Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Modelos de Fragilidade com Aplicações em Análise de Ligação

Benilton de Sá Carvalho
Orientadora: Profa. Dra. Hildete Prisco Pinheiro

Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de Mestre em Estatística.

Campinas
2003

Modelos de Fragilidade com Aplicações em Análise de Ligação

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Benilton de Sá Carvalho e aprovada pela comissão julgadora.

Campinas, 28 de fevereiro de 2003.

Profa. Dra. Hildete Prisco Pinheiro
Orientadora

Banca Examinadora:

1. Profa. Dra. Hildete Prisco Pinheiro - IMECC/UNICAMP
2. Dra. Mariza de Andrade - Clínica Mayo, MN, EUA
3. Prof. Dr. Antônio Carlos Pedroso de Lima - IME/USP

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP.

C253m

Carvalho, Benilton de Sá

Modelos de fragilidade com aplicações em análise de ligação / Benilton de Sá Carvalho – Campinas, [S.P. :s.n.], 2003.

Orientador: Hildete Prisco Pinheiro

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Teste de sobrevivência acelerados. 2. Genética quantitativa. 3. Tábua de Mortalidade. I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Dissertação de Mestrado defendida em 28 de fevereiro de 2003 e aprovada pela Banca Examinadora composta pelos Profs. Drs.

Profa. Dra. Hildete Prisco Pinheiro

Dra. Mariza de Andrade

Prof. Dr. Antônio Carlos Pedroso de Lima

Resumo

Este trabalho apresenta o Modelo de Fragilidade com Risco Logístico, uma proposta de análise de sobrevivência agregada à análise de ligação. A estrutura do modelo trata-se de uma extensão do modelo de sobrevivência proposto por Cox. Assume-se uma forma paramétrica para a função risco, possuindo forma logística (Mackenzie, 1996). A censura é definida de acordo com a observação da idade do paciente no momento do diagnóstico da doença e a idade atual do indivíduo, desta forma, uma idade de diagnóstico maior que a idade atual caracteriza a censura. As fragilidades, construídas de forma aditiva a partir de densidades gama, incorporam contribuições genéticas e também fatores ambientais que influem no comportamento da característica de interesse (neste caso, uma doença).

O uso de tais técnicas de modo unificado tem se mostrado bastante eficaz, pois a idade de diagnóstico é geralmente coletada em estudos de mapeamento genético, além do que a mesma apresenta-se muitas vezes associada a doenças complexas. Conseqüentemente, tem-se um modelo versátil, capaz de análises nos moldes do modelo de Cox, além de prover meios para mapeamento genético. O modelo é ajustado pela maximização da verossimilhança retrospectiva (Whittemore, 1996), por meio de um algoritmo iterativo baseado nas equações de Kuhn-Tucker. A validade da proposta é averiguada por meio de análise de dados simulados com o software G.A.S.P. e comparações de resultados obtidos com análises conjuntas por meio dos softwares SAS e GeneHunter.

Abstract

This research work presents the Frailty Model using the logistic form to the hazard function, a survival analysis method which agregates linkage analysis role. The model structure is based on a Cox model extension. The hazard function has a parametric form, the logistic one (Mackenzie, 1996). The censure is defined using the current age and the age at onset, thus if one has the age at onset greater then the current age, then a censured observation is characterized. The additive frailties are constructed from gamma densities and incorporate genetic and environmental contributions to the trait of interest (in this case, a disease).

The use of these techniques in a unified way has been shown to be efficient, once the age at onset is usually collected in genetic mapping studies, moreover it's shown to be associated to complex diseases. Then, this is a useful model, which is able to be applied to Cox model situations and genetic mapping cases. The model is adjusted by the maximization of retrospective likelihood (Whittemore, 1996), using an iterative algorithm based on the Kuhn-Tucker equations. The proposed model is ascertained by the analysis of simulated data created using G.A.S.P. package and by comparisons with the results from the joint analyses with SAS and GeneHunter softwares.

Aos meus pais.

Agradecimentos

À Deus, por Sua companhia, indispensável para alcançar este objetivo.

Os mais sinceros agradecimentos aos meus pais que, incondicionalmente, abriram mãos de seus sonhos, para realização dos meus. Esta vitória pertence a vocês, pois é reflexo de seus esforços, que me permitiram um ensino de qualidade, fator essencial para a completa formação de um indivíduo.

Meu especial obrigado a Tatiana Benaglia, por sua companhia nos bons momentos e incentivo nos difíceis, sua compreensão e alegria. Suas contribuições acadêmicas, muitas delas involuntárias, foram muito importantes para o desenvolvimento deste trabalho; suas contribuições não-acadêmicas foram por demais valiosas para que eu não esquecesse que há vida fora da Universidade.

À Profa. Hildete Pinheiro, por sua ajuda contínua e, principalmente, por sua paciência em meus instantes de nervosismo. Muito obrigado, também, por suas orientações nos diversos âmbitos, acadêmicos ou não, que me ajudaram indubitavelmente na minha formação pessoal e acadêmica.

Aos Professores Mariza de Andrade e Antônio Carlos, por suas inúmeras contribuições e grande cooperação no decorrer deste trabalho, importantíssimas para o melhor desenvolvimento desta dissertação.

Ao Departamento de Estatística, especialmente aos Profs. Aluísio Pinheiro, Filidor Labra, Eliana e Mauro Marques, Luiz Hotta, Nancy Garcia e Ronaldo Dias, por suas contribuições, disponibilidade, orientações e amizade ao longo destes anos.

Aos Professores Sérgio F. dos Reis, Alberto Saa, Hongzhe Li, Gilbert MacKenzie, Alice Whittemore, Mike Miller e Mark Daly, por todas as contribuições diretas e indiretas, que propiciaram a conclusão deste trabalho.

Aos amigos Hugo Siqueira, Janaína Rossati, Juliana Francisco, Aline Mazzi, Artur Iuri, Cézar Anselmo, Laura Lupicínia, Fernando Ferrari e Dionysio Moriconi, que, mesmo seguindo caminhos distintos dos meus, sempre se mostraram presentes naqueles instantes em que a gente menos espera e mais necessita. Muito obrigado à Michelle (*in memorian*) por seu exemplo de vida e luta por seus ideais.

Àqueles que acompanharam meu cotidiano de forma bastante próxima, no dia-a-dia do LABEST-PÓS e também fora dele, Helder Palaro, Alexandre Rübesam, Ricardo Takeyama, Nelson Lopes, Samara Kiihl, Rodrigo Tsai, Clécio Ferreira, Rossana Lopez, Roberta de Souza e Lílian Hanamoto, que tornaram-se parte da minha família em Campinas.

À FAPESP, Fundação responsável pelo suporte financeiro, essencial para o desenvolvimento deste Projeto, por confiar a mim contribuições à Ciência do País.

"(...) Genética e Estatística, então, têm em comum o fato de que, cada um em seu campo, representa um ponto de vista distinto, o qual influencia profundamente nos processos intelectuais com os quais o trabalho científico é realizado. (...)"
Fisher, R.A., $Heredity$, 6 (1952).

Sumário

Lista de Tabelas

Lista de Figuras

Introdução

Vários estudos de doenças complexas incluem o estudo da variável idade de diagnóstico, como por exemplo os cânceres de mama e próstata, além do diabetes tipo I e o mal de Alzheimer. Entre os membros de uma família, é possível observar correlação entre as idade de diagnóstico dos mesmos; além disso, uma idade de diagnóstico precoce pode estar relacionada à mudança dos riscos relativos do indivíduo.

Uma especial atenção tem sido dada a extensões do modelo de Cox com efeitos aleatórios. Uma vasta literatura (Vaupel *et. al.*, 1979; Clayton and Cuzick, 1985; Oakes, 1989; Nielsen *et. al.*, 1982) pode ser encontrada a respeito dos modelos de fragilidade, capazes de considerar variáveis não-observáveis (como características genéticas).

Este trabalho apresenta uma proposta paramétrica para a realização de análises de sobrevivência e de ligação conjuntamente. A flexibilidade apresentada pela família logística generalizada (Mackenzie, 1996) justifica o emprego de tal forma funcional para a função risco utilizada em análise de sobrevivência. O Capítulo 1 é iniciado com conceitos básicos de análise de sobrevivência e estende-se até a formulação do modelo de risco logístico.

O modelo assume que a construção das fragilidades, apresentada de forma detalhada no Capítulo 2, é feita de forma aditiva por basicamente duas componentes: a genética e a ambiental. Neste mesmo capítulo, é apresentado um teste para ligação baseado na verossimilhança retrospectiva (Whittemore, 1996).

No Capítulo 3, é introduzida uma curta apresentação dos conceitos mais empregados em análise de ligação, seus propósitos e métodos, afim de que o leitor que não esteja familiarizado com tal método possa situar-se diante da proposta aqui discutida.

Diante da complexidade computacional enfrentada afim de efetuar a estimação dos parâmetros, o Capítulo 4 discute maiores detalhes do procedimento computacional empregado quando da maximização da razão de verossimilhança, sob a hipótese de inexistência de ligação. Deve-se enfatizar, também, que além da difícil tarefa de maximizar uma razão,

o trabalho computacional exigido por dados de família sempre constitui um agravante.

Dados de famílias nucleares com dois filhos foram gerados, usando o pacote G.A.S.P. (para geração dos marcadores) e o *software* MATLAB 6.1 (para geração de idades), sob duas condições: uma sob ligação e a outra sob a ausência de ligação. As análises referentes a estes conjuntos de dados simulados encontram-se no Capítulo 5. Deve-se ressaltar que os dados foram assim gerados (pares de irmãos - *sibpair data*) por motivos de simplicidade computacional.

Detalhes a respeito da Função de Sobrevivência Conjunta são apresentados no Anexo A. As informações aí contidas são essenciais para a extensão da atual proposta para casos mais realistas, como famílias com mais de dois filhos e heredogramas compostos por várias gerações. O Anexo B apresenta definições de termos não muito comuns em sua maioria extraídos do dicionário *on-line* MedicineNet.com (2002).

1 *Conceitos em Análise de Sobrevivência*

Aqui, serão apresentados aspectos básicos de análise sobrevivência, iniciando pela definição de censura e funções empregadas (densidade, sobrevivência e risco), finalizando com as relações existentes entre essas três funções.

A proposta deste trabalho é realizar análise de sobrevivência (para dados genéticos sob o contexto de fragilidade) utilizando-se a função risco na forma logística. Portanto, neste capítulo, serão apresentados sucintamente o modelo de riscos proporcionais de Cox, amplamente empregado e, em seguida, a construção da função risco na forma logística, que será associada ao modelo de sobrevivência posteriormente.

1.1 Conceitos Básicos em Análise de Sobrevivência

Tempo de sobrevivência pode ser definido como o tempo até a ocorrência de um determinado evento (que pode ser desenvolvimento de uma doença, resposta a um tratamento, morte ou nascimento, etc.). A ocorrência do referido evento é chamada *falha*. *Dados de sobrevivência* podem incluir tempo de sobrevivência, resposta a um tratamento e outras características de um paciente, por exemplo.

Uma característica importante em dados de sobrevivência é o fato de, após a finalização do estudo, ainda haver indivíduos para os quais o evento ainda não aconteceu. A essa ocorrência denomina-se *censura*. Conseqüentemente, para indivíduos censurados, o tempo exato de falha é desconhecido.

As censuras podem ser classificadas, de modo bastante simples, em três grupos:

1. **Censura à direita**: Tudo o que sabe-se a respeito de T, o tempo de ocorrência do

evento, é que ele é maior que um certo valor c. Geralmente, este tipo de censura é observado nos casos de perda de acompanhamento do indivíduo, término do estudo ou abandono do estudo.

2. **Censura à esquerda**: A informação sobre T é de que ele é menor que algum valor.

3. **Censura intervalar**: É uma combinação de censura à esquerda e à direita, de modo que a informação sobre T é que $a < T < b$ para algum a e b.

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

1. Censura Tipo I:

   Estudos com animais, habitualmente, iniciam-se com um número fixo de animais, aos quais aplicam-se tratamentos a serem analisados. Devido a fatores como custo e tempo disponível (considerando que o interesse seja o estudo do tempo de sobrevivência até a morte dos animais), o pesquisador não pode aguardar pela morte de todas as cobaias. Uma opção é fazer a observação por um tempo pré-fixado, após o qual os animais sobreviventes são sacrificados. Os tempos de sobrevivência para os animais que faleceram durante o período em que o experimento foi realizado são medidos a partir do início do estudo até o momento da morte da cobaia. Essas observações são ditas ser *não-censuradas*. Os tempos de sobrevivência para os animais sacrificados são desconhecidos e são anotados como, pelo menos, o tempo de duração do estudo. Essas são observações *censuradas*. Alguns animais perdem-se ou morrem acidentalmente. Seus tempos de sobrevivência, do início até a perda ou morte, também são censurados. Em estudos com Censura Tipo I, se não existem perdas acidentais, todos os indivíduos censurados têm seus tempos de sobrevivência iguais ao comprimento do período de estudo.

