UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE CIÊNCIAS EXATAS E DA TERRA

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA APLICADA E ESTATÍSTICA

PATRÍCIA BORCHARDT SANTOS

ESTIMAÇÃO EM MODELOS DE TEMPO DE FALHA ACELERADO PARA DADOS DE SOBREVIVÊNCIA CORRELACIONADOS

Natal, dezembro de 2009

ESTIMAÇÃO EM MODELOS DE TEMPO DE FALHA ACELERADO PARA DADOS DE SOBREVIVÊNCIA CORRELACIONADOS

PATRÍCIA BORCHARDT SANTOS

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática Aplicada e Estatística - CCET - UFRN, como requisito parcial para obtenção do título de Mestre em Matemática Aplicada e Estatística.

Área de Concentração: Probabilidade e Estatística

Orientador: Profª Drª Dione Maria Valença

Natal, dezembro de 2009

Resumo

Apresentamos neste trabalho dois métodos de estimação para modelos de tempo de falha acelerado com efeito aleatório para tratar de dados de sobrevivência correlacionados. O primeiro método, que está implementado no *software* SAS, através do procedimento NLMIXED, utiliza a quadratura Gauss-Hermite adaptada para obter a verossimilhança marginalizada. O segundo método, implementado no *software* livre R, está baseado no método da verossimilhança penalizada para estimar os parâmetros do modelo. No primeiro caso descrevemos os principais aspectos teóricos e, no segundo, apresentamos brevemente a abordagem adotada juntamente com um estudo de simulação para investigar a performance do método. Realizamos uma aplicação dos modelos usando dados reais sobre o tempo de funcionamento de poços petrolíferos da Bacia Potiguar (RN/CE).

PALAVRAS CHAVE. Modelos de tempo de falha acelerado. Dados correlacionados. Petróleo.

Abstract

We presented in this work two methods of estimation for accelerated failure time models with random effects to process grouped survival data. The first method, which is implemented in software SAS, by NLMIXED procedure, uses an adapted Gauss-Hermite quadrature to determine marginalized likelihood. The second method, implemented in the free software R, is based on the method of penalized likelihood to estimate the parameters of the model. In the first case we describe the main theoretical aspects and, in the second, we briefly presented the approach adopted with a simulation study to investigate the performance of the method. We realized implement the models using actual data on the time of operation of oil wells from the Potiguar Basin (RN / CE).

KEYWORDS. Accelerated failure time models. Grouped data. Oil.

Sumário

Capítulo 1

Introdução

Em estudos de análise de sobrevivência, estamos interessados no tempo até que determinado evento ocorra, geralmente chamado de tempo até a falha, tempo de sobrevivência ou mesmo tempo de vida. Como exemplos, podemos ter o tempo até a cura de uma doença ou o tempo até a morte de um paciente. Os modelos clássicos consideram a suposição de que os tempos de vida são todos independentes. Entretanto, existem situações em que tal condição não pode ser assumida. A dependência dos tempos pode ser causada, por exemplo, quando observamos tempos de sobrevivência em grupos, podendo ser indivíduos de uma mesma família. Outra situação que sugere dependência é quando observamos mais de um tempo para um único indivíduo.

Neste trabalho tratamos da estimação em uma particular classe de modelos de sobrevivência para dados correlacionados. Apresentamos, a seguir, uma idéia de como esses dados surgem na prática.

1.1 Dados de Sobrevivência Correlacionados

Alguns autores preferem a expressão *multivariados* para designar dados em que a suposição de independência não é apropriada, dentre eles destacamos Hougaard (2000) que faz uma ampla discussão sobre esses dados, cuja dependência se dá basicamente por dois motivos:

1. Os indivíduos em estudo fazem parte de algum tipo de agrupamento (natural

ou artificial). Podemos pensar, por exemplo, no tempo até a recuperação de pacientes, tratados em diferentes hospitais. É possível que características como a rapidez no atendimento, a higiene e a infra-estutura do hospital, a equipe médica, dentre outras, contribuam para que haja semelhança entre os tempos de recuperação dos pacientes dentro de cada hospital.

2. Cada indivíduo em estudo está sujeito a experimentar mais de uma vez o evento de interesse. Este item pode ser subdividido em *eventos recorrentes* e *medidas repetidas*.

   (a) Eventos recorrentes ocorrem quando o número de vezes que o indivíduo vai experimentar o evento é aleatório. Pensemos no número de ataques epiléticos que indíviduos sob estudo podem experimentar em um período pré-fixado. Outros exemplos podem ser o número de ataques cardíacos ou o número de recorrências de tumores.

   (b) Para medidas repetidas o número de eventos medido para cada indivíduo é fixo. Como exemplo podemos ter um estudo em que indivíduos, sob o mesmo tratamento, terão suas pressões arteriais medidas, duas vezes ao dia, durante uma semana.

Em estudos de sobrevivência é comum que informações concomitantes sobre os indivíduos auxiliem na explicação dos tempos de vida. Por exemplo, em um estudo sobre o tempo de recuperação de dependentes químicos é possível que o tipo de tratamento (médico ou grupo de auto-ajuda), a idade, a classe social, o sexo, dentre outros, influenciem no tempo até a recuperação dos dependentes. Neste trabalho chamaremos tais informações de covariáveis.

Embora sejam conhecidas covariáveis de interesse para cada indivíduo, na prática, é razoável que existam informações importantes que não puderam ser observadas. A ausência de tais informações pode proporcionar heterogeneidade entre grupos. Apresentamos a seguir uma breve descrição do conjunto de dados que motivou este trabalho.

1.2 Motivação

Consideremos um estudo retrospectivo sobre o tempo de funcionamento de poços de petróleo terrestres localizados nos estados do Rio Grande do Norte e Ceará (Bacia Potiguar), no qual foi coletada uma amostra composta por 616 poços-coluna.

No período de janeiro de 2000 à dezembro de 2006 foi observado o tempo de funcionamento de cada um dos poços-coluna dentro da sua normalidade até apresentarem falha relacionada a equipamentos de sub-superfície, que cause parada total no funcionamento do poço-coluna. Dantas (2008) analisou esse conjunto de dados através de um modelo de regressão Weibull com o objetivo de verificar a existência do relacionamento entre o *tempo até a primeira falha* do poço-coluna e algumas características destes.

No entanto, após detectada a falha do poço-coluna, os equipamentos eram consertados e o tempo até uma nova falha foi observado. Como esse procedimento foi feito para toda a amostra durante o período observado, o evento de interesse pôde ser medido mais de uma vez em cada poço-coluna, caracterizando eventos recorrentes e totalizando 2374 tempos observados, sendo 563 censurados [1] (23,7%). Entendemos por censura os casos em que o poço-coluna foi desativado ou estava em operação quando o estudo terminou, de forma que tivemos apenas um conhecimento parcial da resposta.

Supondo que existe correlação entre os tempos de funcionamento referentes ao mesmo poço-coluna, este trabalho visa avaliar a influência de características tais como o método utilizado para a elevação dos fluidos, a profundidade da instalação da bomba, dentre outras, no tempo de sobrevivência desses poços-coluna.

1.3 Objetivos

Com base em um modelo de tempo de falha acelerado com efeito aleatório para tratar a depedência entre tempos de sobrevivência correlacionados, os objetivos deste trabalho são:

[1] Ver seção 2.1.

- Descrever os principais aspectos teóricos da abordagem usada por Valença (2003) e Lambert et al. (2004), implementada através do procedimento NLMIXED do *software* SAS, que utiliza uma técnica de integração numérica chamada quadratura Gauss-Hermite adaptada, para obter estimativas dos parâmetros do modelo.

- Apresentar brevemente uma outra abordagem de estimação, implementada no *software* R, que se baseia na verossimilhança penalizada.

- Fazer um estudo de simulação para avaliar como o método que utiliza a penalização da verossimilhança está implementado no R e o comparar com o método que ignora a dependência dos tempos de sobrevivência, também implementado no R.

- Verificar a influência de covariáveis, tais como, produção e idade do poço, no tempo de funcionamento de poços de petróleo da Bacia Potiguar.

1.4 Conteúdo dos Capítulos

Os capítulos deste trabalho estão dispostos da seguinte forma: no Capítulo 2 introduzimos conceitos iniciais em Análise de Sobrevivência juntamente com a descrição de dois modelos de regressão amplamente usados, sob a hipótese de independência: modelos de tempo de falha acelerado e modelos de riscos proporcionais. A seguir apresentamos algumas referências sobre a utilização desses modelos no enfoque de dados de sobrevivência correlacionados. Finalizamos esse Capítulo apresentando o modelo de tempo de falha acelerado e a introdução de um efeito aleatório como alternativa para tratar a dependência dos tempos de vida. No Capítulo 3 descrevemos uma proposta de estimação dos parâmetros através do *software* SAS, usando o procedimento NLMIXED, que utiliza uma técnica de integração numérica chamada Quadratura Gauss-Hermite Adaptada. No Capítulo 4 uma outra proposta de estimação é discutida, agora usando o *software* livre R e um estudo de simulação é feito para investigar a implementação do método proposto, juntamente com o procedimento convencional que ignora a dependência dos tempos. No Capítulo 5 realizamos uma aplicação do modelo utilizando

dados reais cedidos pela PETROBRAS sobre o tempo de funcionamento de poços de petróleo. O Capítulo 6 é o último capítulo deste trabalho e nele estão as conclusões além de possibilidades de pesquisas futuras. Por fim, o Apêndice A fornece o algoritmo do programa utilizado na simulação e no Apêndice B constam os programas utilizados na aplicação dos dados.

Capítulo 2

Tópicos em Análise de Sobrevivência

O objetivo deste capítulo é fazer uma breve revisão dos conceitos iniciais da análise de sobrevivência, além de apresentar duas classes de modelos de regressão amplamente usadas para dados de sobrevivência, sob a suposição de independência: modelos de riscos proporcionais e modelos de tempo de falha acelerado. Trazemos algumas referências desses modelos no enfoque de dados correlacionados e finalizamos este capítulo com a adição de um efeito aleatório em um modelo de tempo de falha acelerado, para dados dessa natureza.

