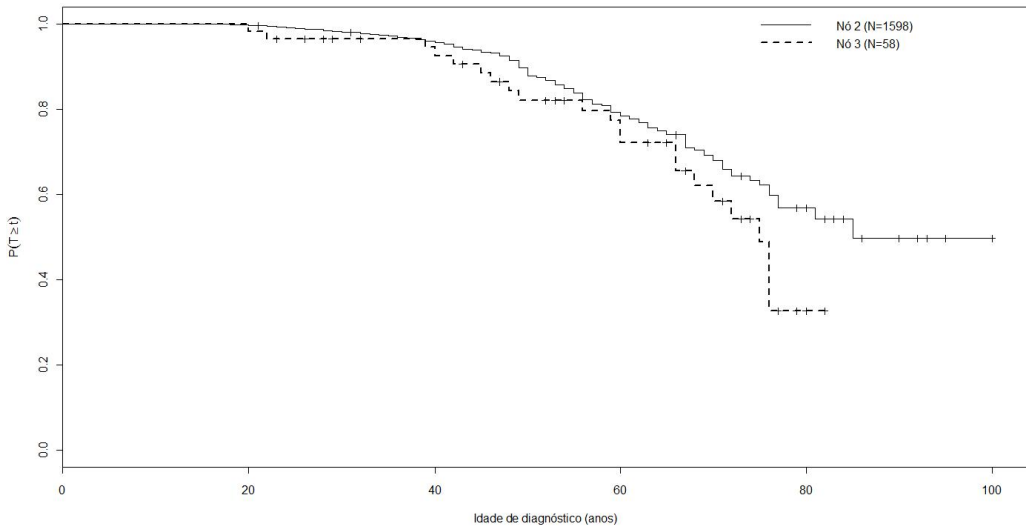


Figura 5.29: Curva de sobrevivência estimada para cada um dos nós terminais da Figura (5.28). **Nó3**: Passou por angioplastia. **Nó4**: Não passou por angioplastia.

5.4 Considerações Finais

Árvore de regressão para dados censurados é uma ferramenta que pode ser aliada à regressão de Cox, para ajudar na interpretação dos resultados ou mesmo ajudar o pesquisador a definir as categorizações das variáveis.

Neste estudo os algoritmos das duas metodologias selecionaram variáveis diferentes, para construir as árvores que consideram a estrutura de correlação entre os indivíduos de uma mesma família e as árvores que não consideram essa possível correlação, mostrando a importância de um modelo bem definido.

Inicialmente foram construídas árvores considerando a idade do indivíduo no momento da pesquisa como uma covariável, em todas as árvores, com exceção da que foi construída considerando correlação entre indivíduos de uma mesma família, com a variável resposta sendo a idade de diagnóstico de colesterol alto, a idade foi a covariável que mais diferenciou os indivíduos. O resultado observado foi que o risco de apresentar hipertensão, diabetes tipo II e colesterol alto diminui com o aumento da idade, que também foi observado por Giolo et al. (2009) nesse mesmo conjunto de dados. Porém optou-se por apresentar neste trabalho as árvores construídas sem inserir a idade do indivíduo, pois devida a alta correlação dessa covariável com as variáveis respostas, a interpretação dos resultados ficaram prejudicadas e alguns resultados importantes foram mascarados, como a relação entre cor da pele e hipertensão. É sabido que pessoas com cor da pele preta tem maior chance de apresentar hipertensão, e a doença se apresenta em níveis mais graves nesse grupo de indivíduos.

Para cada uma das árvores geradas nesse trabalho foi gerado também um modelo de Cox, considerando as mesmas covariáveis e categorizações das árvores. Para as árvores que consideram a correlação entre indivíduos de uma mesma família foi inserida a variável de fragilidade no modelo de Cox. Todas as covariáveis foram significativas (p-valores $< 0{,}038$) e a suposição de proporcionalidade dos riscos não foi rejeitada para nenhuma das covariáveis consideradas nos modelos (p-valores $> 0{,}436$). O teste para fragilidade mostrou haver associação significativa entre as idades de diagnóstico dos indivíduos de uma mesma família (p-valores$<0{,}028$).

Para o conjunto de dados estudado nesse trabalho, o algoritmo que considera correlação entre

indivíduos de uma mesma família demorou, em média, 10 horas para rodar com 100 amostras *bootstrap*, enquanto que o tempo requerido pelo algoritmo que desconsidera essa correlação é de poucos segundos, ou seja, inserir a estrutura de correlação entre os indivíduos de uma mesma família requer um esforço computacional muito maior.

A medida descritiva de risco relativo, apresentada nesse trabalho, é bastante fácil de calcular e interpretar, porém não substitui os gráficos de sobrevivência, a análise conjunta dessas duas medidas descritivas enriquece a interpretação dos resultados.