2. Censura Tipo II:

   Outra opção em estudos com animais é aguardar até que uma porção fixa dos animais falhe; após isso, os animais sobreviventes são sacrificados. Nesse caso, se não há perdas acidentais, as observações censuradas são iguais a maior observação não-censurada.

3. Censura Tipo III:

Na maioria de 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 sobreviveram até o fim do estudo, os tempos de sobrevivência são, pelo menos, o período entre suas chegadas até 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 Tipo III, visto que a resposta de interesse é a idade de diagnóstico de doença observada em estudos clínicos.

Uma característica muito importante, amplamente empregada em trabalhos de análise de sobrevivência, incluindo este, é a hipótese de censura não-informativa. Se o instante de censura é definido como C e o tempo de sobrevivência é T, então um indivíduo é censurado se $T > C$ e não-censurado caso contrário. Uma condição suficiente para que a censura seja não-informativa é T ser independente de 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, vista 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.

1.2 Funções do Tempo de Sobrevivência

Tempos de sobrevivência são dados que medem o tempo até a ocorrência de um determinado evento (falha, morte, desenvolvimento de uma doença). Esses tempos estão sujeitos a variações aleatórias e, portanto, possuem uma distribuição de probabilidade associada. A distribuição dos tempos de sobrevivência é caracterizada por três funções matematicamente equivalentes:

- função de sobrevivência,
- função densidade de probabilidade 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.

Seja T o tempo até a ocorrência de um determinado evento. A distribuição dessa variável aleatória pode ser descrita por alguma das três funções que seguem:

1. **Função de Sobrevivência**: Definida como a probabilidade de um indivíduo sobreviver mais que t unidades de tempo, é denotada por $S(t)$. Assim:

$$\begin{aligned} S(t) &= P(T > t) = 1 - F(t) \\ S(0) &= 1 \\ S(\infty) &= 0 \end{aligned} \tag{1.1}$$

2. **Função Densidade** (Taxa Incondicional de Falha): Definida como o limite da probabilidade de um indivíduo *falhar* no pequeno intervalo $(t,\ t + \Delta t)$ por unidade Δt de tempo, é expressa como:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(\text{um indivíduo } \textit{falhar} \text{ no intervalo } (t, t + \Delta t))}{\Delta t}$$

3. **Função Risco** (Taxa Condicional de Falha): Definida como a probabilidade de falha num pequeno intervalo de tempo, dado que o indivíduo sobreviveu até o início do

intervalo, é expressa por:

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \end{aligned} \tag{1.2}$$

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

1.2.1 Equivalência entre Funções

1. De (1.1) e (1.2), tem-se que:
$$\lambda(t) = \frac{f(t)}{S(t)} \tag{1.3}$$

2. Usando a relação entre densidade e distribuição, obtém-se:
$$f(t) = -S'(t) \tag{1.4}$$

3. Empregando (1.4) em (1.3), observa-se:
$$\lambda(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log_e S(t) \tag{1.5}$$

4. Resolvendo a equação diferencial acima,
$$S(t) = \exp\left[-\int_0^t \lambda(x)\, dx\right] \tag{1.6}$$

5. De (1.6) e (1.3),
$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(x)\, dx\right] \tag{1.7}$$

6. Define-se o risco acumulado como:
$$\Lambda(t) = \int_0^t \lambda(x)\, dx \tag{1.8}$$

O risco acumulado pode ser interpretado como a probabilidade de falha no instante t dada a sobrevivência até t.

Maiores detalhes a respeito de conceitos e aplicações de análise de sobrevivência podem ser obtidos em Lee (1992).

1.3 Modelo de Riscos Proporcionais

Uma aproximação padrão aplicada em estudos médicos para a análise de dados de sobrevivência é o emprego de modelos de riscos proporcionais de Cox e Oakes (1972), em que a função risco tem uma forma multiplicativa, como segue:

$$\lambda(t_i|\mathbf{x}_i) = \lambda_0(t_i)\exp(\mathbf{x}_i\boldsymbol{\beta}),$$

na qual,

$\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})$ é um vetor linha de covariáveis medidas para o i-ésimo indivíduo;

$\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ é um vetor de parâmetros desconhecidos a serem estimados e

$\lambda_0(t_i)$ é uma função risco arbitrária.

Então, o risco é log-linear nos parâmetros da regressão e uma característica especial desse modelo é o uso da verossimilhança parcial para $\boldsymbol{\beta}$, derivada tratando λ_0 como um parâmetro de penalização. Tal aproximação permite estimar $\boldsymbol{\beta}$ independente da forma funcional de $\lambda_0(t)$, uma generalização muito bem sucedida.

1.4 Formulação do Modelo de Risco Logístico

1.4.1 Densidade de Referência

Seja T uma variável aleatória não negativa representando o tempo de falha e seja a taxa instantânea de falha, ou risco, definido como:

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

Então, o risco de referência tem a seguinte forma funcional, conforme Mackenzie (1996):

$$\lambda_0(t|\zeta, \alpha, \gamma) = \frac{\zeta \exp(t\alpha + \gamma)}{1 + \exp(t\alpha + \gamma)}, \tag{1.9}$$

em que $\zeta > 0$ e $\alpha, \gamma \in \mathbb{R}$. Então, a expressão (1.9) caracteriza uma família de três parâmetros de variáveis aleatórias não-negativas com densidade de referência dada por:

$$f_0(t|\zeta,\alpha,\gamma) = \lambda_0(t|\zeta,\alpha,\gamma) S_0(t|\zeta,\alpha,\gamma), \tag{1.10}$$

na qual $S_0(t|\zeta,\alpha,\gamma)$ é a função de sobrevivência definida por:

$$\begin{aligned} S_0(t|\zeta,\alpha,\gamma) &= \exp\left\{-\int_0^t \lambda_0(u|\zeta,\alpha,\gamma)\, du\right\} \\ &= e^{-\Lambda_0(t)} \end{aligned} \tag{1.11}$$

Desenvolvendo $\Lambda_0(t)$, obtém-se:

$$\begin{aligned} \Lambda_0(t) &= \zeta \int_0^t \frac{e^{u\alpha+\gamma}}{1+e^{u\alpha+\gamma}}\, du \\ &= \frac{\zeta}{\alpha} \ln\left(1+e^{u\alpha+\gamma}\right)|_0^t \\ &= \frac{\zeta}{\alpha} \ln\left(\frac{1+e^{t\alpha+\gamma}}{1+e^{\gamma}}\right) \\ &= \ln\left(\frac{1+e^{t\alpha+\gamma}}{1+e^{\gamma}}\right)^{\frac{\zeta}{\alpha}} \end{aligned} \tag{1.12}$$

Portanto, aplicando (1.12) em (1.11), tem-se:

$$S_0(t|\zeta,\alpha,\gamma) = \left(\frac{1+e^{t\alpha+\gamma}}{1+e^{\gamma}}\right)^{-\frac{\zeta}{\alpha}} \tag{1.13}$$

Assim, empregando (1.13) em (1.10), conclui-se que:

$$f_0(t|\zeta,\alpha,\gamma) = \frac{\zeta \exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)} \left\{\frac{1+\exp(t\alpha+\gamma)}{1+\exp(\gamma)}\right\}^{-\frac{\zeta}{\alpha}}, \tag{1.14}$$

para $t > 0$, em que $S_0(0|\zeta,\alpha,\gamma) = 1$ e $S_0(\infty|\zeta,\alpha,\gamma) = 0$, para $\alpha > 0$. A densidade é imprópria para $\alpha < 0$ e a função de sobrevivência de referência $S_0(\infty|\zeta,\alpha,\gamma)$ tende ao limite positivo

$$a_0(\zeta,\alpha,\gamma) = \frac{1}{(1+\exp\gamma)^{\zeta/\alpha^*}} \quad ,$$

com $\alpha^* = |\alpha|$, pois

$$\begin{aligned}\lim_{t\to\infty,\ \alpha<0} S_0(t|\zeta,\alpha,\gamma) &= \lim_{t\to\infty,\ \alpha<0}\left(\frac{1+e^{t\alpha+\gamma}}{1+e^{\gamma}}\right)^{-\frac{\zeta}{\alpha}} \\ &= \left(\frac{1+\lim\limits_{t\to\infty,\ \alpha<0} e^{t\alpha+\gamma}}{1+e^{\gamma}}\right)^{\frac{\zeta}{|\alpha|}} \\ &= \left(\frac{1}{1+e^{\gamma}}\right)^{\frac{\zeta}{|\alpha|}}\end{aligned}$$

Dado que este limite é positivo, tem-se que a função distribuição associada não converge para a unidade. Ou seja, a função densidade não integra 1, o que indica que ela é imprópria. Assim, é importante a leitura da Seção 1.5, na qual são apresentados maiores detalhes a respeito de variáveis aleatórias impróprias.

A equação (1.14) é, a partir de agora, chamada *distribuição logística generalizada dependente do tempo*.

1.4.2 Família Reduzida

Quando $\zeta = 1$, uma família biparamétrica de variáveis aleatórias não-negativas é obtida. Nela, a função risco de referência é dada por:

$$\lambda_0^*(t|\alpha,\gamma) = \frac{\exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)}, \tag{1.15}$$

que é a função logística no tempo. Este modelo reduzido é de interesse, pois o logito da função é linear no tempo:

$$\psi_0^*(t) = \log\left\{\frac{\lambda_0^*(t)}{1-\lambda_0^*(t)}\right\} = t\alpha+\gamma. \tag{1.16}$$

A equação (1.16) pode ser vista como uma forma de modelar uma seqüência de probabilidades dependentes do tempo.

1.4.3 Algumas Propriedades

Observe que se $\zeta = 1 + e^{t\alpha+\gamma}$, i.e. $\zeta > 1$, então a expressão (1.9) torna-se:

$$\begin{aligned} \lambda_0(t|\zeta,\alpha,\gamma) &= \frac{\zeta \exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)} \\ &= \exp(t\alpha+\gamma), \end{aligned} \tag{1.17}$$

que corresponde à função risco para uma distribuição de Gompertz.

Se $e^{t\alpha+\gamma}$ é suficientemente grande, então $\frac{\exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)} \to 1$ e

$$\begin{aligned} \lambda_0(t|\zeta,\alpha,\gamma) &= \frac{\zeta \exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)} \\ &= \zeta, \end{aligned} \tag{1.18}$$

que caracteriza uma distribuição exponencial dos tempos de sobrevivência.

Deste modo, as funções risco para o modelo logístico generalizado são muito flexíveis, no sentido de que podem tomar diferentes formas.

1.5 Variáveis Aleatórias Impróprias

Geralmente, assume-se implicitamente que o evento de interesse é obrigado a ocorrer, de modo que $S(\infty) = 0$, o que implica no fato de que o risco acumulado deve divergir, isto é $\Lambda(\infty) = \infty$. Intuitivamente, o evento ocorrerá com certeza se o risco acumulado durante um longo período for suficientemente alto.

É necessário observar que há eventos de interesse que podem não ocorrer, como é o caso aqui discutido: uma pessoa pode nunca ser diagnosticada como portadora da doença. Outros exemplos de eventos como este são pessoas que nunca se casam e pessoas que nunca mudarão de emprego, pois estão satisfeitas com a atual situação.

Uma opção para análise é observar que as funções risco e sobrevivência ainda podem ser calculadas e estas ainda são bem definidas mesmo com a não obrigatoriedade de ocorrência do evento, conforme Rodriguez (2001). Por exemplo, estuda-se uma determinada doença para toda uma população (o que significa que nesse grupo há pessoas já diagnosticadas, outras que serão diagnosticadas e outras que nunca serão diagnosticadas afetadas), caracterizando-se a

censura por meio do uso da idade de diagnóstico e idade atual (se a idade de diagnóstico for maior que a idade atual, tem-se a censura). Nesse caso, $S(t)$ representa a proporção de pessoas com t anos de idade não diagnosticadas e $S(\infty)$ representaria a proporção de pessoas que nunca serão diagnosticadas.

Uma limitação deste caminho é que se o evento não é obrigado a acontecer, então o tempo de espera T pode ser não definido (ou infinito) e, então, não caracterizaria uma variável aleatória própria. Sua densidade, calculada a partir das funções risco e sobrevivência, seria imprópria, pois não integraria a unidade. É direto observar que o tempo de espera não seria definido. No exemplo citado, é impossível calcular a idade média de diagnóstico, pois nem todas as pessoas são diagnosticadas. Tal limitação não se mostra um problema, pois é possível calcular a idade mediana de diagnóstico.

2 *Modelo de Sobrevivência com Fragilidade Genética*

A seguir, são apresentados os modelos multivariados de sobrevivência induzidos por fragilidades genéticas. Tais modelos são aplicáveis em análise de ligação, como Li e Zhong (2002).

Modelos de fragilidades são modelos de efeitos aleatórios delineados para trabalhar com dados de sobrevivência censurados, nos quais a diferença entre grupos homogêneos é modelada pelo acréscimo de um fator não-observável na função risco.