2.1 Conceitos Iniciais em Análise de Sobrevivência

Embora os métodos de sobrevivência sejam bastante utilizados na área médica, eles também se aplicam a outras situações. Como exemplo, pode ser de interesse estudar o tempo de duração de uma greve, o tempo até a conclusão de um curso, o tempo entre a liberação de um preso e seu retorno à prisão ou ainda o tempo em que determinado aparelho eletrônico permaneceu funcionando.

Entretanto, é possível (e bastante comum) a presença de *censura* em dados de sobrevivência. Isso ocorre quando temos algum conhecimento do tempo de sobrevivência de um indivíduo ou item, mas não conhecemos o seu valor exato. Na prática, a cen-

sura pode acontecer por várias razões, dentre elas, o paciente pode morrer de uma causa diferente da estudada, pode se mudar e perder o contato com os pesquisadores ou mesmo pode ocorrer de um item não falhar até o final do estudo.

Trataremos a seguir dos conceitos de *censura aleatória*, *à direita* e *não informativa*. Na *censura aleatória* um indivíduo ou item pode sair no decorrer do estudo sem ter ocorrido a falha. Para representar esse mecanismo de censura, consideremos as variáveis aleatórias T e C, independentes, contínuas e não-negativas, representando o tempo até a falha de um indivíduo e o tempo até a censura associado a este indivíduo, respectivamente. O que se observa para ele é, portanto,

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

e uma variável aleatória δ indicando se o tempo é de vida ou de censura,

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

Na *censura à direita* o verdadeiro tempo de sobrevivência ocorreria em um instante posterior ao tempo observado (censurado) e a censura é considerada *não informativa* quando não há razão para suspeitar que o motivo da perda de informação esteja relacionado ao evento de interesse. Considere, como exemplo, que o evento de interesse é a remissão de uma doença e um paciente que está sendo acompanhado no estudo morre atropelado. O tempo registrado de acompanhamento para este paciente representa uma censura não informativa.

Suponha que o tempo até a falha, T, possui uma função densidade $f(t)$ e função de distribuição acumulada $F(t)$. Definimos a *função de sobrevivência*, $S(t)$, como a probabilidade de o evento ocorrer após o instante t, ou seja, a probabilidade de o "indivíduo sobreviver" pelo menos até t. Esta função é representada por

$$S(t) = P(T \geq t) = 1 - F(t),$$

em que $S(t)$ é uma função monótona, contínua não-crescente, com $\lim_{t\to 0} S(t) = 1$ e $\lim_{t\to\infty} S(t) = 0$. Na ausência de censura, a função de sobrevivência pode ser estimada pela função de sobrevivência empírica:

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

Quando temos dados censurados, uma das formas de estimarmos $S(t)$ é através do estimador não-paramétrico de Kaplan-Meier (KAPLAN; MEIER, 1958). A Figura 2.1 mostra as curvas de sobrevivência estimadas pelo método de Kaplan-Meier para um conjunto de dados sobre pacientes divididos em dois grupos distintos de tratamento.

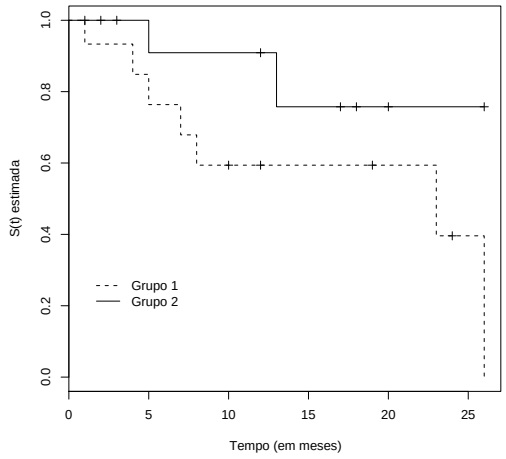

Figura 2.1: Estimativas de Kaplan-Meier para dois grupos. Os tempos indicados por + representam ocorrências de censuras.

A *função de taxa de falha* ou *de risco*, $h(t)$, é definida como a taxa de falha instantânea no tempo t dado que até o tempo t o evento não tenha ocorrido. Matematicamente, pode ser expressa da seguinte forma

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

A Figura 2.2 mostra uma situação em que a função de risco aumenta ao longo do tempo. Essa situação pode ser esperada, por exemplo, em um estudo sobre o tempo

até a solidificação de uma fratura óssea, já que se espera que a chance de solidificação cresça com o passar do tempo (CARVALHO et al., 2005).

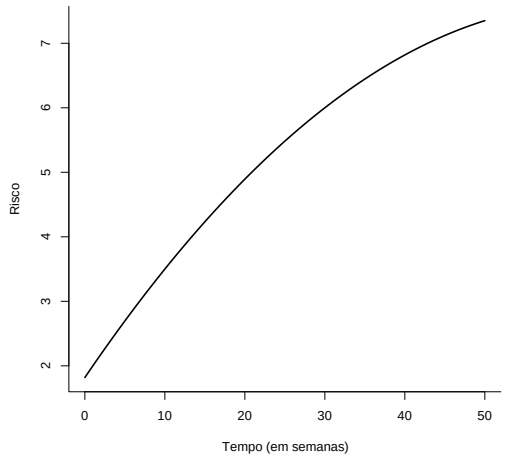

Figura 2.2: Função de risco crescente

Existem algumas relações entre as funções $f(t)$, $F(t)$, $S(t)$ e $h(t)$, de forma que com o conhecimento de uma delas, as outras poderão ser encontradas. Abaixo apresentamos estas relações:

$$f(t) = -\frac{d}{dt}S(t),$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t)),$$

e

$$S(t) = exp\left\{-\int_0^t h(u)du\right\}.$$

Considerando uma abordagem paramétrica, dentre as distribuições usadas para modelar o tempo de vida, podemos citar as distribuições exponencial, Weibull, log-normal e gama. Pela relevância da distribuição Weibull neste trabalho, finalizaremos esta seção com alguns resultados para esta distribuição.

Considere que a variável aleatória T tem distribuição Weibull, cuja função densidade

de probabilidade é dada por

$$f(t) = \frac{\rho}{\alpha}\left(\frac{t}{\alpha}\right)^{\rho-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\rho}\right], \quad t > 0, \tag{2.1}$$

com $\alpha > 0$ e $\rho > 0$ representando os parâmetros de escala e forma, respectivamente. Quando $\rho = 1$ a expressão (2.1) se reduz à função densidade da distribuição exponencial com parâmetro $\frac{1}{\alpha}$, cuja função de risco é constante, conforme mostra a Figura 2.3.

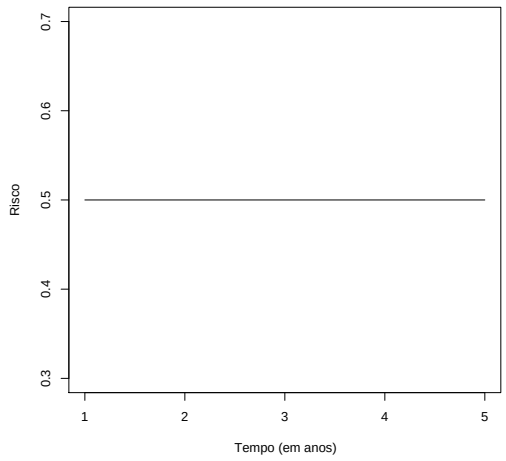

Figura 2.3: Forma típica da função de risco da distribuição exponencial para $\alpha = 2$.

As funções de sobrevivência e de risco da distribuição Weibull são, respectivamente,

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\rho}\right]$$

e

$$h(t) = \frac{\rho}{\alpha}\left(\frac{t}{\alpha}\right)^{\rho-1} .$$

Lembrando que $\Gamma(k)$ é a função gama definida como

$$\Gamma(k) = \int_0^{\infty} x^{k-1} \exp\{-x\}\, dx, \quad k > 0,$$

a média e a variância da distribuição Weibull são dadas por

$$E(T) = \alpha\Gamma\left(1 + \frac{1}{\rho}\right)$$

e

$$Var(T) = \alpha^2\left[\Gamma\left(1 + \frac{2}{\rho}\right) - \Gamma\left(1 + \frac{1}{\rho}\right)^2\right].$$

Quando tomamos o logaritmo de uma variável com distribuição Weibull, surge uma distribuição importante, chamada valor extremo (ou de Gambel). De forma que $Y = \log(T)$ tem função densidade

$$f(y) = \frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}, \quad y \in \mathbb{R}, \tag{2.2}$$

com $\mu \in \mathbb{R}$ parâmetro de locação e $\sigma > 0$ parâmetro de escala, cujas relações com os parâmetros da distribuição Weibull são $\mu = \log(\alpha)$ e $\sigma = \frac{1}{\rho}$.

As funções de sobrevivência e de risco da variável Y são, respectivamente,

$$S(y) = \exp\left\{-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$$

e

$$h(y) = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\},$$

com média e variância dadas por

$$E(Y) = \mu - c\sigma$$

e

$$Var(Y) = \frac{\sigma^2\pi^2}{6},$$

sendo $c = 0.5772...$, conhecida como constante de Euler.

Se $\mu = 0$ e $\sigma = 1$, temos a distribuição valor extremo padrão, com função densidade

$$f(\epsilon) = \exp\{\epsilon - \exp(\epsilon)\}, \quad \epsilon \in \mathbb{R}. \tag{2.3}$$

Pode-se mostrar que Y também pode ser escrita como $Y = \mu + \sigma\epsilon$.