Comparando as árvores construídas para idade de diagnóstico de pressão alta, pelo método que considera correlação entre os indivíduos de uma mesma família e pelo método que não considera essa correlação, tem-se que o primeiro mostrou que pessoas com colesterol alto tem maior risco de apresentar hipertensão, o que tem explicação científica, já que colesterol alto forma placas na parede vascular, aumentando a rigidez dessa parede e sua resistência, e consequentemente é necessária maior pressão para a circulação sanguínea. O segundo método mostrou que quanto maior o IMC maior o risco de pressão alta, e a explicação para esse fato segue o mesmo caminho, já que maiores valores de IMC geralmente indicam maior porcentagem de gordura, o que também ajuda a formar placas na parede vascular.

As árvores construídas sem considerar a correlação entre os indivíduos de uma mesma família para idade de diagnóstico de diabetes tipo II (Figura 5.20) e de colesterol alto (Figura 5.22) selecionaram algumas covariáveis que são parâmetros do eletrocardiograma (Duração QRS, Índice de Cornell e SAQRS), nesses casos a relação causa-efeito é contrária, assim como entre angioplastia e colesterol alto, observada na Figura 5.28, ou seja, diabetes tipo II e colesterol alto podem levar à doenças obstrutivas coronárias e consequentemente à alterações nos parâmetros do eletrocardiograma.

Para trabalhos futuros fica a sugestão de desenvolvimento de medida multivariada para realizar a divisão dos nós e a poda da árvore, dessa forma as três variáveis respostas poderão ser analisadas conjuntamente, já que elas são correlacionadas entre si. Outra sugestão de trabalho é o desenvolvimento de medida de divisão para variável resposta binária que incorpore a correlação entre indivíduos de uma mesma família. Assim haverá outra maneira de analisar os dados, modelando a presença ou ausência da doença, desconsiderando a idade de diagnóstico.

Bibliografia

[1] Breiman, L., Friedman, J.H., Oslhen, R.A. e Stone, C.J.. *Classification and Regression Trees*. Belmont, Wadsworth, 1984.

[2] Breslow, N.E.. *Contribution to the discussion of the paper by DR Cox*. Journal of the Royal Statistical Society, **34**, 02, 1972.

[3] Ciampi, A., Thiffault, J., Nakache, J.-P. e Asselain, B.. *Stratification by stepwise regression, correspondence analysis and recursive partition*. Computational Statistics and Data Analysis **04**, 03, 1985.

[4] Clayton, D.. *Fitting a general family of failure-time distributions using GLIM*. Applied Statistics **32**, 02, 1983.

[5] Clayton, D. e Cuzick, J.. *The EM algorithm for Cox's regression model using GLIM*. Applied Statistics **32**, 02, 1985.

[6] Colosimo E.A. e Giolo S.R.. *Análise de Sobrevivência Aplicada*. Edgard Blücher, 2006.

[7] Cox, D.R.. *Regression Models and Life-tables*. Journal of the Royal Statistical Society, **34**, 02, 1972.

[8] Davis, R. e Anderson, J.. *Exponential survival trees*. Statistics in Medicine **08**, 08, 1989.

[9] Dempster, A.P., Laird, N. M. e Rubin D.B.. *Maximum Likelihood from Incomplete Data via the EM-algorithm*. Biometrics, **39**, 01, 1977.

[10] Efron, B.. *Estimating the error rate of a prediction rule: improvements on cross-validation*. Journal of the American Statistical Association **78**, 1983.

[11] Efron, B., Tibshirani, R.J.. *An Introduction to the Bootstrap*. Chapman e Hall, 1993.

[12] Fan, J., Su, X.-G., Levine, R., Nunn, M. e LeBlanc, M.. *Trees for Censored Survival Data by Goodness of Split, with Application to Tooth Prognosis*. Journal of American Statistical Association **101**, 475, 2006.

[13] Fan, J., Nunn, M. E. e Su, X.. *Multivariate Exponential Survival Trees and Their Application to Tooth Prognosis*. Computational Statistics and Data Analysis **53**, 04, 2009.

[14] Friedman, M.. *Piecewise Exponential Models for Survival Data with Covariates*. The Annals of Statistics **10**, 01, 1982.

[15] Geisser, S.. *Predictive Inference*. Chapman e Hall, 1993.

[16] Gordon, L. e Olshen, R. A.. *Tree-structured survival analysis*, Cancer Treatments Reposts **69**, 10, 1985.

[17] Gao, F., Manatunga, A. K. e Chen, S.. *Identification of Prognostic Factors with Multivariate Survival Data*. Biometrics **45**, 04, 2004.