Este capítulo apresenta detalhes da abordagem de fragilidade, estendendo-se da sua construção até sua inclusão na função risco, de forma multiplicativa. O modelo de sobrevivência para idade de diagnóstico de doença é apresentado em seguida, quando define-se a função conjunta de densidade e sobrevivência (densidade para indivíduos afetados e sobrevivência para censurados). A função conjunta de densidade e sobrevivência caracteriza a verossimilhança retrospectiva a ser empregada na estimação dos parâmetros. Define-se uma medida de *Lod Score*, a ser empregada na análise de ligação (ferramenta utilizada em mapeamento genético), como função da razão de verossimilhança.

2.1 Modelo de Fragilidade Aditiva Gama

2.1.1 Construção de Fragilidades Genéticas para Famílias

Seja uma irmandade com n irmãos e denote seus pais por F para pai e M para mãe. Assumindo-se independência entre pai e mãe, existem apenas quatro alelos distintos por descendência para um dado *locus*. Suponha uma série de marcadores numa região cromossômica suspeita de ter o *locus*/*loci* da doença em estudo. Se d é um ponto nessa região, deseja-se

saber quando há um gene susceptível à doença (DS) no *locus d*. Designam-se os cromossomos paternos que contêm o *locus* de interesse por $(1, 2)$ e os maternos, por $(3, 4)$. O vetor de herança alélica de uma irmandade no *locus d* é o vetor

$$A_d = (a_1, a_2, \ldots, a_{2j-1}, a_{2j}, \ldots, a_{2n-1}, a_{2n}),$$

no qual $a_{2j-1} = 1$ ou 2 e $a_{2j} = 3$ ou 4, isto é, os índices ímpares indicam a herança paterna e os pares, a materna. O vetor de herança indica que partes do genoma no *locus d* são transmitidas para os n filhos a partir dos pais.

É importante definir o conceito de IBD_d(*identical by descendent - idêntico por descendência*), que indica o número de alelos compartilhados por grupos de indivíduos no locus d, conforme Andrade e Pinheiro (2002). Aqui, os grupos de indivíduos são pares de irmãos. Desta forma, eles podem compartilhar 0, 1 ou 2 alelos, conforme a Figura 2.1.

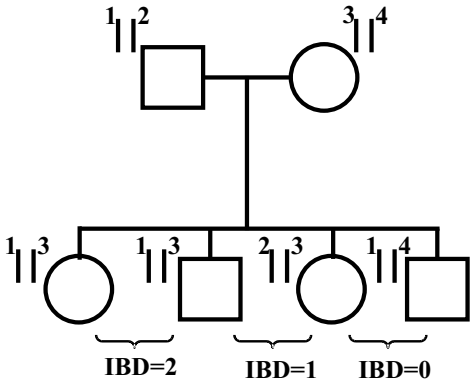

Figura 2.1: Identidade por Descendência

Definem-se as fragilidades genéticas devidas ao *locus d* para o pai e mãe como

$$\begin{cases} Z_{dF} = U_{d1} + U_{d2} \\ Z_{dM} = U_{d3} + U_{d4}, \end{cases}$$

nas quais U_{d1} e U_{d2} são as fragilidades genéticas devidas a parte do genoma nos dois cromossomos paternos no *locus d*; U_{d3} e U_{d4} possuem interpretação análoga, só que para a mãe. Além disso, assume-se que as fragilidades paternas sejam independentes das maternas.

Para um dado vetor de herança v_d no *locus d* para uma irmandade, define-se a fragilidade para o j-ésimo parente como

$$Z_{dj} = U_{da_{2j-1}} + U_{da_{2j}}; \;\; j = 1, \ldots, n.$$

Assume-se que U_{d1}, U_{d2}, U_{d3} e U_{d4} são independentes identicamente distribuídos segundo uma distribuição Gama, $\Gamma(v_d/2, \eta)$, para a qual η é o inverso do parâmetro de escala e v_d é o parâmetro de forma. Então,

$$Z_{dj} \sim \Gamma(v_d, \eta); \quad j = 1, \ldots, n.$$

Contribuições genéticas para a doença não devidas ao único *locus d* da doença (por exemplo, devidas a *loci* não ligados a d) ou contribuições para efeitos familiares compartilhados são consideradas pela adição de outro termo aleatório de fragilidade, U_p, à fragilidade genética e, então, define-se a fragilidade genética para o j-ésimo indivíduo como:

$$\begin{aligned} Z_j &= Z_{dj} + U_p \\ &= U_{da_{2j-1}} + U_{da_{2j}} + U_p, \end{aligned}$$

na qual $U_p \sim \Gamma(v_p, \eta)$ em diferentes famílias. Portanto, $Z_j \sim \Gamma(v_d + v_p, \eta)$. Então, as médias das fragilidades são

$$E(Z_1) = E(Z_2) = \cdots = E(Z_n) = \frac{v_p + v_d}{\eta}$$

e as variâncias

$$V(Z_1) = V(Z_2) = \cdots = V(Z_n) = \frac{v_p + v_d}{\eta^2}.$$

Assim, o parâmetro v_d indicará a proporção de variância da fragilidade genética explicada pelo *locus d*, visto que, para identificabilidade do modelo, a restrição $v_d + v_p = \eta$ será empregada, como indicado na Seção 2.1.2. Deste modo, a variância da fragilidade será $1/(v_d + v_p)$ e v_p será interpretado como a porção de variância devida ao *locus d*.

As fragilidades para uma irmandade podem ser escritas de forma matricial,

$$\mathbf{Z} = \mathbf{HU}, \tag{2.1}$$

expressão na qual

$$\begin{aligned} \mathbf{Z} &= (Z_1, Z_2, \ldots, Z_n)' \\ \mathbf{H} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & 1 \end{pmatrix} \\ \mathbf{U} &= (U_{d1}, U_{d2}, U_{d3}, U_{d4}, U_p)', \end{aligned}$$

com

$$\begin{aligned} a_{j1} &= I(a_{2j-1} = 1) \\ a_{j2} &= I(a_{2j-1} = 2) \\ a_{j3} &= I(a_{2j} = 3) \\ a_{j4} &= I(a_{2j} = 4); \quad j = 1, \ldots, n, \end{aligned}$$

tal que

Z representa o vetor de fragilidade para os n irmãos,

H é a matriz indicadora de alelos herdados pelos indivíduos em questão,

U é o vetor das fragilidades genéticas e ambiental e

$I(\mathcal{A})$ é definida como a função indicadora de $\mathcal{A}$ (ou seja, $I(\mathcal{A}) = 1$ se o evento $\mathcal{A}$ ocorre e $I(\mathcal{A}) = 0$, caso contrário).

2.1.2 Modelo Genético de Fragilidade Aditiva Gama

Seja uma irmandade com n irmãos, T_j a variável aleatória idade de diagnóstico da doença para o j-ésimo irmão e (t_j, δ_j) o conjunto de dados, no qual t_j é a idade de diagnóstico da doença (se $\delta_j = 1$) ou idade de censura (se $\delta_j = 0$). Assume-se que a função risco de desenvolvimento de doença para o j-ésimo indivíduo com idade t_j seja modelada segundo o modelo de riscos proporcionais com efeito aleatório Z_j,

$$\lambda_j(t|Z_j) = \lambda_0(t) e^{\mathbf{X}_j' \boldsymbol{\beta}} Z_j; \quad j = 1, \ldots, n, \tag{2.2}$$

para o qual λ_0 é a função base de risco, $\mathbf{X}_j$ é um vetor de covariáveis observadas para o j-ésimo indivíduo e $\boldsymbol{\beta}$ é um vetor de parâmetros de regressão associado às covariáveis, Z_j é a fragilidade não observada construída por (2.1). Como $Z_1, \ldots, Z_n$ são dependentes devido à segregação genética e fragilidade compartilhada, então $T_1, \ldots, T_n$ também o são. Para que o modelo seja identificável, considera-se $v_d + v_p = \eta$, então $E(Z_j) = 1$, $j = 1, \ldots, n$.

2.2 Modelo de Sobrevivência e Idade de Diagnóstico

2.2.1 Modelo de Sobrevivência e Fragilidade Genética

Assumindo independência condicional ($T_j|Z_j$ são independentes) e baseando-se no modelo (2.2), observa-se que condicionando-se no vetor de fragilidade $\mathbf{Z}$, a função conjunta de sobrevivência pode ser escrita como

$$S(t_1,\ldots,t_n|Z_1,\ldots,Z_n) = e^{-\Lambda_1(t_1)Z_1-\cdots-\Lambda_n(t_n)Z_n},$$

na qual $\Lambda_j(t_j) = \Lambda_0(t_j)e^{\mathbf{X}_j'\boldsymbol{\beta}}; \quad j = 1,\ldots,n.$

Integrando-se em $Z_1,\ldots,Z_n$, obtém-se a função de sobrevivência conjunta marginal, dada por (conforme demonstrado no Apêndice):

$$\begin{aligned} S(t_1,\ldots,t_n) &= \left\{\prod_{i=1}^{4}\frac{\eta^{v_d/2}}{\left[\sum_{j=1}^{n}\Lambda_j(t_j)a_{ji}+\eta\right]^{v_d/2}}\right\}\times \\ &\quad\times\left\{\frac{\eta^{v_p}}{\left[\sum_{j=1}^{n}\Lambda_j(t_j)+\eta\right]^{v_p}}\right\}. \end{aligned} \tag{2.3}$$

Na prática, as observações são geralmente censuradas e, então, necessita-se tanto da função de sobrevivência quanto de combinações das funções densidades e sobrevivência. Para uma irmandade com a irmãos afetados ($j = 1,\ldots,a$) e $n-a$ não afetados, a função conjunta de densidade e sobrevivência é

$$P(t_1,\delta_1=1,\ldots,t_a,\delta_a=1,t_{a+1},\delta_{a+1}=0,\ldots,t_n,\delta_n=0) = (-1)^a\frac{\partial^a S(t_1,\ldots,t_n)}{\partial t_1\ldots\partial t_a}.$$

Para o caso de todos os irmãos não afetados, i.e. $a = 0$, emprega-se a própria função de sobrevivência conjunta marginal, pois a função densidade é apenas para casos nos quais há censura.

Para famílias com todos os irmãos afetados, a densidade conjunta é:

$$P(t_1,\delta_1=1,\ldots,t_n,\delta_n=1) = (-1)^n\frac{\partial^n S(t_1,\ldots,t_n)}{\partial t_1\ldots\partial t_n}.$$

2.2.2 Caso Bivariado

Para uma irmandade com dois irmãos (pares de irmãos), deriva-se a função conjunta de densidade e sobrevivência para um par de irmãos que compartilham 0, 1 e 2 alelos idênticos por descendência (IBD) no *locus d*. Essas funções conjuntas são apresentadas na Tabela 2.1[1], na qual empregam-se as notações $\Lambda_j^* = \Lambda_j(t_j) + \eta$, $j = 1, 2$ e $\Lambda_{12} = \Lambda_1 + \Lambda_2 + \eta$ para simplificação das expressões. Observa-se que, quando $v_d = 0$, a função conjunta de sobrevivência não depende do número de alelos IBD no *locus d*, sugerindo que não há ligação entre a doença e o *locus d*.

Tabela 2.1: Função Conjunta de Densidade e Sobrevivência - Caso Bivariado

Função Conjunta de Densidade e Sobrevivência			
$P(t_1, \delta_1 = 0, t_2, \delta_2 = 0) = S(t_1, t_2)$			
$P(t_1, \delta_1 = 1, t_2, \delta_2 = 0) = C_1(t_1, t_2)\lambda_1(t_1)S(t_1, t_2)$			
$P(t_1, \delta_1 = 0, t_2, \delta_2 = 1) = C_2(t_1, t_2)\lambda_2(t_2)S(t_1, t_2)$			
$P(t_1, \delta_1 = 1, t_2, \delta_2 = 1) = [C_1(t_1, t_2)C_2(t_1, t_2) + C(t_1, t_2)]\,\lambda_1(t_1)\lambda_2(t_2)S(t_1, t_2)$			
	$IBD_d = 0$	$IBD_d = 1$	$IBD_d = 2$
$S(t_1, t_2)$	$\left(\frac{\eta^2}{\Lambda_1^*\Lambda_2^*}\right)^{v_d}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$	$\left(\frac{\eta^3}{\Lambda_1^*\Lambda_2^*\Lambda_{12}}\right)^{v_d/2}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$	$\left(\frac{\eta}{\Lambda_{12}}\right)^{v_d+v_p}$
$C_1(t_1, t_2)$	$\frac{v_d}{\Lambda_1^*} + \frac{v_p}{\Lambda_{12}}$	$\frac{v_d/2}{\Lambda_1^*} + \frac{v_d/2+v_p}{\Lambda_{12}}$	$\frac{v_d+v_p}{\Lambda_{12}}$
$C_2(t_1, t_2)$	$\frac{v_d}{\Lambda_2^*} + \frac{v_p}{\Lambda_{12}}$	$\frac{v_d/2}{\Lambda_2^*} + \frac{v_d/2+v_p}{\Lambda_{12}}$	$\frac{v_d+v_p}{\Lambda_{12}}$
$C(t_1, t_2)$	$\frac{v_p}{\Lambda_{12}^2}$	$\frac{v_d/2+v_p}{\Lambda_{12}^2}$	$\frac{v_d+v_p}{\Lambda_{12}^2}$

2.2.3 A Função Razão de Risco Condicional

O risco de recorrência, λ_s, é definido como a razão da probabilidade de desenvolvimento da doença para o l-ésimo irmão dado que o m-ésimo foi afetado e a probabilidade de desenvolvimento da doença na população.