2.2 Modelos de Regressão

Como vimos na Seção 1.1, estudos de sobrevivência muitas vezes envolvem covariáveis que estão relacionadas com o tempo de sobrevivência. Uma maneira de modelar o relacionamento entre essas covariáveis e a variável resposta é através de modelos de regressão. As classes de modelos de regressão mais comumente usadas para analisar dados de sobrevivência independentes são a que envolve o modelo de riscos proporcionais (COX, 1972) e a que trabalha com o modelo de tempo de falha acelerado - MTFA. Informações detalhadas sobre esses modelos podem ser encontradas em Lawless (2003), Colosimo e Giolo (2006), dentre outros. A seguir faremos uma breve descrição desses modelos. Consideremos, para isso, um vetor de covariáveis $\mathbf{x}$, de ordem $p \times 1$, associado com o tempo de vida T para cada indivíduo.

Modelos de Riscos Proporcionais

A característica básica deste modelo é a proporcionalidade dos riscos durante todo o período de observação, como o próprio nome sugere. É o modelo mais utilizado na análise de dados de sobrevivência e não exige que se suponha uma distribuição para o tempo de vida T. O modelo de riscos proporcionais pode ser definido por

$$h(t;\mathbf{x}) = h_0(t)\exp(\boldsymbol{\varphi}^{\mathsf{T}}\mathbf{x}), \tag{2.4}$$

sendo $h_0(t)$ a função de risco basal e $\boldsymbol{\varphi}$ o vetor de parâmetros associado às covariáveis. A razão entre o risco de ocorrência do evento para dois indivíduos com covariáveis $\mathbf{x_1} = (x_{11}, x_{12}, \cdots, x_{1p})^T$ e $\mathbf{x_2} = (x_{21}, x_{22}, \cdots, x_{2p})^T$ é

$$\frac{h(t;\mathbf{x_1})}{h(t;\mathbf{x_2})} = \frac{h_0(t)\exp(\boldsymbol{\varphi}^{\mathsf{T}}\mathbf{x_1})}{h_0(t)\exp(\boldsymbol{\varphi}^{\mathsf{T}}\mathbf{x_2})} = \frac{\exp(\boldsymbol{\varphi}^{\mathsf{T}}\mathbf{x_1})}{\exp(\boldsymbol{\varphi}^{\mathsf{T}}\mathbf{x_2})}$$

que não depende do tempo.

Quando não assumimos uma distribuição para $h_0(t)$ temos o modelo de riscos proporcionais de Cox. Se considerarmos uma representação paramétrica para $h_0(t)$ obtemos a família paramétrica de riscos proporcionais, da qual a distribuição Weibull faz

parte.

Modelos de Tempo de Falha Acelerado

Esta classe se caracteriza pelo fato da variável $Y = \log(T)$ ter uma distribuição com parâmetro de locação $\mu(\mathbf{x})$, geralmente definido como $\mu(\mathbf{x}) = \alpha + \boldsymbol{\beta}^T\mathbf{x}$, e um parâmetro de escala σ constante. Pode ser representada por:

$$Y = \log(T) = \alpha + \boldsymbol{\beta}^T\mathbf{x} + \sigma\epsilon, \tag{2.5}$$

com $\sigma > 0$, α e $\boldsymbol{\beta}$=$(\beta_1, \cdots, \beta_p)^T$ parâmetros desconhecidos e ϵ um erro aleatório cuja distribuição não depende de $\mathbf{x}$.

Este modelo recebe este nome porque a função das covariáveis é acelerar ou desacelerar o tempo de vida. Nesta classe de modelos é necessário supor uma distribuição para o tempo de vida T, de forma que são menos flexíveis que os modelos de riscos proporcionais. Dentre as distribuições assumidas para T podemos citar a log-normal, gama, log-logística e Weibull, consequentemente as distribuições para ϵ são, respectivamente, normal padrão, log-gama, logística e valor extremo padrão.

2.3 Modelos de Regressão para dados de Sobrevivência Correlacionados

Quando estamos tratanto de dados de sobrevivência correlacionados, os modelos vistos anteriormente não podem ser diretamente usados para a análise dos dados. Uma alternativa é incluir no modelo um efeito aleatório para cada grupo de indivíduos, a fim de tratar a correlação existente entre os tempos. Em modelos de riscos proporcionais o efeito aleatório é geralmente chamado fragilidade.

Neste enfoque Clayton e Cuzik (1985) estendem o modelo de riscos proporcionais pela adição de um termo de fragilidade multiplicativo, com distribuição gama. Klein (1992) propôs um modelo de riscos proporcionais de Cox com fragilidade gama, em que as estimativas dos parâmetros foram obtidas usando um algoritmo EM. Therneau

e Grambsch (2000) fornecem uma ampla discussão sobre modelos de Cox aplicados a dados de sobrevivência correlacionados em que utilizam o método da verossimilhança penalizada para estimar os parâmetros do modelo.

Por outro lado, quando dados correlacionados são analisados através de um modelo de tempo de falha acelerado, podemos citar Pan (2001) que incorpora uma fragilidade com distribuição gama ao termo do erro de um MTFA não paramétrico. A estimação dos parâmetros foi feita via algoritmo EM e o modelo foi aplicado a um conjunto de dados referente a pacientes com infecção renal. Como uma alternativa ao MTFA, Ha, Lee e Song (2002) propõem um modelo linear misto normal, com efeito e erro aleatórios seguindo uma distribuição normal com média zero e variâncias desconhecidas. Esse modelo torna-se um MTFA com efeito aleatório se a transformação logaritma for usada. Para a estimação dos parâmetros do modelo, desenvolvem um método inferencial via verossimilhança hierárquica (LEE; NELDER, 1996).

Valença (2003) apresenta um MTFA para tratar de dados correlacionados e considera uma distribuição normal para o efeito aleatório. A autora também propõe um procedimento para obtenção de estimadores de máxima verossimilhança dos parâmetros da verossimilhança marginalizada. Considera uma aproximação da integral por uma quadratura adaptada, e a maximização da verossimilhança através de um método iterativo. O procedimento NLMIXED do SAS e o algoritmo quase-Newton foram usados para obter as estimativas. Lambert et al. (2004) em um trabalho com o mesmo enfoque, utiliza também um MTFA com um efeito aleatório. Os autores avaliam diferentes combinações da distribuição assumida para o efeito aleatório e para a função de risco basal, para ajustar dados de sobrevivência de pacientes com transplante renal agrupados em diferentes centros de transplante. Para o efeito aleatório os autores consideram as distribuições gama, normal inversa e log-normal e usam também o procedimento NLMIXED do SAS.

Kelly (2004) faz uma revisão de seis *softwares* para ajustar dados de sobrevivência correlacionados. Foram considerados, entre outros, os softwares SAS, WinBUGS, SPlus e R. Cada software foi revisado considerando o MTFA e o modelo de Cox. O autor

conclui que pode-se usar os softwares S-Plus e R para ajustar um modelo com um efeito aleatório simples e para se usar o SAS é necessário alguma programação. Também verificou que o WinBUGS é indicado para ajustar um modelo com mais de um efeito aleatório

Ha, Lee e Pawitan (2007) propoem um modelo linear misto genético para analisar dados de sobrevivência de gêmeos, em que os tempos de sobrevivência poderiam ser truncados à esquerda e censurados à direita. Nesse modelo foram considerados dois efeitos aleatórios, um para o efeito genético e outro para o ambiente compartilhado entre indivíduos de cada grupo. Os efeitos aleatórios, bem como o erro aleatório, seguem distribuição normal com média zero e variância desconhecida. Zhang e Peng (2007) propoem um novo método de estimação não paramétrico baseado nos M estimadores (RITOV, 1990) e no algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977). Para avaliar a performance desse modelo, realizam um estudo de simulação e, por fim, uma aplicação usando dados reais. Além disso, oferecem uma vasta revisão de publicações que analisam dados de sobrevivência correlacionados.

Neste trabalho estudamos duas abordagens para estimação dos parâmetros em MTFA com efeito aleatório. Para a primeira, dada por Valença (2003) e Lambert et al. (2004), descrevemos alguns procedimentos teóricos e a implementação via *software* SAS. A segunda abordagem é apresentada por Therneau, Grambsch e Pankratz (2003). Nesse caso descrevemos brevemente o procedimento teórico e a implementação no R, e realizamos simulações para suprir a deficiência bibliográfica na descrição do método para MTFA. Na próxima seção apresentamos o MTFA com a adição de um efeito aleatório, que será posteriormente utilizado nos procedimentos de estimação discutidos nesse trabalho.

2.4 Modelo de Tempo de Falha Acelerado com Efeito Aleatório

Considere uma amostra dividida em k grupos. Seja T_{ij} o tempo de vida do indivíduo j no grupo i, com $j = 1, \ldots, n_i$; $i = 1, \ldots, k$ e $n = \sum_{i=1}^{k} n_i$. Os tempos de sobrevivência T_{ij} podem ser parcialmente observados devido a presença de censuras, representadas por variáveis aleatórias C_{ij}, independentes dos tempos de vida. Assumimos censuras à direita, com mecanismo aleatório e não informativas. Seja $Y_{ij} = min(\log(T_{ij}), \log(C_{ij}))$ o tempo observado para o indivíduo j no grupo i. O indicador de falhas δ_{ij} é definido como

$$\delta_{ij} = \begin{cases} 1, & T_{ij} \leq C_{ij} \\ 0, & T_{ij} > C_{ij} \end{cases}$$

Denotemos por $\mathbf{x}_{ij}$ o vetor de covariáveis de ordem $p \times 1$ medido para o $j-$ésimo indivíduo, $j = 1, \ldots, n_i$, no grupo i, $i = 1, \ldots, k$. Os dados são então representados por (Y_{ij}, δ_{ij}), e pelas covariáveis $\mathbf{x}_{ij}$.