[18] Gao, F., Manatunga, A. K. e Chen, S.. *Developing Multivariate Survival Trees with a Proportional Hazards Structure*. Journal of the American Statistical Association **04**, 2006.

[19] Giolo, S.R., Pereira, A.C., de Andrade, M., Oliveira, C.M., Krieger, J. K. e Soler, J.M.P.. *Genetic Analysis of Age-at-Onset for Cardiovascular Risk Factors in a Brazilian Family Study*. Human Heredity **68**, 02, 2009.

[20] Hastings, W.K.. *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. Biometrika **57**, 01, 1970.

[21] Holford, T.R.. *The analysis of rates and survivorship using log-linear models*. Biometrics **36**, 02, 1980.

[22] Hougaard, P.. *Analysis of Multivariate Survival Data*. New York: Springer-Verlag, 2000.

[23] Hunt, E.B., Marin, J. e Stone, P.J.. *Experiments in Induction*. New York: Academic Press., 1966.

[24] Ibrahim, J.G., Chen, M.H. e Sinha, D.. *Bayesian Survival Analysis*. Springer, 2001.

[25] Klein, J.P.. *Semiparametric estimation of random effects using Cox model based on the EM algorithm*. Biometrics, **48**, 03, 1992.

[26] Laird, N.. *Covariance analysis of censored survival data using log-linear analysis techniques*. Journal of American Statistical Association **76**, 374, 1981.

[27] LeBlanc, M. e Crowley, J.. *Relative Risk Trees for Censored Survival Data*. Biometrics **48**, 02, 1992.

[28] LeBlanc, M. e Crowley, J.. *Survival Trees by Goodness of Split*. Journal of the American Statistical Association **81**, 422, 1993.

[29] Mantel, N.. *Evaluation of survival data and two new rank order statistics arinsing in its consideration*. Cancer Chemotherapy Reports, **50**, 03, 1966.

[30] Messenger, R.C. e Mandell, M.L.. *A model Search Technique for Predictive Nominal Scale Multivariate Analysis*. Journal of the American Statistical Association, **67**, 340, 1972.

[31] Morgan, J.N. e Sonquist, J.A.. *Problems in the Analysis of Survey Data and a Proposal*. Journal of the American Statistical Association, **58**, 302, 1963.

[32] Morgan, J.N. e Messenger, R.C.. *HAID: a Sequential Search Program for the Analysis of Nominal Scale Dependent Variables*. Ann Arbor: Institute for Social Research, University of Michigan, 1973.

[33] Nelson, W.. *On estimating the distribution of random vector when only the coordinate is observable*. Technometrics **12**, 1969.

[34] Oliveira, C.M., Pereira, A.C., de Andrade M., Soler J.M.P. e Krieger J.E. *Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study*. BMC Med Genet, **09**, 32, 2008.

[35] Quinlan, J.R.. *Induction of Decision Trees*. Machine Learning, **01**, 01, 1986.

[36] Segal, M.. *Regression trees for censored data*. Biometrics **44**, 01, 1988.

[37] Sonquist, J.A.. *Multivariate Model Building: the Validation of a Search Strategy*. Ann Arbor: Institute for Social Research, University of Michigan, 1970.

[38] Sonquist, J.A., Baker, E.L. e Morgan, J.N.. *Searching for Structure*. Ann Arbor: Institute for Social Research, University of Michigan, 1973.

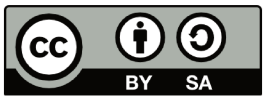
[39] Su, X. e Fan, J.. *Multivariate Survival Trees: A Maximum Likelihood Approach Based on Frailty Models*. Biometrics **60**, 01, 2004.

[40] Therneau, T.M. e Grambsch, P.M.. *Modeling Survival Data: Extending the Cox Model*. (New York: Springer-Verlag), 2000.

[41] Van Eck, A.N.. *Statistical Analysis and Data Management Highlights of OSIRIS IV*. The American Statistician **34**, 02, 1980.

Capítulo 6

Licença



6.1 Sobre a licença dessa obra

A licença Creative Commons Atribuição-CompartilhaIgual 3.0 Não Adaptada utilizada nessa obra diz que:

1. Você tem a liberdade de:

    - Compartilhar — copiar, distribuir e transmitir a obra;
    - Remixar — criar obras derivadas;

- fazer uso comercial da obra.

2. Sob as seguintes condições:

   - Atribuição — Você deve creditar a obra da forma especificada pelo autor ou licenciante (mas não de maneira que sugira que estes concedem qualquer aval a você ou ao seu uso da obra).
   - Compartilhamento pela mesma licença — Se você alterar, transformar ou criar em cima desta obra, você poderá distribuir a obra resultante apenas sob a mesma licença, ou sob uma licença similar à presente.