Esse parâmetro é importante na determinação do poder do método "par de irmãos afetados"(*ASP - affected sib-pair*). Deve-se observar que para doenças com idades de início e penetrância dependente de idade, λ_s ignora tanto a idade atual do indivíduo l como a idade de início da doença do irmão m. Para tais doenças, uma medida de agregação familiar adequada é a razão de riscos condicionais. Considere um par de irmãos (l, m). Sejam (T_l, T_m)

[1]Demonstrações na Seção A.2.

as variáveis aleatórias idade de início de doença para l e m, define-se

$$\phi(t_l, t_m) = \frac{\lambda(t_l|T_m = t_m)}{\lambda(t_l|T_m > t_m)}$$

como a razão de risco condicional, na qual $\lambda(t_l|T_m = t_m)$ é a probabilidade instantânea de o l-ésimo irmão ter a doença com idade t_l, dado que o m-ésimo irmão está afetado com idade t_m e $\lambda(t_l|T_m > t_m)$ define-se similarmente, dado que o m-ésimo irmão está são na idade t_m.

$$\begin{aligned}
\phi(t_l, t_m) &= \frac{P(t_l, \delta_l = 0, t_m, \delta_m = 0)P(t_l, \delta_l = 1, t_m, \delta_m = 1)}{P(t_l, \delta_l = 0, t_m, \delta_m = 1)P(t_l, \delta_l = 1, t_m, \delta_m = 0)} \\
&= \frac{C_l(t_l, t_m)C_m(t_l, t_m) + C(t_l, t_m)}{C_l(t_l, t_m)C_m(t_l, t_m)} \\
&= 1 + \frac{C(t_l, t_m)}{C_l(t_l, t_m)C_m(t_l, t_m)}, \qquad (2.4)
\end{aligned}$$

na qual $C_l(t_l, t_m)$, $C_m(t_l, t_m)$ e $C(t_l, t_m)$ encontram-se definidos na Tabela 2.1 para pares de irmãos que compartilham 0, 1 e 2 alelos IBD no *locus d*. Deve-se observar que quando $v_d = 0$, $\phi(t_l, t_m)$ não depende do número de alelos IBD no *locus d*.

2.3 Verossimilhança e Análise de Ligação

2.3.1 Teste da Razão de Verossimilhança Retrospectiva

O modelo de sobrevivência proposto pode ser usado para construir um teste baseado em verossimilhança para análise de ligação. Conforme a Tabela 2.1, quando $v_d = 0$, a função razão de risco condicional (2.4) entre um par de irmãos, a densidade conjunta e função de sobrevivência para uma irmandade não dependem do número de alelos IBD no *locus d* ou do vetor de herança no mesmo *locus*. Conseqüentemente, o teste de ligação entre o *locus d* e a doença pode ser feito testando-se $H_0 : v_d = 0$.

Sejam a i-ésima irmandade com n_i irmãos e $(t_i, \delta_i) = (t_{i1}, \delta_{i1}, \ldots, t_{in_i}, \delta_{in_i})$ a idade de início da doença ou censura. Considera-se, também, um marcador M_i para a i-ésima irmandade. Os dados (M_i, t_i, δ_i) podem ser tratados na verossimilhança retrospectiva da informação marcadora M_i condicionada nos fenótipos (t_i, δ_i), como em Whittemore (1996). Uma vantagem de usar a verossimilhança retrospectiva é que a estatística do teste daí provinda é livre de vícios, se as famílias são corrigidas por seus fenótipos.

A verossimilhança retrospectiva para a i-ésima irmandade é

$$\begin{aligned} L_i(v_d, v_p, \Lambda_0(t), \boldsymbol{\beta}) &= P(M_i|t_i, \delta_i) \\ &= \frac{\sum_{a_d} P(t_i, \delta_i|\ A_d = a_d)P(A_d = a_d|M_i)}{\sum_{a_d} P(t_i, \delta_i|A_d = a_d)P(A_d = a_d)} P(M_i), \end{aligned}$$

na qual $P(t_i, \delta_i|A_d = a_d)$ é apresentada no Apêndice para o caso bivariado sob a notação $P(t_i, \delta_i|IBD = k)$, $P(A_d = a_d)$ é a probabilidade *a priori* do vetor de herança A_d e $P(A_d = a_d|M_i)$ pode ser calculada usando métodos multiponto, segundo apresentado em Kruglyak (1996). Para dados de pares de irmãos, a verossimilhança retrospectiva é:

$$\begin{aligned} L_i(v_d, v_p, \Lambda_0(t), \boldsymbol{\beta}) &= P(M_i|t_i, \delta_i) \\ &= \frac{\sum_{k=0}^{2} P(t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}|IBD_d = k)P(IBD_d = k|M_i)}{\sum_{k=0}^{2} P(t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}|IBD_d = k)P(IBD_d = k)} \\ &\times \ P(M_i), \end{aligned}$$

na qual $P(t_{i1}, \delta_{i1}, t_{i2}, \delta_{i2}|IBD_d = k)$ é dada na Tabela 2.1 e $P(IBD_d = k)$ é a probabilidade *a priori* de um par de irmãos compartilhar k alelos IBD.

Essa função de verossimilhança depende apenas do risco acumulado e quando $v_d = 0$, $L_i(0, v_p, \Lambda_0(t)) = P(M_i)$, então a estatística da razão de verossimilhanças é dada por

$$LR_i(v_d, v_p, \boldsymbol{\beta}) = \frac{\sum_{a_d} P(t_i, \delta_i|A_d = a_d)P(A_d = a_d|M_i)}{\sum_{a_d} P(t_i, \delta_i|A_d = a_d)P(A_d = a_d)}.$$

Assumindo que existem K famílias, define-se uma medida de Lod (*logarithm of odds*), vide Olson, Witte e Elston (1999), no *locus* d como

$$Lod_d = \max_{v_d, v_p, \beta} \sum_{i=1}^{K} \log_{10} LR_i(v_d, v_p, \boldsymbol{\beta}). \tag{2.5}$$

A medida de Lod escore é empregada em análise de ligação, com o objetivo de realizar o mapeamento genético. Sua construção, uso e maiores detalhes sobre análise de ligação são apresentados no Capítulo 3. Baseando-se na teoria de testes de razão de verossimilhanças quando a hipótese nula encontra-se na fronteira do espaço espaço paramétrico, conforme Self e Liang (1987), tem-se que, sob H_0, $2Lod_d \ln(10)$ distribui-se de acordo com uma mistura de igual probabilidade de massa pontual em zero e uma qui-quadrado com um grau de liberdade.

2.4 Modelo de Fragilidade com Risco Logístico

A proposta de modelo de fragilidade com risco logístico é construída aplicando a função (1.15) em (2.2), o que resulta num caso paramétrico de (2.2). A função risco acumulado, representada por $\Lambda_j(t_j)$, empregada na construção da função de sobrevivência (2.3) deve ser substituída pela expressão (1.12), resultando:

$$\lambda_j(t_j|Z_j) = \lambda_0(t)e^{\mathbf{X}_j'\boldsymbol{\beta}}Z_j; \quad j = 1, \ldots, n,$$

para a qual,

$$\lambda_0(t|\alpha, \gamma) = \zeta\frac{\exp(t\alpha + \gamma)}{1 + \exp(t\alpha + \gamma)},$$

e

$$\begin{aligned} S(t_1, \ldots, t_n) &= \left\{\prod_{i=1}^{4} \frac{\eta^{v_d/2}}{\left[\sum_{j=1}^{n}\Lambda_j(t_j)a_{ji} + \eta\right]^{v_d/2}}\right\} \times \\ &\quad \times \left\{\frac{\eta^{v_p}}{\left[\sum_{j=1}^{n}\Lambda_j(t_j) + \eta\right]^{v_p}}\right\}. \end{aligned} \tag{2.6}$$

em que,

$$\Lambda_0(t_j) = \ln\left(\frac{1 + e^{t_j\alpha+\gamma}}{1 + e^{\gamma}}\right)^{\frac{\zeta}{\alpha}} \text{ e} \tag{2.7}$$

$$\Lambda_j(t_j) = \Lambda_0(t_j)e^{\mathbf{X}_j'\boldsymbol{\beta}}. \tag{2.8}$$

O fato de a função risco representar a probabilidade instantânea de falha justifica a proposta de utilização de uma função risco na forma logística, pois função logística é muito empregada na modelagem de probabilidades.

Todas as características deste modelo são análogas às características apresentadas desde o início deste capítulo.

3 Análise de Ligação

Este capítulo apresenta conceitos de análise de ligação. Estão inclusas descrições do mecanismo de recombinação genética e construção de teste de ligação.

3.1 Conceitos de Análise de Ligação

Cromossomos homólogos segregam de modo independente. Alelos para *locus* de um mesmo cromossomo podem co-segregar para uma razão relacionada com a distância entre eles no cromossomo. Tal razão é a probabilidade do evento recombinante ocorrer entre os dois *loci* ou fração de recombinação, denotada θ.

A fração de recombinação varia entre zero (quando os *loci* estão muito próximos) e 0,5 (quando estão muito distantes ou em cromossomos diferentes). Portanto, pode ser empregada como uma medida de distância genética, bastante funcional para pequenas distâncias. Entretanto, a fração de recombinação não é uma medida de distância aditiva (devido à possibilidade de ocorrência de múltiplos *crossing overs*). A unidade de medida da fração de recombinação é 1 unidade map $\equiv$ 1 centiMorgam (cM), cerca de 1% de fração de recombinação.

Dois *loci* são ditos geneticamente ligados quando $\theta \approx 0$. O objetivo da análise de ligação é estimar θ e testá-lo contra a hipótese $H_1 : \theta < 0,50$. A estimativa da fração de recombinação, $\hat{\theta}$, é a proporção de recombinações (proporção de indivíduos que possuem um cromossomo recombinado) em todas as oportunidades para recombinação e, inicialmente, varia no intervalo $[0, 1]$. Entretanto, a estimativa de máxima verossimilhança é definida no conjunto de valores admissíveis do parâmetro, assim não excede 0,50, de acordo com Olson, Witte e Elston (1999). Quando há um *crossing over*, metade dos gametas resultantes continuam sendo não recombinantes. Portanto, se um *crossing over* ocorre a cada meiose, metade dos gametas continuam originais. Assim θ pode alcançar no máximo o valor 0,5.

O termo ligação refere-se a *locus* e não para associar alelos a *locus*. Não é correto dizer que o gene de uma determinada doença está ligado com o alelo A pelo *locus* marcador.

Um teste de ligação pode ser feito com um teste qui-quadrado (k recombinações e $n-k$ recombinações comparadas com $n/2$ recombinações sob H_0). Mas, geralmente, não é possível contar o número de recombinações em heredogramas humanos. Além disso, há formas mais eficazes de realizar testes de ligação, de forma que este teste não será empregado aqui.

Um teste uniformemente mais poderoso, pode ser encontrado utilizando o teste da razão de verossimilhanças, conjuntamente com o Lema de Neyman-Pearson. Como tem-se

$$\begin{aligned} H_0 : \theta &= 0,50 && (3.1)\\ H_1 : \theta &< 0,50, \text{ utiliza-se} \\ \Lambda &= \frac{L(\theta_1)}{L(\theta_0)}, && (3.2) \end{aligned}$$

expressão em que θ_1 representa o estimador de verossimilhança da fração de recombinação. Este teste é comumente expresso em termos do logaritmo na base 10 da razão de verossimilhança. Assim, define-se a estatística Lod Escore (*logarithm of odds*) do seguinte modo:

$$\begin{aligned} \text{Lod Escore} = Lod &= Z(\theta) = \log_{10}\left[\frac{L(\theta_1)}{L(\theta_0)}\right] && (3.3)\\ &= \log_{10}[L(\theta_1)] - \log_{10}[L(\theta_0)]. && (3.4) \end{aligned}$$

A aplicação mais comum de análise de ligação é localizar, no genoma, um gene responsável por uma doença herdada de acordo com as leis Mendelianas. O teste de ligação será aqui expresso em termos do Lod Escore, por meio do teste de hipótese $H_0 : v_d = 0$.

3.1.1 Recombinação Genética

Este fenômeno está intimamente ligado com a meiose celular. É devido a ocorrência de recombinação que existe um aumento na variabilidade genética, conferindo igual variação aos descendentes de uma espécie formados a partir dessas células.

Pode-se dizer que a recombinação baseia-se em quebras que ocorrem quando os cromossomos homólogos são emparelhados, sendo que tais quebras sempre atingem duas cromátides irmãs em pontos correspondentes e são seguidas de soldadura. Sua localização é casual, variando de célula para célula e o número de recombinações é muito irregular. As cromátides

que trocam pedaços, na seqüência da meiose, serão os novos cromossomos que se distribuirão entre as células filhas e, dessa forma, o conjunto gênico recebido pelos descendentes depende do resultado das trocas ocorridas durante o processo de divisão celular.