Com base nos modelos propostos por Valença (2003) e Lambert et al. (2004), que incluem um efeito aleatório no MTFA (2.5), consideramos a seguinte representação:

$$\log(T_{ij}) = \alpha + U_i + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij} + \sigma\varepsilon_{ij}, \tag{2.6}$$

sendo ε_{ij} erros aleatórios independentes e identicamente distribuídos com distribuição valor extremo padrão com densidade (2.3). O vetor $\boldsymbol{\beta}$ e σ são parâmetros desconhecidos e para cada grupo temos um efeito aleatório (não observável) U_i representado por variáveis aleatórias independentes e identicamente distribuídas com função densidade $g(\cdot)$, com média zero e variância θ. Observe que a covariância entre o logaritmo dos tempos de sobrevivência dentro de cada grupo é dada por θ, portanto o modelo reduz-se ao MTFA usual quando $\theta = 0$. Assumimos, ainda, que os efeitos aleatórios U_i são independentes dos tempos de censura e que $\mathrm{Cov}(U_i, \varepsilon_{ij}) = 0$. Nos próximos capítulos veremos dois procedimentos para estimar os parâmetros do modelo acima.

Capítulo 3

Estimação via Verossimilhança Marginal

Neste capítulo descrevemos um procedimento de estimação implementado no SAS, baseado na verossimilhança marginal para o ajuste do modelo. Apresentamos, também, um estudo sobre métodos numéricos, usados na aproximação de integrais que não podem ser resolvidas analiticamente. Nosso enfoque aqui é a quadratura Gauss-Hermite adaptada, que é utilizada no procedimento NLMIXED do SAS.

3.1 Procedimento de Estimação

Com o intuito de eliminarmos o efeito aleatório, variável não observável, utilizaremos a verossimilhança marginal na maximização da função de verossimilhança.

Considere o modelo descrito em (2.6) e seja $\lambda = (\alpha, \boldsymbol{\beta}^{\mathsf{T}}, \sigma, \theta)^{\mathsf{T}}$ o vetor de parâmetros desconhecidos que desejamos estimar. A verossimilhança condicional ao efeito do grupo, para o indivíduo j no grupo i é dada por

$$L_{ij}(\alpha, \boldsymbol{\beta}, \sigma | u_i) = f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}},$$

com $f(\cdot)$ e $S(\cdot)$ denotando, respectivamente, as funções densidade e sobrevivência condicionais de $\log T_{ij}$ dado o efeito do i-ésimo grupo U_i.

Denotamos o vetor de tempos observados no grupo i como $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^{\mathsf{T}}$,

e consideramos δ_i e $\mathbf{x}_i$ definidos analogamente. Pela suposição de independência condicional ao efeito do grupo, temos que a verossimilhança condicional para os indivíduos do grupo i é da forma

$$L_i(\alpha, \boldsymbol{\beta}, \sigma|u_i) = \prod_{j=1}^{n_i} f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}},$$

para $i = 1, \ldots, k$. Assim, a verossimilhança relativa à distribuição marginal de (Y_i, δ_i), denotada por $L_i(\lambda)$, é representada por

$$L_i(\lambda) = \int_{-\infty}^{\infty} L_i(\alpha, \boldsymbol{\beta}, \sigma|u_i) g(u_i; \theta) du_i,$$

sendo $g(\cdot)$ a função densidade da variável U_i.

Apesar da possível correlação existente dentro dos grupos, assumimos aqui a independência entre os vetores $(Y_1, \delta_1), \ldots, (Y_k, \delta_k)$. Desta forma, o logaritmo da verossimilhança marginal que desejamos obter para toda a amostra, é dado por

$$\begin{aligned} l(\lambda) &= \log \prod_{i=1}^{k} L_i(\lambda) \\ &= \sum_{i=1}^{k} \log \int_{-\infty}^{\infty} L_i(\alpha, \boldsymbol{\beta}, \sigma|u_i) g(u_i; \theta) du_i. \end{aligned} \tag{3.1}$$

A solução da integral em (3.1) e a maximização de $l(\lambda)$ com respeito a λ, dependem das distribuições assumidas para o efeito aleatório e para os tempos de vida. No nosso caso, em que consideramos a distribuição normal para o efeito aleatório e a Weibull para os tempos de vida, a integral em (3.1) não pode ser resolvida analiticamente. A abordagem adotada em Valença (2003) e Lambert et al. (2004) é a aproximação da integral por uma quadratura Gauss-Hermite adaptada. Dedicamos a próxima Seção para apresentar diversos conceitos relacionados a aproximação de integrais, descritos com mais detalhes em Barroso et. al (1987), Hildebrand (1974) e Atkinson (1978).

3.2 Integração Numérica

Existem várias razões que justificam a necessidade do recurso a métodos numéricos para aproximar integrais.

- Há funções cuja primitiva não é uma função que se possa expressar em termos de funções elementares;
- Às vezes, embora a primitiva da função a integrar seja conhecida, a sua expressão é de tal modo complicada que não é eficiente o seu uso;
- Além disso, poderemos ter necessidade de integrar uma função da qual conhecemos apenas uma tabela de valores, obtidos, por exemplo, experimentalmente.

Antes de prosseguirmos com o estudo de métodos numéricos apresentamos uma breve revisão de Interpolação.

3.2.1 Interpolação

A interpolação é uma técnica usada quando uma função é conhecida apenas em um conjunto finito e discreto de pontos de um intervalo. Para trabalharmos com uma função sem dispormos de sua forma analítica, podemos substituí-la por outra função que é uma aproximação da função dada e que é deduzida a partir dos dados tabelados.

Outra aplicação da interpolação é a aproximação de funções complexas por funções mais simples. Suponha que tenhamos uma função cuja forma analítica é bastante complicada. Podemos, então, escolher alguns dados pontuais da função complicada e tentar interpolar esses dados para construir uma função cujo manuseio seja bem mais simples.

Obviamente, quando utilizamos a função mais simples para calcular novos dados, em geral não se obtém o mesmo resultado da função original. Contudo, dependendo do problema, o ganho de simplicidade pode compensar o erro.

Interpolação de Lagrange

Para encontrarmos um polinômio de grau no máximo $n-1$ que passa por n pontos $(x_i, f(x_i)), i = 1, 2, ..., n$, de uma função $f(x)$, definiremos as funções auxiliares:

$$\pi(x) = (x - x_1)(x - x_2)...(x - x_n)$$

e

$$l_i(x) = \frac{\pi(x)}{(x - x_i)\pi'(x)} = \frac{(x - x_1)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}, \tag{3.2}$$

com

$$\pi'(x_i) = (x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)$$

e

$$l_i(x_j) = \delta_{ij}, \tag{3.3}$$

sendo δ_{ij} o delta de Kronecker,

$$\delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases},$$

para $i = 1, \ldots, n$ e $j = 1, \ldots, n$.

Assim, o polinômio interpolador de Lagrange pode ser escrito como

$$P(x) = \sum_{i=1}^{n} l_i(x) f(x_i).$$

Como exemplo, a Figura 3.1 compara a função $f(x) = 0,5e^{-(x-1)^2} + e^{-(x+1)^2}$ com o polinômio interpolador de Lagrange nos nós (-2,02; -0,96; 0; 0,96; 2,02).

Interpolação de Hermite

Suponhamos que sejam conhecidas as triplas $(x_i, f(x_i), f'(x_i)), i = 1, 2, ..., n$, em que $f'(x_i)$ é a 1ª derivada da função f desconhecida, no ponto x_i.

Desejamos encontrar uma função polinomial de ordem no máximo $2n-1$ que passa por todos os valores $(x_i, f(x_i))$, com 1ª derivada igual a $f'(x_i)$ em $x_i, i = 1, 2, ..., n$.

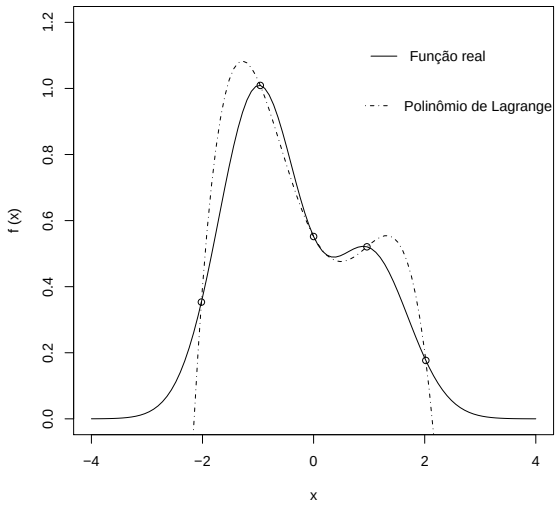

Figura 3.1: Comparação entre o polinômio de Lagrange e a função $f(x)$.

Para determinar tal polinômio assumimos que ele pode ser expresso na forma:

$$P(x) = \sum_{i=1}^{n} h_i(x) f(x_i) + \sum_{i=1}^{n} \overline{h}_i(x) f'(x_i), \tag{3.4}$$

sendo $h_i(x)$ e $\overline{h}_i(x), i = 1, 2, ..., n$ polinômios de ordem no máximo $2n - 1$.