No cálculo da distância entre genes ao longo de um cromossomo, emprega-se a freqüência de recombinação, pois esta depende da distância entre os pontos nos quais ocorrem as quebras e permutas. Na recombinação, os alelos apenas trocam de posição dentro do par de cromossomos homólogos, de modo que a estrutura e a função cromossômica permanecem inalteradas. Esse processo não deve ser confundido com mutação.

Na formação de um gameta, os dois homólogos são copiados de cada par de cromossomos. Na distribuição de cromossomos homológos, a seleção de qualquer um deles proveniente do pai ou da mãe para uma célula filha é aleatória. Quando os pares de cromossomos homólogos alinham-se, pode ocorrer um processo chamado de *crossing-over*, o qual resulta na recombinação genética.

Recombinações ocorrem freqüentemente e o número de *crossing-over* depende do tamanho do cromossomo. Assim, pode-se relacionar fração de recombinação com distância genética.

O fundamento da análise de ligação é que eventos de recombinação ocorrem entre dois *loci* genéticos (genes, marcadores, aberrações cromossômicas, etc) segundo uma razão relacionada com a distância entre eles em um mesmo cromossomo, isto é, *loci* que estão muito próximos tendem a serem herdados juntos, conforme a Figura 3.1.

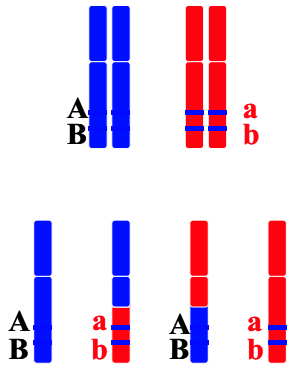

Figura 3.1: Recombinação - *Loci* Próximos

Quando os *loci* são fisicamente distantes, maior torna-se a chance de *crossing-over* e, conseqüentemente, a de recombinação. Assim, de acordo com a Figura 3.2, a presença de recombinação é um indicador da ausência de ligação.

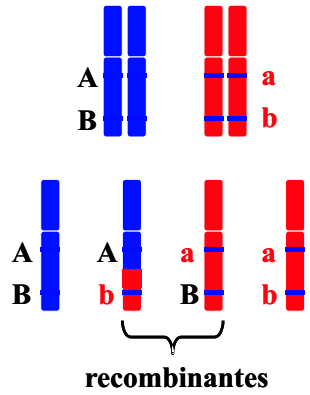

Figura 3.2: Recombinação - *Loci* Distantes

A Tabela 3.1 apresenta um sumário das possibilidades em uma análise de ligação.

Tabela 3.1: Resumo de Recombinação

	Mesmo Cromossomo			Cromossomos Diferentes
	Muito Próximo	Próximo	Distante	
Freqüência de *Crossing-Over*	Rara	Pouca	Freqüente	Freqüente
Ligação	Sim	Sim	Não	Não
θ	0%	1-49%	50%	50%

3.1.2 Teste e Estimação

Com um teste de ligação, deseja-se saber se os dados possuem informações suficientes para afirmar-se da existência de ligação entre dois genes. Usualmente, possuem-se *loci* marcadores com localização genética conhecida e uma doença para a qual deseja-se encontrar a causa genética. Portanto, constrói-se um teste baseando-se na possibilidade de ligação entre o gene de doença e os *loci* marcadores. Geralmente, uma estatística do teste superior ao valor crítico três é aceito como evidência significante de ligação, a um nível de significância aproximado de 5%. Para doenças complexas, este valor crítico pode apresentar-se pequeno.

Após encontradas evidências significativas de ligação da doença com um marcador, passa-se a procurar a localização deste gene. Sabe-se da correspondência entre a fração de recombinação e a distância física no cromossomo. O ponto $\hat{\theta}$ que maximiza o *lod score*, o EMV, é uma estimativa da fração de recombinação.

4 Procedimentos Computacionais

A seguir serão apresentados os procedimentos computacionais empregados na maximização da função *Lod Score* baseada na verossimilhança retrospectiva (2.5). O programa para a estimação foi criado em MATLAB 6.1, compilado em MEX e C, tendo por finalidade a otimização do processo.

4.1 Método para Maximização

A função (2.5), também apresentada a seguir, é não-linear e, portanto, procedimentos adequados devem ser aplicados. Além disso, existem restrições a serem satisfeitas: v_d, v_p e ζ devem ser não-negativos e $v_d + v_p = \eta$.

$$Lod_d = \max_{v_d, v_p, \beta} \sum_{i=1}^{K} \log_{10} LR_i(v_d, v_p, \boldsymbol{\beta})$$

4.1.1 Otimização com Restrições

Em uma otimização com restrições, o objetivo geral é transformar o problema em um sub-problema mais simples que possa, então, ser solucionado e usado como parte de um processo iterativo. Uma característica de uma ampla classe de métodos é a tradução do problema com restrições para um problema básico sem restrições por meio de uma função de penalização para as restrições. Dessa forma, o problema com restrições é solucionado usando uma seqüência de otimizações reparametrizadas sem restrições que, no limite da seqüência, converge para o problema com restrições. Esses métodos são agora considerados relativamente ineficientes e foram substituídos por métodos baseados na solução das equações de Kuhn-Tucker (KT). As equações de KT são condições necessárias para a otimalidade para

um problema de otimização com restrições.

Considere o problema geral (PG) de otimização:

$$\min_{x \in \Re^n} f(x), \tag{4.1}$$

tal que:

$$\begin{aligned} G_i(x) &= 0, \qquad i = 1, \ldots, m_e \\ G_i(x) &\leq 0, \quad i = m_e + 1, \ldots, m \\ x_l \leq x &\leq x_u, \end{aligned}$$

no qual x é um vetor de parâmetros, $x \in \Re^n$, $f(x)$ é a função a ser minimizada ($f(x) : \Re^n \rightarrow \Re$), e a função vetorial $G_i(x)$ retorna os valores das restrições de igualdade e desigualdade calculadas em x ($G_i(x) : \Re^n \rightarrow \Re^m$).

Uma solução precisa para este problema não depende apenas do tamanho do problema em termos de número de restrições e parâmetros, mas também das características da função e restrições. A solução de um problema não-linear geralmente requer um procedimento iterativo para estabelecer uma direção de busca em cada passo da iteração principal. Isto é geralmente alcançado pela solução de um problema de programação linear, programação quadrática ou de um sub-problema sem restrições.

Baseado no problema geral (4.1), as equações de Kuhn-Tucker são definidas como:

$$\begin{aligned} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla G_i(x^*) &= 0, \\ \lambda_i^* G_i(x^*) &= 0, \qquad i = 1, \ldots, m \\ \lambda_i^* &\geq 0, \quad i = m_e + 1, \ldots, m. \end{aligned}$$

A primeira equação descreve um cancelamento entre os gradientes da função a ser minimizada e as restrições ativas no ponto da solução. Para que os gradientes sejam cancelados, Multiplicadores de Lagrange ($\lambda_i, i = 1, \ldots, m$) precisam balancear os desvios na magnitude da função objetivo e dos gradientes das restrições. Visto que apenas restrições ativas são inclusas no procedimento de cancelamento dos gradientes, restrições que não são ativas não devem ser acrescidas no procedimento descrito e, então, os multiplicadores de Lagrange são iguais a zero (implícito nas duas últimas equações).

A solução das equações de Kuhn-Tucker formam a base de vários algoritmos de programação não-linear. Tais algoritmos procuram calcular diretamente os multiplicadores de Lagrange. Métodos quasi-Newton para casos com restrições garantem uma convergência super-linear por meio do acúmulo da informação de segunda ordem com respeito às equações de KT, usando um método quasi-Newton de atualização. Estes métodos são conhecidos como Métodos de Programação Quadrática Seqüencial (PQS), visto que um sub-problema de programação quadrática é solucionado a cada iteração principal (também conhecido como métodos de Programação Quadrática Iterativa e Programação Quadrática Recursiva).

4.1.2 Programação Quadrática Seqüencial

A implementação de Programação Quadrática Seqüencial disponível no MATLAB fundamenta-se em três passos:

- Atualização da matriz Hessiana da função Lagrangiana;
- Solução do problema de programação quadrática e
- Busca linear e cálculo da função mérito.

4.1.2.1 Atualização da Hessiana

Em cada iteração principal, uma aproximação quasi-Newton positiva-definida da Hessiana da função Lagrangiana, H, é calculada usando o método BFGS (variação do algoritmo de otimização de Newton, na qual uma aproximação da Hessiana é obtida a partir de gradientes calculados em cada iteração do algoritmo), para o qual $\lambda_i, (i = 1, \ldots, m)$ é uma estimativa dos multiplicadores de Lagrange.

$$\begin{aligned} H_{k+1} &= H_k + \frac{q_k q_k^T}{q_k s_k^T} - \frac{H_k^T H_k}{s_k^T H_k s_k}, \text{ em que} \\ s_k &= x_{k+1} - x_k \\ q_k &= \nabla f(x_{k+1}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x_{k+1}) - \left(\nabla f(x_k) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x_k) \right). \end{aligned}$$

Powell (1978) recomenda manter a Hessiana positiva definida, mesmo que ela possa ser positiva indefinida no ponto de solução. Uma Hessiana positiva definida é mantida com $q_k^T s_k$

positivo a cada atualização e H inicializada com uma matriz positiva definida. Quando $q_k^T s_k$ não é positiva, q_k é modificada de forma que $q_k^T s_k > 0$. O objetivo geral desta modificação é alterar os elementos de q_k, que contribuem para uma atualização positiva definida, o mínimo possível. Portanto, numa fase inicial da modificação, o elemento mais negativo de $q_k \times s_k$ é repetidamente diminuído. Este procedimento é continuado até que $q_k^T s_k$ seja maior ou igual a 10^{-5}. Se após esse procedimento, $q_k^T s_k$ ainda não é positivo, q_k é modificado adicionando um vetor v multiplicado por um escalar w, ou seja:

$$\begin{aligned} q_k &= q_k + wv, \text{ para o qual} \\ v_i &= \nabla g_i(x_{k+1}) g_i(x_{k+1}) - \nabla g_i(x_k) g_i(x_k), \text{ se } (q_k)_i w < 0 \text{ e } (q_k)_i (s_k)_i < 0, \quad (i = 1, \ldots, m) \\ v_i &= 0, \text{ caso contrário,} \end{aligned}$$

e w é sistematicamente aumentado até que $q_k^T s_k$ seja positivo.

4.1.2.2 Solução do Problema de Programação Quadrática

A cada iteração principal do método de Programação Quadrática Seqüencial, um problema de Programação Quadrática é resolvido de forma tal que A_i refere-se a i-ésima linha da matrix $A_{m \times n}$.

$$\begin{aligned} \min_{d \in \Re^n} q(d) &= \frac{1}{2} d^T H d + c^T d \\ A_i d &= b_i \qquad i = 1, \ldots, m_e \\ A_i d &\leq b_i \qquad i = m_e + 1, \ldots, m. \end{aligned}$$

O procedimento de solução possui duas fases: a primeira envolve o cálculo de um ponto possível (se existir); a segunda fase envolve a geração de uma seqüência iterativa de pontos possíveis que convergem para a solução. Neste método um conjunto ativo é mantido, $\bar{A}_k$, que é uma estimativa das restrições ativas no ponto de solução. Virtualmente, todos os algoritmos de Programação Quadrática são métodos de conjunto ativo. Este ponto é enfatizado porque existem diferentes métodos que são muito similares em estrutura mas que são descritos em termos muito diferentes.

$\bar{A}_k$ é atualizado em cada iteração, k, e é usado para formar uma base para a direção de procura $\hat{d}_k$. Restrições de igualdade sempre permanecem no conjunto ativo, $\bar{A}_k$. A notação

para a variável $\hat{d}_k$ é usada para diferenciá-la de d_k nas iterações principais do método de Programação Quadrática Seqüencial. A direção de busca, $\hat{d}_k$, é calculada e minimiza a função objetivo de acordo com os limites das restrições. O subespaço possível para $\hat{d}_k$ é formado a partir de uma base, Z_k, cujas colunas são ortogonais à estimativa do conjunto ativo $\bar{A}_k$ (i.e., $\bar{A}_k Z_k = 0$). Portanto, uma direção de busca, que é formada a partir de uma soma linear de qualquer combinação das colunas de Z_k, permanece nos limites das restrições ativas.

A matriz Z_k é formada a partir das $(m - l)$ colunas da decomposição QR da matriz $\bar{A}_k^T$, em que l é o número de restrições ativas e $l < m$. Isto é, Z_k é dada por:

$$\begin{aligned} Z_k &= Q[:, l+1 : m], \qquad paraaqual \\ Q^T \bar{A}_k^T &= \begin{bmatrix} R \\ 0 \end{bmatrix}. \end{aligned}$$

Encontrado Z_k, uma nova direção de busca $\hat{d}_k$ é determinada de modo a minimizar $q(d)$, tal que $\hat{d}_k$ está no espaço nulo das restrições ativas, isto é, $\hat{d}_k$ é uma combinação linear das colunas de Z_k, i.e. $\hat{d}_k = Z_k p$ para algum vetor p.