Para que $P(x_j) = f(x_j)$ seja satisfeito, é necessário que

$$h_i(x_j) = \delta_{ij} \quad \overline{h}_i(x_j) = 0, \tag{3.5}$$

pois,

$$\begin{aligned} P(x_j) &= \sum_{i=1}^{n} h_i(x_j) f(x_i) + \sum_{i=1}^{n} \overline{h}_i(x_j) f'(x_i) \\ &= \sum_{i=1}^{n} \delta_{ij} f(x_i) + \sum_{i=1}^{n} 0.f'(x_i) \\ &= \sum_{i=1, i\neq j}^{n} \delta_{ij} f(x_i) + \delta_{jj} f(x_j) + 0 \\ &= 0 + ... + 0 + f(x_j) \\ &= f(x_j), \quad \forall j = 1, ..., n. \end{aligned}$$

Ao passo que o requisito $P'(x_j) = f'(x_j)$ será satisfeito se

$$h_i'(x_j) = 0 \quad \overline{h'}_i(x_j) = \delta_{ij}, \tag{3.6}$$

pois,

$$\begin{aligned} P'(x_j) &= \sum_{i=1}^{n} h_i'(x_j) f(x_i) + \sum_{i=1}^{n} \overline{h'}_i(x_j) f'(x_i) \\ &= \sum_{i=1}^{n} 0.f(x_i) + \sum_{i=1}^{n} \delta_{ij} f'(x_i) \\ &= 0 + \sum_{i=1, i\neq j}^{n} \delta_{jj} f'(x_j) \\ &= 0 + ... + 0 + f'(x_j) \\ &= f'(x_j), \quad \forall j = 1, ..., n. \end{aligned}$$

Sendo $l_i(x)$ um polinômio de grau $n-1$, como dado em (3.2), então $[l_i(x)]^2$ que também satisfaz (3.3) é um polinômio de grau $2n-2$ e cuja derivada desaparece em x_j quando $i \neq j$. Assim, $[l_i(x)]^2$ parece ser um bom candidato para h_i e $\overline{h}_i$. Portanto, para que h_i e $\overline{h}_i$ sejam polinômios de grau $2n-1$, assumimos:

$$h_i(x) = r_i(x)[l_i(x)]^2 \qquad \overline{h}_i(x) = s_i(x)[l_i(x)]^2, \tag{3.7}$$

sendo $r_i(x)$ e $s_i(x)$ funções lineares de x. Assim, $r_i(x)$ e $s_i(x)$ são tais que h_i e $\overline{h}_i$ satisfazem (3.5) e (3.6) para $i \neq j$. No caso em que $i = j$,

$$r_i(x_i) = 1 \qquad r_i'(x_i) + 2l_i'(x_i) = 0$$

e

$$s_i(x_i) = 0 \qquad s_i'(x_i) = 1.$$

De onde segue que

$$r_i(x) = 1 - 2l_i'(x_i)(x - x_i) \qquad s_i(x) = x - x_i. \tag{3.8}$$

Assim, de (3.4), (3.7) e (3.8), obtemos o polinômio desejado

$$P(x) = \sum_{i=1}^{n} h_i(x)f(x_i) + \sum_{i=1}^{n} \overline{h}_i(x)f'(x_i),$$

sendo

$$h_i(x) = [1 - 2l_i'(x_i)(x - x_i)][l_i(x)]^2$$

e

$$\overline{h}_i(x) = (x - x_i)[l_i(x)]^2.$$

3.2.2 Quadratura

Assumimos uma integral na forma $\int_a^b f(x)w(x)dx$, sendo $f(x)$ uma função integrável; $w(x)$, função peso e $g(x) = f(x)w(x)$ é chamada função produto.

O método da quadratura é empregado quando $f(x)$ é complexa ou mesmo analiticamente intratável. Nesses casos, substituímos $f(x)$ pelo polinômio interpolador $P(x)$ (polinômio que tem o mesmo valor que a função $f(x)$ em um conjunto discreto e finito de pontos desta função), no intervalo $[a, b]$ e resolvemos a integral $\int_a^b P(x)w(x)dx$.

Quadratura de Hermite

Para aproximar uma integral da forma $\int_a^b f(x)w(x)dx$ usando Quadratura de Hermite, a função $f(x)$ é substituída pelo polinômio interpolador de Hermite. Assim,

$$\int_a^b f(x)w(x)dx \approx \int_a^b P(x)w(x)dx = \sum_{i=1}^{n} H_i f(x_i) + \sum_{i=1}^{n} \overline{H}_i f'(x_i), \tag{3.9}$$

sendo H_i e $\overline{H}_i$ conhecidos como pesos de quadratura e expressos da forma

$$H_i = \int_a^b h_i(x)w(x)dx = \int_a^b [1 - 2l_i'(x_i)(x - x_i)][l_i(x)]^2 w(x)dx,$$

e

$$\overline{H}_i = \int_a^b \overline{h}_i(x)w(x) = \int_a^b (x - x_i)[l_i(x)]^2 w(x)dx. \tag{3.10}$$

Quadratura Gaussiana

A quadratura Gaussiana considera uma outra formulação de (3.9) na situação em que $\overline{H}_i$ desaparece e o polinômio continua com $2n-1$ graus.

Para isso, reescreveremos a definição (3.10) de forma equivalente

$$\overline{H}_i = \frac{1}{\pi'(x_i)} \int_a^b w(x)\pi(x)l_i(x)dx.$$

Se $\pi(x)$ é um polinômio ortogonal a $l_1(x), ..., l_n(x)$ em $[a, b]$, com relação à função peso $w(x)$, então

$$\int_a^b w(x)\pi(x)l_i(x)dx = 0,$$

e a equação (3.9) reduz-se a

$$\int_a^b f(x)w(x)dx \approx \sum_{i=1}^{n} H_i f(x_i).$$

Quadratura Gauss-Hermite

Nesta fórmula serão empregados os polinômios de Hermite, $H_n(x)$, os quais são

Tabela 3.1: Polinômios de Hermite

n	$H_n(x)$
0	1
1	$2x$
2	$4x^2 - 2$
3	$8x^3 - 12x$
4	$16x^4 - 48x^2 + 12$
5	$32x^5 - 160x^3 + 120x$

ortogonais no intervalo $(-\infty, \infty)$ frente à função peso $w(x) = e^{-x^2}$.

Relação de Recorrência:

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x).$$

Utilizaremos a quadratura Gauss-Hermite por sua relação com a densidade normal (gaussiana). Considere uma integral na forma

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx. \tag{3.11}$$

Na quadratura Gauss-Hermite a integral (3.11) é aproximada por

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx \approx \sum_{i=1}^{n} f(x_i)H_i,$$

sendo as abscissas x_i, $i = 1, \ldots, n$, os zeros do polinômio de Hermite de ordem n, chamadas de pontos da quadratura e H_i os pesos dados por

$$H_i = \frac{2^{n-1}n!\sqrt{\pi}}{n^2[H_{n-1}(x_i)]^2}.$$

A Tabela 3.2 fornece os valores de x_i e H_i para $n = 2$, 3 e 4. Uma listagem mais completa pode ser encontrada em Abramowitz e Stegun (1972, pag. 924, tab.25.10).

Tabela 3.2: Abcissas x_i (raízes dos polinômios de Hermite) e pesos H_i para integração de Gauss-Hermite

n	$\pm x_i$	H_i
2	0,707106781186548	0,886226925452
3	0,000000000000000	1,181635900604
	1,224744871391589	0,295408975150
4	0,524647623275290	0,804914090005
	1,650680123885785	0,081312835447

Quadratura Gauss-Hermite Adaptada

Liu e Pierce (1994) argumentam que os valores selecionados para $f(x)$ em (3.11), podem não ser de interesse e sugerem uma adaptação. Consideremos a seguinte integral:

$$\int_{-\infty}^{\infty} g(t)dt,$$

com $g(t) = L(t)\phi(t; \gamma, \theta)$, positiva, sendo ϕ a densidade normal (caso mais geral em que a média pode ser diferente de zero). Assim,

$$\int_{-\infty}^{\infty} g(t)dt = \int_{-\infty}^{\infty} L(t)\frac{1}{\sqrt{2\pi\theta}}exp\left\{-\frac{1}{2}\left(\frac{t-\gamma}{\sqrt{\theta}}\right)^2\right\}dt.$$

Chamando $x = \dfrac{t-\gamma}{\sqrt{2\theta}}$, consequentemente, $dt = \sqrt{2\theta}dx$. Logo,

$$\begin{aligned}\int_{-\infty}^{\infty} g(t)dt &= \int_{-\infty}^{\infty} L(\gamma+\sqrt{2\theta}x)\frac{1}{\sqrt{\pi}\sqrt{2\theta}}e^{-x^2}\sqrt{2\theta}dx \\ &= \int_{-\infty}^{\infty} L(\gamma+\sqrt{2\theta}x)\frac{1}{\sqrt{\pi}}e^{-x^2}dx\end{aligned}$$

Utilizando a quadratura Gauss-Hermite, temos

$$\int_{-\infty}^{\infty} L(t)\phi(t;\gamma,\theta)dt \approx \sum_{i=1}^{n} L(\gamma+\sqrt{2\theta}x_i)\frac{H_i}{\sqrt{\pi}}.$$

Para que os pontos de $g(t)$ fiquem em uma região adequada, Liu e Pierce (1994) sugerem obter as quantidades $\hat{\mu} = \arg\max g(t)$ e

$$\hat{\upsilon} = \left(-\frac{\partial^2 \log g(t)}{\partial t^2}|_{t=\hat{\mu}}\right)^{-1/2}.$$

Considerando $h(t) = \dfrac{g(t)}{\phi(t;\hat{\mu},\hat{\upsilon})}$, temos que

$$\begin{aligned}\int_{-\infty}^{\infty} g(t)dt &= \int_{-\infty}^{\infty} \frac{g(t)}{\phi(t;\hat{\mu},\hat{\upsilon})}\ \phi(t;\hat{\mu},\hat{\upsilon})dt \\ &= \int_{-\infty}^{\infty} h(t)\ \phi(t;\hat{\mu},\hat{\upsilon})dt \\ &= \int_{-\infty}^{\infty} h(t)\frac{1}{\sqrt{2\pi\hat{\upsilon}}}exp\left\{-\frac{1}{2}\left(\frac{t-\hat{\mu}}{\sqrt{\hat{\upsilon}}}\right)^2\right\}dt.\end{aligned}$$

Chamando, novamente, $x = \dfrac{t-\hat{\mu}}{\sqrt{2\hat{v}}}$, consequentemente, $dt = \sqrt{2\hat{v}}dx$. Logo,

$$\begin{aligned}
\int_{-\infty}^{\infty} g(t)dt &= \int_{-\infty}^{\infty} h(\hat{\mu}+\sqrt{2\hat{v}}x)\frac{1}{\sqrt{\pi}\sqrt{2\hat{v}}}e^{-x^2}\sqrt{2\hat{v}}dx \\
&= \int_{-\infty}^{\infty} h(\hat{\mu}+\sqrt{2\hat{v}}x)\frac{1}{\sqrt{\pi}}e^{-x^2}dx \\
&\approx \sum_{i=1}^{n}\frac{1}{\sqrt{\pi}}H_i h(\hat{\mu}+\sqrt{2\hat{v}}x_i) \\
&= \sqrt{2}\hat{v}\sum_{i=1}^{n} g(\hat{\mu}+\sqrt{2\hat{v}}x_i)e^{x_i^2}H_i.
\end{aligned}$$

Com o resultado acima, temos que a aproximação da integral pela quadratura Gauss-Hermite adaptada é dada por

$$\int_{-\infty}^{\infty} L(t)\ \phi(t;\gamma,\theta)dt \approx \sqrt{2\hat{v}}\sum_{i=1}^{n} L(\hat{\mu}+\sqrt{2\hat{v}}x_i)\ \phi(\hat{\mu}+\sqrt{2\hat{v}}x_i;\gamma,\theta)e^{x_i^2}H_i. \qquad (3.12)$$

O resultado acima pode ser implementado através do procedimento NLMIXED do *software* SAS para obter a solução da integral em (3.1). A maximização da verossimilhança com respeito ao vetor de parâmetros pode ser obtida através do algoritmo quase-Newton. No apêndice B listamos o programa usado para avaliar o conjunto de dados apresentado no Capítulo 5.

Capítulo 4

Estimação via Verossimilhança Penalizada

Neste capítulo apresentamos brevemente a abordagem utilizada pelo *software* R para estimar um MTFA com efeito aleatório. Realizamos um estudo de simulação para investigar a implementação desse método juntamente com o procedimento usual que ignora a dependência dos tempos.

4.1 Procedimento de Estimação

O procedimento de estimação que está implementado no *software* R ajusta modelos de regressão de Cox com fragilidade e MTFA com efeito aleatório, usando os comandos *coxph* e *survreg*, respectivamente, através da função de verossimilhança penalizada. Como opção para o efeito aleatório o R oferece as distribuições gama, gaussiana e t.

Na formulação de regressão penalizada o vetor de efeitos aleatórios $u = (u_1, ..., u_k)$ é tratado como um vetor de parâmetros que são controlados por uma função de penalização adicionada à verossimilhança correspondente à distribuição conjunta dos dados, como função dos parâmetros $\boldsymbol{\beta}$, σ e u.

Através de estudos de simulação, Zhang e Peng (2007) comparam esse método paramétrico, e o intitulam como método de Therneau, com um método não paramétrico proposto por Pan (2001) e com um novo método, também não paramétrico, proposto

por eles. Nesse artigo, o método que utiliza a penalização da verossimilhança foi empregado no modelo sugerido por Lambert et al. (2004). Zhang e Peng (2007) argumentam que Therneau, em comunicação pessoal, alerta que se a distribuição da fragilidade for gaussiana é necessário usar a opção AIC no código do ajuste para que os resultados sejam corretos.

Para o modelo de Cox, o procedimento de estimação é feito maximizando o logaritmo da verossimilhança parcial penalizada, lp_{pen}, que é dado por

$$lp_{pen} = lp(\boldsymbol{\beta}, u; \text{dados}) - h(u; \theta),$$

sendo lp o logaritmo da verossimilhança parcial de Cox e h é uma função de penalidade escolhida por investigação para restringir os valores de u.

O procedimento da verossimilhança penalizada para o modelo de regressão de Cox assim como sua implementação no *software* R encontram-se devidamente documentados (THERNEAU; GRAMBSCH, 2000; THERNEAU; GRAMBSCH; PANKRATZ, 2003). Contudo, não encontramos na literatura descrição abrangente para o MTFA, embora o R disponibilize a opção de incluir fragilidade nesse modelo.

Seguindo o raciocínio usado para o modelo de Cox, descrevemos a seguir como esse procedimento estaria supostamente implementado no *software* R para estimar um MTFA com um efeito aleatório. Assim, com a mesma notação empregada na Seção 3.1, em que $\lambda = (\alpha, \boldsymbol{\beta}^\mathsf{T}, \sigma, \theta)^\mathsf{T}$ é o vetor de parâmetros desconhecidos, temos que a verossimilhança condicional ao efeito aleatório e a verossimilhança condicional para os indivíduos do grupo i são, respectivamente,

$$L_{ij}(\alpha, \boldsymbol{\beta}, \sigma | u_i) = f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}}$$

e

$$L_i(\alpha, \boldsymbol{\beta}, \sigma | u_i) = \prod_{j=1}^{n_i} f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}}.$$

Assumindo independência entre os grupos, a verossimilhança condicional para toda a amostra é

$$L(\alpha, \boldsymbol{\beta}, \sigma|u_i) = \prod_{i=1}^{k}\prod_{j=1}^{n_i} f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}}.$$

Daí segue que o logaritmo da função de verossimilhança condicional para toda a amostra é da forma

$$\begin{aligned} l(\alpha, \boldsymbol{\beta}, \sigma|u_i) &= \log \prod_{i=1}^{k}\prod_{j=1}^{n_i} f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \log f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}}. \end{aligned} \tag{4.1}$$

Na abordagem de verossimilhança penalizada é necessário incluir em (4.1) uma função de penalidade. De forma que a função logaritmo da verossimilhança penalizada, l_{pen}, é dada por

$$\begin{aligned} l_{pen}(\lambda) &= l(\alpha, \boldsymbol{\beta}, \sigma|u_i) - h(u;\theta) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \log f(y_{ij}|u_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|u_i, \mathbf{x}_{ij})^{1-\delta_{ij}} - h(u;\theta). \end{aligned}$$

Em nosso caso, Y_{ij} tem distribuição valor extremo, como dada em (2.2), com $\mu = \alpha + u_i + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}$. Na página de discussão do R, Therneau justifica que quando o efeito aleatório tem distribuição normal, a função de penalização é escrita como

$$h(u;\theta) = \frac{1}{2\theta}\sum_{i=1}^{n} u_i^2.$$

No processo de estimação descrito em Therneau, Grambsch e Pankratz (2003) o parâmetro θ é considerado constante. Para cada θ fixo o modelo penalizado é resolvido pelo método iterativo de Newton-Raphson que fornece as estimativas dos parâmetros assim como o valor da verossimilhança penalizada ajustada. Os autores argumentam que a estimativa de θ é obtida pela maximização do logaritmo da verossimilhança marginal perfilada (integrada com respeito aos u_i's). Para a verossimilhança parcial de Cox eles usam os resultados dados em Ripatti e Palmgren (2000) baseados na apro-

ximação de Laplace para obter o logaritmo da verossimilhança marginal aproximada. Para suprir a deficiência de bibliografia, apresentamos na próxima seção um estudo de simulação para verificar a validade da implementação desse método no *software* R.

4.2 Simulação

Apresentamos um estudo de simulação considerando um MTFA para avaliar o desempenho da implementação do método que utiliza a verossimilhança penalizada, considerando a distribuição normal para o efeito aleatório, e comparamos com um método que supõe que os tempos de vida são todos independentes, ambos implementados no *software* R (versão 2.7.1).

Para este estudo assumimos grupos com três tamanhos distintos: 10, 100 e 500, com 5 indivíduos em cada grupo. Consideramos que temos uma única covariável x_{ij} associada a cada indivíduo j no grupo i, com distribuição normal padrão. Os tempos de censura foram gerados a partir de uma distribuição uniforme $(0,\ q)$, onde q é escolhido de forma a fornecer 0%, 10%, 30% e 60% de censura na amostra. Os erros aleatórios foram gerados a partir de variáveis aleatórias independentes e identicamente distribuídas com distribuição valor extremo padrão. Realizamos 1000 réplicas para cada amostra e os verdadeiros valores para os coeficientes do modelo foram $\alpha = 5$, $\beta = 0.5$ e $\sigma = 1$. Além disso, consideramos cinco valores para a variância do efeito aleatório: θ=0 (sem dependência entre os tempos), θ=0.25, θ=0.5, θ=0.75 e θ=1.

As estimativas da média, erro padrão (EP) e raiz do erro quadrático médio (REQM) para os dados simulados estão nas Tabelas 4.1, 4.2, 4.3 e 4.4.

4.2.1 Resultados

As Tabelas 4.1, 4.2, 4.3 e 4.4 mostram que, em geral, as estimativas fornecidas pelo método que supõe independência são maiores que as dadas pelo método que trata a dependência entre os tempos. A maior diferença entre os dois métodos foi observada quanto às estimativas de σ. Com o aumento da variância do efeito aleatório, as es-

Tabela 4.1: Estimativas da média, erro padrão e raiz do erro quadrático médio dos parâmetros com $\alpha = 5$, $\beta = 0.5$, $\sigma = 1$, 1000 réplicas, amostras divididas em 10, 100 e 500 grupos, sem censura e variando o valor do parâmetro θ.