Então, se se observa a função quadrática como função de p, substituindo por $\hat{d}_k$, tem-se:

$$q(p) = \frac{1}{2} p^T Z_k^T H Z_k p + c^T Z_k p.$$

Efetuando a diferenciação com respeito a p:

$$\nabla q(p) = Z_k^T H Z_k p + Z_k^T c.$$

$\nabla q(p)$ é o gradiente projetado da função quadrática no subespaço definido por Z_k. O termo $Z_k^T H Z_k$ é chamado Hessiana projetada. Assumindo que a matriz Hessiana H é positiva definida, então o mínimo da função $q(p)$ no subespaço definido por Z_k ocorre quando $\nabla q(p) = 0$, que é a solução do sistema de equações lineares:

$$Z_k^T H Z_k p = -Z_k^T c.$$

Um passo é, então, tomado na seguinte forma:

$$x_{k+1} = x_k + \alpha \hat{d}_k, \quad \text{no qual: } \hat{d}_k = Z_k^T p.$$

A cada iteração, por conta da natureza quadrática da função objetivo, existem apenas duas escolhas para a largura do passo α. Um passo de unidade em $\hat{d}_k$ é o passo exato para o mínimo da função restrito ao espaço nulo de $\bar{A}_k$. Se este passo pode ser dado, sem violar nenhuma restrição, então esta é a solução para (4.2). Caso contrário, o passo na direção $\hat{d}_k$ para a restrição mais próxima é menor que a unidade e uma nova restrição é incluída no conjunto ativo na próxima iteração. A distância para os limites de restrições em qualquer direção $\hat{d}_k$ é dada por:

$$\alpha \quad = \quad min_i \frac{-(A_i x_k - b_i)}{A_i \hat{d}_k}, \qquad (i = 1, \ldots, m),$$

que é definido para restrições que não estão no conjunto ativo e no qual a direção $\hat{d}_k$ vá rumo ao limite de restrição, i.e., $A_i \hat{d}_k > 0$, $i = 1, \ldots, m$.

Quando n restrições independentes são incluídas no conjunto ativo, sem localização do mínimo, multiplicadores de Lagrange, λ_k, são calculados para satisfazer o conjunto não-singular de equações lineares:

$$\bar{A}_k^T \lambda_k \quad = \quad c.$$

Se todos os elementos de λ_k são positivos, x_k é a solução ótima da programação quadrática (4.2). Entretanto, se qualquer componente de λ_k é negativo, e isso não corresponde a uma restrição de igualdade, então o elemento correspondente é retirado do conjunto ativo e uma nova iteração é realizada.

Inicialização:

O algoritmo requer um ponto possível para iniciar. Se o ponto atual do método de Programação Quadrática Seqüencial não é possível, então um ponto pode ser encontrado solucionando o problema de programação linear:

$$\min \gamma,$$

considerando-se as seguintes restrições:

$$\begin{aligned} \gamma &\in \Re \\ x &\in \Re^n \\ A_i x &= b_i \qquad i = 1, \ldots, m_e \\ A_i x - y &\leq b_i \qquad i = m_e + 1, \ldots, m. \end{aligned}$$

Um ponto possível (se existir) para (4.2) pode ser encontrado ajustando x ao valor que satisfaz as restrições de igualdade. Isto pode ser alcançado resolvendo um conjunto de equações lineares sub- e sobredeterminado a partir do conjunto de restrições de igualdade. Se existir uma solução para este problema, então a variável γ é definida como a restrição de desigualdade máxima neste ponto.

O algoritmo de programação quadrática acima é modificado para problemas de programação linear ajustando a direção de busca para a direção de descida mais rápida a cada iteração na qual g_k é o gradiente da função objetivo (igual aos coeficientes da função objetivo linear).

$$\hat{d}_k = -Z_k Z_k^T g_k.$$

Se um ponto possível é encontrado usando o método de programação linear acima, entra a fase principal da programação quadrática. A direção de busca $\hat{d}_k$ é inicializada com uma direção de busca $\hat{d}_1$ encontrada a partir da solução do conjunto de equações lineares:

$$H\hat{d}_1 = -g_k,$$

para o qual g_k é o gradiente da função objetivo na iteração atual x_k (i.e. $Hx_k + c$).

Se uma solução possível não for encontrada para o problema de programação quadrática, a direção de busca para a rotina principal do método de Programação Quadrática Seqüencial $\hat{d}_k$ é encontrada como aquela que minimiza γ.

4.1.2.3 Busca Linear e Função Mérito

A solução para o subproblema de programação quadrática produz um vetor d_k, que é usado para formar uma nova iteração:

$$x_{k+1} \quad = \quad x_k + \alpha d_k.$$

O parâmetro de tamanho do passo α_k é determinado afim de produzir uma redução suficiente na função mérito. A função mérito usada por Powell (1978) com a forma abaixo foi usada na implementação:

$$\Psi(x) \quad = \quad f(x) + \sum_{i=1}^{m_e} r_i g_i(x) + \sum_{i=m_e+1}^{m} r_i \max\{0, g_i(x)\}.$$

Powell recomenda que o parâmetro de penalização seja:

$$r_i \quad = \quad (r_{k+1})_i == \max_i \left\{ \lambda_i, \frac{1}{2}((r_k)_i + \lambda_i) \right\}, \quad i = 1, \ldots, m.$$

Isso permite restrições de contribuição positiva que são inativas na solução da programação quadrática mas foram recentemente ativas. Nesta implementação, inicialmente, o parâmetro de penalização é calculado como:

$$r_i \quad = \quad \frac{\|\nabla f(x)\|}{\|\nabla g_i(x)\|},$$

tal que $\|\cdot\|$ representa a norma Euclidiana.

Isso garante maiores contribuições das restrições com gradientes menores para o parâmetro de penalização.

5 Análise de Dados

No presente trabalho, foram analisados dados simulados, gerados por meio do emprego do G.A.S.P. (Genometric Analysis Simulation Program). A geração das idades de diagnóstico foi baseada numa densidade logística, como apresentada na Equação (1.10), com $\alpha = 0,1$, $\gamma = 0,1$ e $\zeta = 0,005$, que implica na densidade apresentada na Figura 5.1; ao passo que a simulação das idades atuais foi feita de acordo com uma U(60,80).

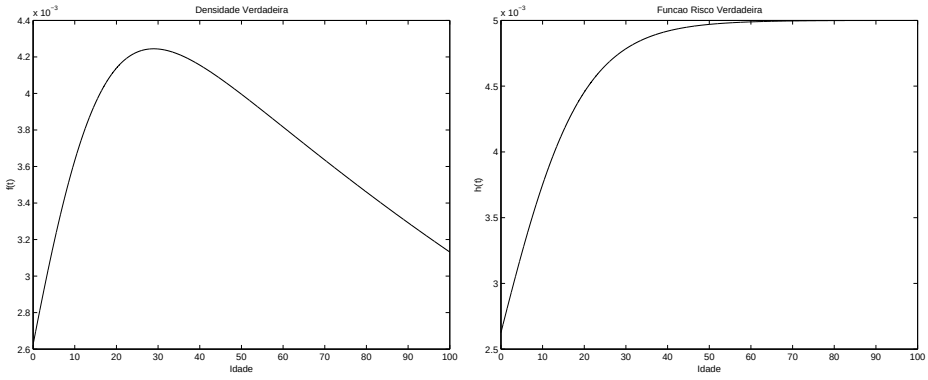

Figura 5.1: Funções Densidade e Risco Verdadeiras

Para o primeiro conjunto de dados, foi simulada uma característica binária devida a um locus com dois alelos que é ligado aos marcadores 1, 2, 3, 4 e 5. Tal característica binária é a indicadora de doença. A partir daí, foram geradas a idade de diagnóstico e a idade atual de acordo com a característica binária. O segundo conjunto de dados foi gerado de modo similar ao primeiro, exceto pelo fato de que o locus responsável pela característica qualitativa não apresenta ligação com os marcadores.

Os dois conjuntos de dados simulados com o G.A.S.P. são constituídos de mil famílias nucleares com dois filhos, portanto, quatro membros na família (pai, mãe, primeiro filho e segundo filho).

Intervalos de confiança para os parâmetros foram construídos com o emprego do método *Bootstrap* de reamostragem.

O pacote GeneHunter ajusta o modelo de análise de ligação utilizando um enfoque diferente do proposto no Modelo de Fragilidade com Risco Logístico. O GeneHunter utiliza como resposta a condição de afetado (ou não) de cada paciente, desconsiderando a estrutura de análise de sobrevivência existente nos dados. Por tal motivo, empregou-se um método alternativo para comparação dos resultados.

A solução encontrada para a validação do Modelo de Fragilidade com Risco Logístico foi ajustar o modelo de Cox para cada conjunto de dados e calcular os resíduos por Martingalas (empregando-se o SAS). A partir daí, empregam-se os resíduos como característica quantitativa a ser analisada pelo GeneHunter, via métodos não-paramétricos, obtendo-se o escore Z (*Z-score*), comumente utilizado para análise de ligação de características quantitativas (QTL). Portanto, comparam-se os modelos e valida-se a proposta do Modelo de Fragilidade com Risco Logístico por meio dos p-valores associados a cada uma das estatísticas Z e Lod escores, uma vez que $Z^2 \sim \chi^2_1$ e $2Lod_d \ln(10) \sim \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$, conforme Kruglyak (1996), Li e Zhong (2002) e Self e Liang (1987).

É possível observar, conforme apresentam as Figuras 5.1 e 5.2, que o método proposto não estimou corretamente os parâmetros, o que causou uma diferença considerável entre a densidade verdadeira e as estimadas. Conseqüentemente, a má estimação dos parâmetros reflete-se nas funções risco estimadas, vide Figuras 5.1 e 5.2.

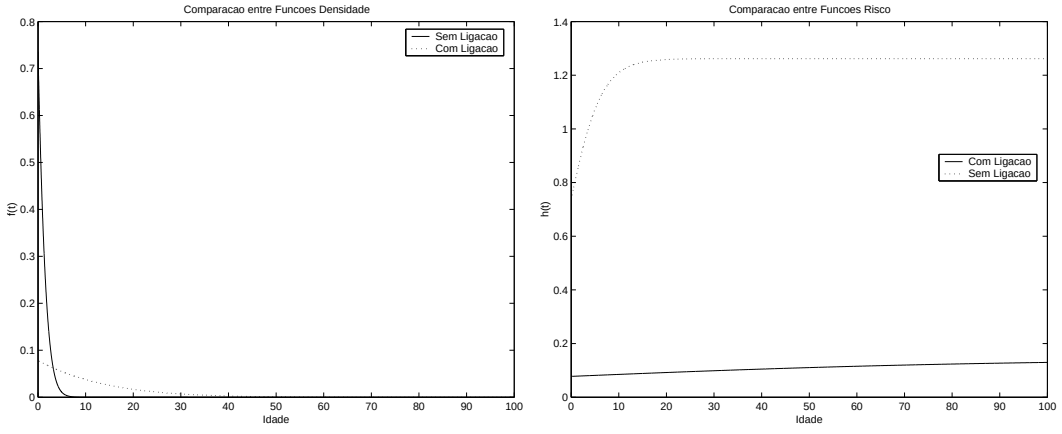

Figura 5.2: Comparações de Densidades e Riscos

5.1 Análise de Dados Simulados com Ligação

A proposta de análise de ligação por meio do Modelo de Fragilidade com Risco Logístico mostra-se concordante com os resultados obtidos com o ajuste do Modelo de Cox com resíduos por Martingalas ajustado pelo par SAS/GeneHunter.

Os cinco marcadores foram simulados com forte ligação, característica atestada por ambos os métodos, SAS/GH (SAS/GeneHunter) e MFRL (Modelo de Fragilidade com Risco Logístico), devido ao fato de o p-valor ser inferior ao limite usual de 5%.

Observa-se uma tendência do método MFRL ser mais conservador quando comparado ao SAS/GH, conforme apresentado na Tabela 5.1. Apesar desta característica, é importante ressaltar a forte evidência de ligação apontada por ambos os métodos na posição 7,38 e vizinhança, onde foi simulado o locus causador da doença. Deve-se observar que, na região onde encontra-se o locus causador, o MFRL apresenta os valores mínimos de significância, rejeitando-se a hipótese nula de inexistência de ligação.

Tabela 5.1: Comparação entre SAS/GH e MFRL - Caso com Ligação

Posição	P-valor		Posição	P-valor		Posição	P-valor	
	SAS/GH	MFRL		SAS/GH	MFRL		SAS/GH	MFRL
0,00	10^{-16}	0	7,38	0	0	14,75	10^{-16}	0
1,05	10^{-16}	0	8,43	0	0	15,80	10^{-16}	0
2,11	0	0	9,48	0	0	16,86	10^{-15}	0
3,16	0	0	10,54	0	0	17,91	10^{-15}	0
4,21	0	0	11,59	0	0	18,96	10^{-15}	0
5,27	0	0	12,64	0	0	20,02	10^{-14}	0
6,32	0	0	13,70	10^{-16}	0	21,07	10^{-14}	0

Tabela 5.2: Parâmetros Verdadeiros e suas Estimativas - Caso com Ligação

	v_d	v_p	α	γ	ζ_0	β
Verdadeiro	0,4000	0,0005	0,1000	0,1000	0,0050	2,0000
Limite Inferior 95%	0,3346	0,0008	0,0202	0,1472	0,1311	0,6634
Estimado	0,3958	0,0010	0,0208	0,1663	0,1429	0,7837
Limite Superior 95%	0,4569	0,0011	0,0215	0,1854	0,1547	0,9039

Apesar das diferenças existentes entre as estimativas dos parâmetros e os valores verdadeiros respectivos (conforme ilustra a Tabela 5.2), é necessário enfatizar que o método proposto foi capaz de detectar o mesmo padrão de ligação detectado pelo GeneHunter/SAS, conforme apresenta a Figura 5.3, pois houve a sobreposição (perfeita) dos p-valores ao longo de toda a região cromossômica analisada.

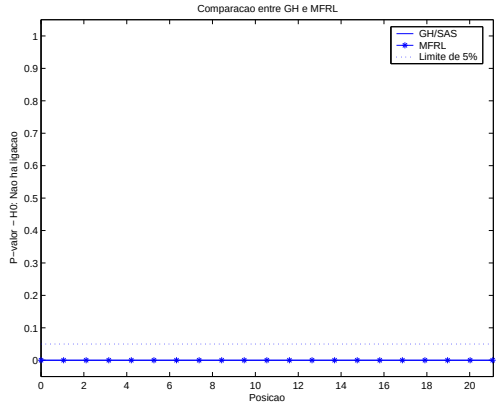

Figura 5.3: P-valores para SAS/GH e MFRL - Caso com Ligação

Conforme ilustra a Figura 5.4, observa-se a congruência entre o padrão de ligação detectado entre o GH/SAS e o MRFRL, sendo que este último apresenta o comportamento de ligação dos dados com bastante mais evidência que o primeiro. A região onde encontra-se o pico de ambas as curvas é o local onde encontra-se a causa da doença.

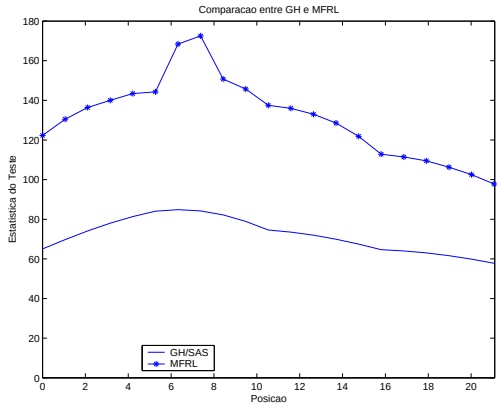

Figura 5.4: Estatísticas do Teste para SAS/GH e MFRL - Caso com Ligação

O comportamento das estimativas dos parâmetros relativos ao modelo proposto não foi de acordo com o esperado, quando comparado ao padrão utilizado para a geração dos dados, de acordo com a Tabela 5.2. O presente conjunto de dados foi gerado de modo tal que

houvesse ligação entre os marcadores e o locus causador da doença, e, de acordo com o citado anteriormente, tal característica revela-se à medida que v_d distancia-se de 0. Verifica-se uma concordância entre os valores verdadeiro e estimado para v_d, v_p e γ. Este problema pode ser causado pelo procedimento de estimação, que é bastante dependente da condição inicial do processo, de modo que diferentes condições iniciais podem gerar diferentes estimativas, além de influenciar bastante na convergência do procedimento.

5.2 Análise de Dados Simulados sem Ligação

Para os dados referentes aos cinco marcadores simulados sem ligação com o locus causador da doença, o MFRL apresenta resultados adequados. Como os dados foram gerados sob a hipótese nula (inexistência de ligação), deseja-se que o teste não rejeite tal hipótese: fato observado com mais evidência no MFRL, conforme apresentado na Tabela 5.3. Destaca-se a capacidade do Modelo de Fragilidade com Risco Logístico detectar os padrões de ligação ao longo do cromossomo, assim como o SAS/GeneHunter, vide Figura 5.5, pois não houve a rejeição da hipótese nula em nenhuma posição. Além disso, é necessário enfatizar que o teste de hipótese por meio do MFRL não rejeitou a hipótese nula com muito mais evidência (segurança) que o SAS/GeneHunter, devido ao fato de os p-valores via MRFRL sempre serem maiores que aqueles via SAS/GeneHunter.

Tabela 5.3: Comparação entre SAS/GH e MFRL - Caso sem Ligação

Posição	Valor-p		Posição	Valor-p		Posição	Valor-p	
	SAS/GH	MFRL		SAS/GH	MFRL		SAS/GH	MFRL
0,00	0,96412	1,00000	7,38	0,61912	1,00000	14,75	0,60724	1,00000
1,05	0,87122	1,00000	8,43	0,67346	1,00000	15,80	0,61870	1,00000
2,11	0,76165	1,00000	9,48	0,75867	1,00000	16,86	0,53346	1,00000
3,16	0,67187	1,00000	10,54	0,84330	1,00000	17,91	0,45898	1,00000
4,21	0,62797	1,00000	11,59	0,75816	1,00000	18,96	0,42841	1,00000
5,27	0,61790	1,00000	12,64	0,67309	1,00000	20,02	0,45041	1,00000
6,32	0,60636	1,00000	13,70	0,61953	1,00000	21,07	0,50032	1,00000

De acordo com a Figura 5.6, observa-se não foi detectada ligação por nenhum dos dois métodos, pois as estatísticas do testes apresentam-se bastante próximas de zero (ao contrário daquelas reportadas no caso de ligação apresentado anteriormente).

Tabela 5.4: Parâmetros Verdadeiros e suas Estimativas - Caso sem Ligação

	v_d	v_p	α	γ	ζ_0	β
Verdadeiro	0,0050	0,0500	0,1000	0,1000	0,0050	2,0000
Limite Inferior 95%	0,0047	0,0422	0,2189	0,2621	0,9793	1,3454
Estimado	0,0061	0,0559	0,2803	0,3356	1,2618	1,5496
Limite Superior 95%	0,0076	0,0696	0,3418	0,4090	1,5463	1,7539

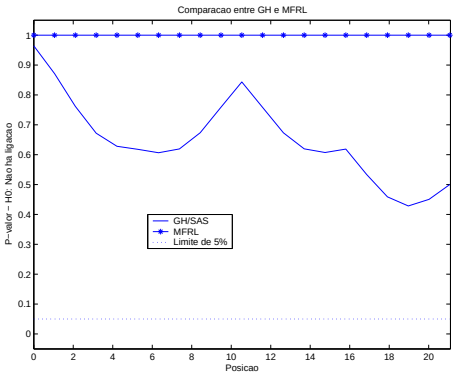

Figura 5.5: P-valores para SAS/GH e MFRL - Caso sem Ligação

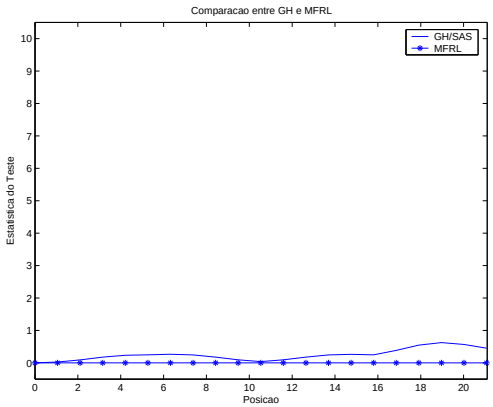

Figura 5.6: Estatísticas do Teste para SAS/GH e MFRL - Caso sem Ligação

Observa-se discordância entre os parâmetros reais e suas estimativas. Mas, a estimativa do parâmetro v_d, responsável pela detecção de ligação, é bastante próxima do valor verdadeiro, o mesmo é observado para v_p, de acordo com a Tabela 5.4. Mais uma vez, o problema de condição inicial e estabilidade do processo diante de condições iniciais distintas pode ser responsável pela estimação incorreta dos parâmetros.

Conclusão

A maximização da razão de verossimilhanças apresentada é uma tarefa muito complexa, pois trata-se da razão de duas combinações convexas das mesmas três funções. Desta forma, o procedimento de maximização não apresenta-se muito estável, pelo fato de a função apresentada possuir máximos locais dentro do espaço paramétrico.

Outros procedimentos para estimação dos parâmetros podem ser empregados como, por exemplo, o Algoritmo EM. Maior estudo a respeito da inferência dos parâmetros exige destaque por diversos motivos, como convergência dos estimadores de máxima-verossimilhança obtidos por maximização numérica, haja visto que os estimadores não possuem forma fechada e a complexidade da estrutura da verossimilhança torna-se bastante alta.

A aplicação de testes assintóticos (Wald, Razão de Verossimilhança e Score) para os parâmetros α, γ e β não foi realizada devido ao fato de não ter sido possível determinar $P(M_i)$, motivo pelo qual optou-se pela maximização da razão de verossimilhança sob a hipótese de não-ligação.

A extensão do Modelo de Fragilidade com Risco Logístico para uso em heredogramas mais complexos (com mais de uma geração e sem restrições quanto ao número de filhos) permitirá uma análise mais verossímil dos dados, por outro lado aumentará significativamente a complexidade computacional do problema.

Análise de dados disponibilizados pelo GAW 12 (*Genetic Analysis Workshop*, 2000) está sendo conduzida, afim de continuar com o processo de averiguação da validade da proposta. Entretanto, encontrou-se uma situação que precisa ser contornada afim de proceder com o estudo. Os dados possuem uma estrutura familiar complexa que, ao ser modificada afim de a presente técnica ser utilizada (ou seja, efetuar a extração de todos os pares de irmãos, de forma a existir apenas famílias nucleares compostas por pais e dois filhos), influi negativamente nos resultados. Se a análise de ligação for conduzida pelos meios usuais, observa-se no cromossomo seis evidências de ligação na região 30-32cM; enquanto que, ao utilizar os dados

no formato de pares de irmãos (*sibpair data*) não é possível detectar tais evidências.

O modelo aqui discutido constitui uma proposta que se mostra muito eficaz e deve ter seu estudo continuado, pois ele foi capaz de detectar os padrões de ligação e não-ligação (como ilustrado nas Figuras 5.3 e 5.5), apesar da diferença entre os modelos ajustados pelo GeneHunter e Modelo de Fragilidade com Risco Logístico. Análises mais detalhadas (outras propostas de maximização, como verossimilhanças perfiladas, além de comparações com os modelos de Li, 2002 e Mackenzie, 1996 afim de melhor validar o modelo aqui proposto) devem ser conduzidas para um melhor desenvolvimento desta proposta, que agrega um ferramental importante para estudos de doenças complexas que exijam análise de sobrevivência e mapeamento genético concomitantes.

ANEXO A

Provas

A.1 Função de Sobrevivência Conjunta Marginal

$$\begin{aligned}
S(t_1,\ldots,t_n|Z_1,\ldots,Z_n) &= \exp\left\{-\Lambda_1(t_1)Z_1-\cdots-\Lambda_n(t_n)Z_n\right\} \\
&= \exp\left\{-\sum_{j=1}^{n}\Lambda_j(t_j)Z_j\right\} \\
&= \exp\left\{-\sum_{j=1}^{n}\Lambda_j(t_j)\left(\sum_{i=1}^{4}a_{ji}U_{di}+U_p\right)\right\} \\
&= \left[\prod_{i=1}^{4}\exp\left\{-\sum_{j=1}^{n}\Lambda_j(t_j)a_{ji}U_{di}\right\}\right]\times \\
&\quad \times\exp\left\{-U_p\sum_{j=1}^{n}\Lambda_j(t_j)\right\}
\end{aligned}$$

Dadas as condições

$$\begin{aligned}
U_{di} &\sim \Gamma(v_d/2,\eta),\forall i; \\
U_p &\sim \Gamma(v_p,\eta); \\
U_{di} &\perp U_{dj}\ \ \forall i\neq j; \\
U_{di} &\perp U_p\ \ \forall i; \\
\Lambda_0(t_j) &= \ln\left(\frac{1+e^{t_j\alpha+\gamma}}{1+e^{\gamma}}\right)^{\frac{\zeta}{\alpha}}\ \text{e} \\
\Lambda_j(t_j) &= \Lambda_0(t_j)e^{\mathbf{X}_j'\boldsymbol{\beta}},
\end{aligned}$$

então a função de sobrevivência é expressa por

$$\begin{aligned} S(t_1,\ldots,t_n) &= \left[\prod_{i=1}^{4}\int_0^\infty \frac{\eta^{\frac{v_d}{2}}}{\Gamma(\frac{v_d}{2})}u_{di}^{\frac{v_d}{2}-1}e^{-u_{di}(\eta+\sum_{j=1}^n\Lambda_j(t_j)a_{ji})}du_{di}\right]\times \\ &\quad \times\int_0^\infty \frac{\eta^{v_p}}{\Gamma(v_p)}u_p^{v_p-1}e^{-u_p(\eta+\sum_{j=1}^n\Lambda_j(t_j))}du_p \\ &= \left\{\prod_{i=1}^{4}\left[\frac{\eta}{\eta+\sum_{j=1}^n\Lambda_j(t_j)a_{ji}}\right]^{\frac{v_d}{2}}\right\}\left[\frac{\eta}{\eta+\sum_{j=1}^n\Lambda_j(t_j)}\right]^{v_p} \end{aligned} \quad (A.1)$$