θ	Par	Método	$k = 10$			$k = 100$			k=500		
			Média	EP	REQM	Média	EP	REQM	Média	EP	REQM
0	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.0768	0.1229	0.1449	0.0196	0.0292	0.0352	0.0070	0.0114	0.0133
	α	Ind	4.9964	0.1510	0.1510	4.9971	0.0481	0.0482	5.0004	0.0209	0.0209
		Dep	4.9731	0.1543	0.1566	4.9889	0.0498	0.0511	4.9972	0.0215	0.0217
	β	Ind	0.5011	0.1908	0.1908	0.4983	0.0436	0.0436	0.4985	0.0205	0.0206
		Dep	0.4977	0.1944	0.1944	0.4986	0.0437	0.0438	0.4985	0.0205	0.0206
	σ	Ind	0.9703	0.1048	0.1089	0.9964	0.0351	0.0353	0.9997	0.0158	0.0158
		Dep	0.9164	0.1216	0.1476	0.9764	0.0441	0.0500	0.9918	0.0206	0.0222
0.25	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.3183	0.2553	0.2643	0.3331	0.0971	0.1278	0.3343	0.0535	0.0998
	α	Ind	5.0399	0.2273	0.2308	5.0538	0.0702	0.0884	5.0572	0.0315	0.0653
		Dep	4.9583	0.2252	0.2290	4.9584	0.0688	0.0804	4.9605	0.0312	0.0503
	β	Ind	0.5048	0.1769	0.1770	0.4984	0.0520	0.0520	0.5000	0.0233	0.0233
		Dep	0.5025	0.1722	0.1722	0.4998	0.0473	0.0473	0.5008	0.0219	0.0219
	σ	Ind	1.0859	0.1336	0.1589	1.1345	0.0429	0.1411	1.1371	0.0195	0.1385
		Dep	0.8992	0.1301	0.1646	0.9042	0.0422	0.1047	0.9023	0.0197	0.0997
0.5	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.5835	0.3706	0.3799	0.5915	0.1101	0.1432	0.5773	0.0482	0.0911
	α	Ind	5.0774	0.2823	0.2928	5.1061	0.0848	0.1359	5.1067	0.0391	0.1137
		Dep	4.9427	0.2803	0.2861	4.9542	0.0827	0.0945	4.9547	0.0380	0.0592
	β	Ind	0.4968	0.2013	0.2013	0.5001	0.0642	0.0642	0.4996	0.0282	0.0283
		Dep	0.4985	0.1760	0.1760	0.4995	0.0523	0.0523	0.5008	0.0230	0.0230
	σ	Ind	1.1856	0.1563	0.2426	1.2476	0.0533	0.2533	1.2550	0.0249	0.2562
		Dep	0.8814	0.1215	0.1698	0.8854	0.0389	0.1210	0.8898	0.0175	0.1116
0.75	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.8880	0.5912	0.6071	0.8547	0.1915	0.2183	0.8587	0.0900	0.1412
	α	Ind	5.1242	0.3118	0.3356	5.1526	0.1002	0.1825	5.1521	0.0468	0.1592
		Dep	4.9448	0.2999	0.3049	4.9525	0.0982	0.1091	4.9499	0.0449	0.0673
	β	Ind	0.4959	0.2579	0.2579	0.5014	0.0677	0.0677	0.5002	0.0302	0.0303
		Dep	0.4920	0.1907	0.1908	0.5016	0.0516	0.0516	0.5005	0.0229	0.0229
	σ	Ind	1.2688	0.1830	0.3252	1.3520	0.0670	0.3583	1.3595	0.0302	0.3608
		Dep	0.8698	0.1184	0.1760	0.8810	0.0389	0.1252	0.8812	0.0167	0.1200
1	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	1.1467	0.7194	0.7342	1.1108	0.2129	0.2400	1.0909	0.1037	0.1379
	α	Ind	5.1717	0.3666	0.4048	5.1944	0.1159	0.2263	5.2004	0.0506	0.2067
		Dep	4.9535	0.3517	0.3548	4.9614	0.1105	0.1171	4.9512	0.0495	0.0695
	β	Ind	0.4903	0.1925	0.1928	0.5008	0.0708	0.0708	0.4989	0.0351	0.0352
		Dep	0.4963	0.1661	0.1661	0.5008	0.0514	0.0514	0.4991	0.0242	0.0242
	σ	Ind	1.3684	0.2202	0.4292	1.4483	0.0762	0.4548	1.4571	0.0362	0.4585
		Dep	0.8675	0.1215	0.1797	0.8751	0.0376	0.1304	0.8780	0.0167	0.1232

Tabela 4.2: Estimativas da média, erro padrão e raiz do erro quadrático médio dos parâmetros com $\alpha = 5$, $\beta = 0.5$, $\sigma = 1$, 1000 réplicas, amostras divididas em 10, 100 e 500 grupos, com 10% de censura e variando o valor do parâmetro θ.

θ	Par	Método	$k = 10$			$k = 100$			k=500		
			Média	EP	REQM	Média	EP	REQM	Média	EP	REQM
0	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.0698	0.1159	0.1353	0.0212	0.0319	0.0383	0.0086	0.0134	0.0159
	α	Ind	4.9895	0.1518	0.1522	4.9983	0.0488	0.0488	4.9986	0.0224	0.0224
		Dep	4.9664	0.1552	0.1588	4.9887	0.0501	0.0514	4.9944	0.0232	0.0239
	β	Ind	0.4949	0.1499	0.1499	0.5006	0.0510	0.0510	0.5001	0.0210	0.0210
		Dep	0.4965	0.1512	0.1512	0.5004	0.0509	0.0509	0.5000	0.0210	0.0210
	σ	Ind	0.9755	0.1191	0.1216	0.9982	0.0366	0.0366	0.9997	0.0172	0.0172
		Dep	0.9277	0.1285	0.1475	0.9777	0.0461	0.0512	0.9905	0.0219	0.0239
0.25	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.3423	0.2849	0.2995	0.3172	0.0972	0.1181	0.3233	0.0580	0.0935
	α	Ind	5.0317	0.2244	0.2266	5.0517	0.0694	0.0865	5.0483	0.0324	0.0582
		Dep	4.9431	0.2214	0.2286	4.9542	0.0693	0.0831	4.9481	0.0323	0.0612
	β	Ind	0.4934	0.1700	0.1701	0.4981	0.0571	0.0571	0.4932	0.0254	0.0263
		Dep	0.4988	0.1614	0.1614	0.5009	0.0539	0.0539	0.4958	0.0239	0.0243
	σ	Ind	1.0835	0.1342	0.1580	1.1134	0.0450	0.1220	1.1188	0.0191	0.1203
		Dep	0.8926	0.1394	0.1760	0.8998	0.0458	0.1102	0.9001	0.0207	0.1020
0.5	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	1.1604	12.9598	12.9767	0.5884	0.1157	0.1456	0.5748	0.0484	0.0891
	α	Ind	5.0784	0.2849	0.2955	5.0922	0.0889	0.1281	5.0942	0.0396	0.1022
		Dep	4.9475	0.2796	0.2844	4.9396	0.0863	0.1053	4.9404	0.0384	0.0708
	β	Ind	0.5107	0.2011	0.2014	0.4914	0.0657	0.0663	0.4883	0.0289	0.0312
		Dep	0.5107	0.1872	0.1875	0.4992	0.0565	0.0565	0.4951	0.0244	0.0249
	σ	Ind	1.1588	0.1566	0.2231	1.2124	0.0501	0.2182	1.2172	0.0222	0.2183
		Dep	0.8768	0.1340	0.1821	0.8780	0.0413	0.1288	0.8816	0.0183	0.1198
0.75	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	2.6041	23.2583	23.3321	0.8749	0.2228	0.2554	0.8805	0.1416	0.1926
	α	Ind	5.1042	0.3188	0.3354	5.1348	0.1005	0.1682	5.1330	0.0461	0.1408
		Dep	4.9276	0.3118	0.3201	4.9381	0.0978	0.1157	4.9347	0.0447	0.0791
	β	Ind	0.5030	0.1973	0.1973	0.4878	0.0667	0.0677	0.4826	0.0327	0.0371
		Dep	0.5036	0.1743	0.1744	0.4974	0.0536	0.0537	0.4954	0.0248	0.0252
	σ	Ind	1.2487	0.1781	0.3059	1.3012	0.0555	0.3063	1.3054	0.0263	0.3066
		Dep	0.8669	0.1269	0.1839	0.8710	0.0394	0.1349	0.8723	0.0183	0.1291
1	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	8.5155	75.9880	76.3588	1.1915	0.2854	0.3437	1.1560	0.1425	0.2112
	α	Ind	5.1146	0.3555	0.3735	5.1727	0.1148	0.2074	5.1684	0.0509	0.1759
		Dep	4.9087	0.3498	0.3615	4.9358	0.1109	0.1281	4.9302	0.0501	0.0859
	β	Ind	0.4905	0.2263	0.2265	0.4842	0.0688	0.0706	0.4802	0.0339	0.0393
		Dep	0.4970	0.1746	0.1746	0.4989	0.0529	0.0529	0.4963	0.0256	0.0258
	σ	Ind	1.3065	0.1968	0.3642	1.3858	0.0647	0.3911	1.3887	0.0284	0.3897
		Dep	0.8618	0.1298	0.1896	0.8701	0.0383	0.1354	0.8686	0.0172	0.1323

Tabela 4.3: Estimativas da média, erro padrão e raiz do erro quadrático médio dos parâmetros com $\alpha = 5$, $\beta = 0.5$, $\sigma = 1$, 1000 réplicas, amostras divididas em 10, 100 e 500 grupos, com 30% de censura e variando o valor do parâmetro θ.