A.2 Funções Conjuntas Sobrevivência e Densidade para Pares de Irmãos

A.2.1 $IBD_d = 0$

$$S(t_1,t_2|IBD=0) = \left(\frac{\eta^2}{\Lambda_1^*\Lambda_2^*}\right)^{v_d}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}$$

Assumindo-se censura não informativa, têm-se:

$$\begin{aligned} P(t_1,\delta_1=1,t_2,\delta_2=0|IBD=0) &\propto -\frac{\partial S(t_1,t_2)}{\partial t_1} \\ &= \left(\frac{\eta^2}{\Lambda_1^*\Lambda_2^*}\right)^{v_d}\left(\frac{\eta}{\Lambda_{12}}\right)^{v_p}\lambda_1(t_1)\left(\frac{v_d}{\Lambda_1^*}+\frac{v_p}{\Lambda_{12}}\right) \\ &= \lambda_1(t_1)S(t_1,t_2)\left(\frac{v_d}{\Lambda_1^*}+\frac{v_p}{\Lambda_{12}}\right) \\ &= \lambda_1(t_1)S(t_1,t_2)C_1(t_1,t_2) \end{aligned}$$

Analogamente,

$$P(t_1,\delta_1=0,t_2,\delta_2=1|IBD=0)\propto\lambda_2(t_2)S(t_1,t_2)C_2(t_1,t_2).$$

Para derivar-se $P(t_1,\delta_1=1,t_2,\delta_2=1|IBD=0)$, tem-se que

$$\begin{aligned} P(t_1, \delta_1 = 1, t_2, \delta_2 = 1 | IBD = 0) &\propto \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} \\ &= \left[\left(\frac{v_d}{\Lambda_1^*} + \frac{v_p}{\Lambda_{12}} \right) \left(\frac{v_d}{\Lambda_2^*} + \frac{v_p}{\Lambda_{12}} \right) + \frac{v_p}{\Lambda_{12}^2} \right] \times \\ &\quad \times \lambda_1(t_1)\lambda_2(t_2) \left(\frac{\eta^2}{\Lambda_1^* \Lambda_2^*} \right)^{v_d} \left(\frac{\eta}{\Lambda_{12}} \right)^{v_p} \\ &= [C_1(t_1, t_2) C_2(t_1, t_2) + C(t_1, t_2)] \times \\ &\quad \times \lambda_1(t_1)\lambda_2(t_2) S(t_1, t_2) \end{aligned}$$

A.2.2 $IBD_d = 1$

$$S(t_1, t_2 | IBD = 1) = \left(\frac{\eta^3}{\Lambda_1^* \Lambda_2^* \Lambda_{12}} \right)^{v_d/2} \left(\frac{\eta}{\Lambda_{12}} \right)^{v_p}$$

Novamente, emprega-se a hipótese de censura não informativa, de modo que obtêm-se:

$$\begin{aligned} P(t_1, \delta_1 = 1, t_2, \delta_2 = 0 | IBD = 1) &\propto -\frac{\partial S(t_1, t_2)}{\partial t_1} \\ &= \left[\frac{v_d/2}{\Lambda_1^*} + \frac{v_d/2 + v_p}{\Lambda_{12}} \right] \lambda_1(t_1) S(t_1, t_2) \\ &= C_1(t_1, t_2) \lambda_1(t_1) S(t_1, t_2) \end{aligned}$$

Analogamente,

$$\begin{aligned} P(t_1, \delta_1 = 0, t_2, \delta_2 = 1 | IBD = 1) &\propto -\frac{\partial S(t_1, t_2)}{\partial t_2} \\ &= \left[\frac{v_d/2}{\Lambda_2^*} + \frac{v_d/2 + v_p}{\Lambda_{12}} \right] \lambda_2(t_2) S(t_1, t_2) \\ &= C_2(t_1, t_2) \lambda_2(t_2) S(t_1, t_2) \end{aligned}$$

Conseqüentemente,

$$\begin{aligned} P(t_1, \delta_1 = 1, t_2, \delta_2 = 1|IBD = 1) &\propto \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} \\ &= \lambda_1(t_1)\lambda_2(t_2)S(t_1, t_2)\left\{\left[\frac{(v_d/2)}{\Lambda_1^*} + \frac{(v_d/2 + v_p)}{\Lambda_{12}}\right]\right. \\ &\quad \left.\times\left[\frac{(v_d/2)}{\Lambda_2^*} + \frac{(v_d/2 + v_p)}{\Lambda_{12}}\right] + \frac{(v_d/2 + v_p)}{\Lambda_{12}^2}\right\} \\ &= [C_1(t_1, t_2)C_2(t_1, t_2) + C(t_1, t_2)] \times \\ &\quad \times\lambda_1(t_1)\lambda_2(t_2)S(t_1, t_2) \end{aligned}$$

A.2.3 $IBD_d = 2$

$$S(t_1, t_2|IBD = 2) = \left(\frac{\eta}{\Lambda_{12}}\right)^{v_d+v_p}$$

Assumindo-se que as censuras são não informativas, conclui-se que:

$$\begin{aligned} P(t_1, \delta_1 = 1, t_2, \delta_2 = 0|IBD = 2) &\propto -\frac{\partial S(t_1, t_2)}{\partial t_1} \\ &= \frac{v_d + v_p}{\Lambda_{12}}\lambda_1(t_1)S(t_1, t_2) \\ &= C_1(t_1, t_2)\lambda_1(t_1)S(t_1, t_2). \end{aligned}$$

Analogamente,

$$\begin{aligned} P(t_1, \delta_1 = 0, t_2, \delta_2 = 1|IBD = 2) &\propto -\frac{\partial S(t_1, t_2)}{\partial t_2} \\ &= \frac{v_d + v_p}{\Lambda_{12}}\lambda_2(t_2)S(t_1, t_2) \\ &= C_2(t_1, t_2)\lambda_2(t_2)S(t_1, t_2). \end{aligned}$$

Conseqüentemente,

$$
\begin{aligned}
P(t_1, \delta_1 = 1, t_2, \delta_2 = 1 | IBD = 2) &\propto \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} \\
&= \frac{1}{\Lambda_{12}^2} \left[(v_p + v_d) S(t_1, t_2)(1 + v_p + v_d) \lambda_1(t_1) \lambda_2(t_2) \right] \\
&= \left[\left(\frac{v_p + v_d}{\Lambda_{12}} \right)^2 + \frac{v_d + v_p}{\Lambda_{12}^2} \right] \lambda_1(t_1) \lambda_2(t_2) S(t_1, t_2) \\
&= \left[C_1(t_1, t_2) C_2(t_1, t_2) + C(t_1, t_2) \right] \times \\
&\quad \times \lambda_1(t_1) \lambda_2(t_2) S(t_1, t_2)
\end{aligned}
$$

ANEXO B

Glossário

Alelos: Diversas formas de um mesmo gene que ocupam o mesmo locus (a mesma região) em cromossomos homólogos e que determinam um caráter e formam um genótipo.

Cromossomo: Estrutura celular alongada presente no núcleo das células eucarióticas formada por proteínas e DNA.

Crossing-Over: A troca de partes correspondentes entre cromossomos homólogos por quebra e reunião. Mais precisamente é a troca de partes entre cromátides não-irmãs, fenômeno típico da meiose.

Fenótipo: Referente às propriedades morfológicas, fisiológicas, bioquímicas, comportamentais e outras de um organismo. O fenótipo se desenvolve pela interação entre os efeitos do gene e os efeitos ambientais.

Gene: Conceito muito complexo que se refere a unidade da hereditariedade, transmitida de uma geração para outra através dos gametas. O gene corresponde a determinado segmento de DNA que codifica proteínas. Todavia o gene não atua sozinho na determinação das características individuais, mas ele interage com outros genes e com o ambiente.

Genoma: Conjunto de genes presentes em todos os cromossomos de um indivíduo.

Genótipo: Termo que se refere ao conjunto de genes de um organismo. Normalmente este termo se refere a composição genética de um indivíduo em um locus especifico ou conjunto de *loci* .

Heredograma: Representação gráfica de uma família, podendo incluir características como fenótipos e genótipos.

Locus: A posição de um gene num cromossomo (pl. *loci*).

Marcador: Gene de *locus* conhecido exatamente sem função de codificação.

Penetrância: Probabilidade de desenvolvimento de uma característica (doença, por exemplo), condicionada no genótipo do indivíduo.

Segregação: Transmissão de caracteres paternos para seus descendentes.

Referências Bibliográficas

AALEN, O. Nonparametric inference for a family of counting processes. *Ann. Statist.*, v. 6, n. 4, p. 701–726, 1978. ISSN 0090-5364.

ANDERSEN, P. K.; GILL, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, v. 10, n. 4, p. 1100–1120, 1982. ISSN 0090-5364.

ANDERSEN, P. K. et al. Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, v. 53, n. 4, p. 1475–1484, 1997. ISSN 0006-341X.

ANDRADE, M.; PINHEIRO, H. P. *Métodos Estatísticos Aplicados em Genética Humana*. São Paulo: 15o. SINAPE - ABE, 2002.

CHANG, I.-S. et al. A unified multipoint linkage analysis of qualitative and quantitative traits for sib-pairs. *Statist. Sinica*, v. 12, n. 1, p. 297–309, 2002. ISSN 1017-0405. Special issue on bioinformatics.

CHICARINO, M. P. Z. *Modelo Semiparamétrico de Fragilidade Gama*. Dissertação (Mestrado) — Universidade de São Paulo, IME, 1999.

CLAYTON, D.; CUZICK, J. Multivariate generalizations of the proportional hazards model. *Jornal of Royal Statistical Society*, A 148, p. 82–108, 1985.

COX, D.; OAKES, D. *Analysis of Survival Data*. London: Chapman Hall, 1972.

KONG, A.; COX, N. J. Allele-sharing models: lod scores and accurate linkage tests. *American Journal of Human Genetics*, v. 61, p. 1179–1188, 1997.

KRUGLYAK, L. et al. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, v. 58, p. 1347–1363, 1996.

LEE, E. T. *Statistical Methods for Survival Data Analysis, 2nd. Edition*. New York: Wiley-Interscience, 1992.

LI, H.; ZHONG, X. Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics*, v. 3,1, p. 57–75, 2002.

MACKENZIE, G. Regression models for survival data: the generalized time-dependent logistic family. *The Statistician*, v. 45, p. 21–34, 1996.

MARKIANOS, K.; DALY, M. J.; KRUGLYAK, L. Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics*, v. 68, p. 963–977, 2001.

MATHWORKS. *Matlab Version 6.1.0.450 Release 12.1*. 2001.

MEDICINENET.COM. *Medical Terms Dictionary*. 2002. Disponível em: <http://www.medterms.com>.

NIELSEN, G. G. et al. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics. Theory and Applications*, v. 19, n. 1, p. 25–43, 1992. ISSN 0303-6898.

OAKES, D. Bivariate survival models induced by frailties. *Journal of American Statistical Association*, v. 84, p. 487–493, 1989.

OLSON, J. M.; WITTE, J. S.; ELSTON, R. C. Tutorial in biostatistics genetic mapping of complex traits. *Statistics in Medicine*, v. 18, p. 2961–2981, 1999.

POWELL, M. A fast algorithm for nonlinearly constrained optimization calculations. *Numerical Analysis*, v. 630, 1978. Lecture Notes in Mathematics, Springer Verlag.

RODRIGUEZ, G. *Survival Analysis - Chapter 7: Survival Models*. 2001. Disponível em: <http://data.princeton.edu/wws509/notes/c7.pdf>.

SAS INSTITUTE INC. *The SAS System for Windows, Release 8.02 TS Level 02M0*. 2001.

SELF, S. G.; LIANG, K. Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of American Statistical Association*, v. 82, p. 605–610, 1987.

TENG, J.; SIEGMUND, D. Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics*, v. 54, n. 4, p. 1247–1279, 1998. ISSN 0006-341X. With discussion and a rejoinder by the authors.

VAUPEL, J.; MANTON, K.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, v. 16, p. 439–454, 1979.

WHITTEMORE, A. S. Genome scanning for linkage: an overview. *American Journal of Human Genetics*, v. 59, p. 704–716, 1996.

WILSON, A. et al. The genometric analysis simulation program (g.a.s.p.): a software tool for testing and investigating methods in statistical genetics. *American Journal of Human Genetics*, v. 630, p. 59–193, 1996.