θ	Par	Método	$k = 10$			$k = 100$			k=500		
			Média	EP	REQM	Média	EP	REQM	Média	EP	REQM
0	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.0888	0.1598	0.1828	0.0242	0.0352	0.0428	0.0104	0.0167	0.0196
	α	Ind	4.9955	0.1730	0.1731	4.9980	0.0579	0.0579	5.0001	0.0234	0.0234
		Dep	4.9626	0.1799	0.1838	4.9848	0.0601	0.0620	4.9940	0.0251	0.0258
	β	Ind	0.5131	0.1929	0.1934	0.5025	0.0538	0.0539	0.5011	0.0243	0.0243
		Dep	0.5092	0.1932	0.1934	0.5012	0.0537	0.0537	0.5005	0.0243	0.0243
	σ	Ind	0.9776	0.1299	0.1319	0.9989	0.0422	0.0423	0.9986	0.0192	0.0193
		Dep	0.9224	0.1427	0.1625	0.9771	0.0522	0.0570	0.9887	0.0247	0.0272
0.25	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	1.6168	13.5970	13.6655	0.2983	0.1052	0.1158	0.2909	0.0528	0.0668
	α	Ind	5.0157	0.2314	0.2319	5.0325	0.0742	0.0810	5.0326	0.0329	0.0463
		Dep	4.9310	0.2292	0.2394	4.9259	0.0729	0.1040	4.9234	0.0326	0.0833
	β	Ind	0.5101	0.2034	0.2036	0.4900	0.0616	0.0624	0.4872	0.0278	0.0306
		Dep	0.5082	0.2014	0.2016	0.4885	0.0578	0.0589	0.4855	0.0258	0.0296
	σ	Ind	1.0527	0.1501	0.1591	1.0847	0.0471	0.0969	1.0875	0.0202	0.0898
		Dep	0.8906	0.1534	0.1884	0.8943	0.0546	0.1190	0.8942	0.0234	0.1084
0.5	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	5.9572	30.9291	31.4069	0.5835	0.1683	0.1879	0.5655	0.0580	0.0875
	α	Ind	5.0469	0.2903	0.2941	5.0631	0.0914	0.1110	5.0593	0.0397	0.0713
		Dep	4.9237	0.2965	0.3061	4.9034	0.0868	0.1299	4.8992	0.0381	0.1078
	β	Ind	0.4958	0.2095	0.2095	0.4808	0.0669	0.0696	0.4787	0.0294	0.0363
		Dep	0.4998	0.2013	0.2013	0.4843	0.0626	0.0646	0.48421	0.0259	0.0304
	σ	Ind	1.1326	0.1598	0.2077	1.1623	0.0512	0.1702	1.1640	0.0237	0.1657
		Dep	0.8704	0.1487	0.1972	0.8673	0.0463	0.1405	0.8671	0.0203	0.1345
0.75	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	11.8264	47.3586	48.6366	0.9275	0.2979	0.3467	0.9059	0.1799	0.2381
	α	Ind	5.0949	0.3331	0.3464	5.0868	0.1021	0.1340	5.0879	0.0468	0.0996
		Dep	4.9463	0.3515	0.3555	4.8923	0.0994	0.1466	4.8927	0.0449	0.1163
	β	Ind	0.5043	0.2211	0.2212	0.4736	0.0686	0.0735	0.4724	0.0325	0.0426
		Dep	0.5080	0.1991	0.1993	0.4875	0.0601	0.0614	0.4853	0.0274	0.0311
	σ	Ind	1.2011	0.1837	0.2724	1.2312	0.0572	0.2382	1.2356	0.0253	0.2369
		Dep	0.8597	0.1509	0.2060	0.8531	0.0443	0.1534	0.8562	0.0189	0.1451
1	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	19.3079	63.3116	65.9055	1.3880	0.5444	0.6685	1.2926	0.2213	0.3669
	α	Ind	5.0824	0.3655	0.3747	5.1166	0.1091	0.1596	5.1140	0.0522	0.1254
		Dep	4.9230	0.3834	0.3911	4.8937	0.1065	0.1504	4.8887	0.0514	0.1226
	β	Ind	0.4785	0.2426	0.2435	0.4694	0.0791	0.0848	0.4678	0.0338	0.0467
		Dep	0.4980	0.2081	0.2081	0.4898	0.0619	0.0628	0.4886	0.0285	0.0306
	σ	Ind	1.2435	0.1941	0.3113	1.2971	0.0638	0.3038	1.2988	0.0277	0.3001
		Dep	0.8524	0.1455	0.2073	0.8501	0.0414	0.1555	0.8500	0.0188	0.1512

Tabela 4.4: Estimativas da média, erro padrão e raiz do erro quadrático médio dos parâmetros com $\alpha = 5$, $\beta = 0.5$, $\sigma = 1$, 1000 réplicas, amostras divididas em 10, 100 e 500 grupos, com 60% de censura e variando o valor do parâmetro θ.

θ	Par	Método	$k = 10$			$k = 100$			k=500		
			Média	EP	REQM	Média	EP	REQM	Média	EP	REQM
0	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	0.4636	3.1762	3.2099	0.0314	0.0444	0.0544	0.0141	0.0220	0.0262
	α	Ind	5.0098	0.2703	0.2705	5.0017	0.0861	0.0861	5.0014	0.0395	0.0395
		Dep	4.9542	0.2737	0.2775	4.9755	0.0898	0.0931	4.9889	0.0433	0.0447
	β	Ind	0.5136	0.3060	0.3063	0.5061	0.0780	0.0782	0.49876	0.0357	0.0357
		Dep	0.5061	0.3067	0.3068	0.5009	0.0775	0.0775	0.4962	0.0357	0.0359
	σ	Ind	0.9778	0.1873	0.1886	0.9981	0.0580	0.0580	0.9997	0.0277	0.0277
		Dep	0.9071	0.1974	0.2182	0.9718	0.0670	0.0726	0.9874	0.0333	0.0357
0.25	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	2.2198	9.9467	10.1399	0.2267	0.1148	0.1171	0.2191	0.0516	0.0602
	α	Ind	5.0165	0.3394	0.3398	4.9879	0.0980	0.0988	4.9926	0.0442	0.0448
		Dep	4.9327	0.3665	0.3726	4.8467	0.0961	0.1809	4.8448	0.0457	0.1618
	β	Ind	0.4977	0.3204	0.3204	0.4891	0.0801	0.0808	0.4878	0.0345	0.0366
		Dep	0.5001	0.3409	0.3409	0.4679	0.0755	0.0820	0.4632	0.0324	0.0490
	σ	Ind	1.0278	0.1977	0.1997	1.0455	0.0611	0.0761	1.0488	0.0279	0.0563
		Dep	0.8791	0.1994	0.2332	0.8887	0.0724	0.1328	0.8874	0.0360	0.1182
0.5	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	6.0173	19.2472	20.0223	0.5149	0.3963	0.3966	0.4682	0.0741	0.0807
	α	Ind	5.0146	0.3923	0.3926	4.9950	0.1114	0.1115	4.9874	0.0489	0.0505
		Dep	4.9248	0.4348	0.4413	4.7902	0.0998	0.2323	4.7761	0.0421	0.2278
	β	Ind	0.4941	0.2993	0.2993	0.4813	0.0860	0.0880	0.4794	0.0380	0.0433
		Dep	0.4981	0.2964	0.2964	0.4584	0.0776	0.0880	0.4565	0.0341	0.0553
	σ	Ind	1.0759	0.2096	0.2229	1.0973	0.0641	0.1165	1.0980	0.0280	0.1020
		Dep	0.8596	0.2016	0.2456	0.8402	0.0623	0.1715	0.8369	0.0271	0.1654
0.75	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	9.5328	24.7951	26.3047	1.0059	1.1267	1.1554	0.8017	0.1890	0.1959
	α	Ind	4.9727	0.3913	0.3923	4.9909	0.1188	0.1191	4.9921	0.0518	0.0524
		Dep	4.8907	0.4502	0.4633	4.7683	0.1108	0.2569	4.7554	0.0450	0.2487
	β	Ind	0.5044	0.2777	0.2777	0.4757	0.0850	0.0885	0.4725	0.0371	0.0462
		Dep	0.5107	0.2838	0.2840	0.4665	0.0792	0.0860	0.4606	0.0327	0.0511
	σ	Ind	1.1084	0.2110	0.2372	1.1448	0.0664	0.1593	1.1472	0.0295	0.1501
		Dep	0.8332	0.1847	0.2489	0.8197	0.0582	0.1894	0.8172	0.0245	0.1845
1	θ	Ind	-	-	-	-	-	-	-	-	-
		Dep	14.1101	35.1998	37.5619	1.9736	2.4153	2.6041	1.3963	0.4265	0.5822
	α	Ind	5.0166	0.5094	0.5096	4.9831	0.1247	0.1258	4.9966	0.0572	0.0573
		Dep	4.9432	0.5467	0.5496	4.7686	0.1323	0.2666	4.7634	0.0549	0.2429
	β	Ind	0.5042	0.3219	0.3220	0.4643	0.0915	0.0983	0.4646	0.0407	0.0539
		Dep	0.5157	0.2983	0.2987	0.4699	0.0850	0.0902	0.4688	0.0355	0.0473
	σ	Ind	1.1603	0.2256	0.2767	1.1888	0.0678	0.2005	1.1919	0.0323	0.1946
		Dep	0.8397	0.1875	0.2467	0.8138	0.0504	0.1930	0.8112	0.0229	0.1902

timativas obtidas pelo modelo sob independência tendem a crescer, tornando-se mais distantes do verdadeiro valor. O contrário acontece com o método que utiliza a penalização da verossimilhança, cujas estimativas de σ diminuem com o aumento de θ, mas também afastam-se do valor desejado. Essa regularidade foi notada para todos os percentuais de censura.

Quanto ao crescimento dos grupos e, consequentemete, com o aumento do tamanho da amostra, evidenciamos que os valores de σ crescem e se afastam do valor almejado se ignorarmos a dependência entre os tempos (exceto quando temos $\theta = 0$) . As estimativas fornecidas supondo dependência também crescem com o aumento dos grupos, mas, ao contrário das obtidas sob independência, tornam-se cada vez mais próximas de 1. Essa tendência é melhor evidenciada na ausência de censura. De acordo com o aumento do percentual de censura, isso só se confirma quando θ é pequeno.

Aumentando a proporção de censura, as estimativas dadas pelos dois métodos tendem a diminuir. Para o modelo com efeito aleatório elas tornam-se mais distantes do verdadeiro valor. O contrário ocorre para o método usual, uma vez que esse método superestima as estimativas de σ. De forma que podemos acreditar que, com exceção do caso em que $\theta = 0$, o aumento do percentual de censura aperfeiçoa a estimava do σ para tal método. É importante ressaltar, porém, que embora as estimativas sob este método estejam se aproximando dos valores desejados, o método que trata a dependência ainda fornece estimativas menos viesadas para σ.

Nas Figuras 4.1, 4.2 e 4.3 apresentamos alguns gráficos para ilustrar as interpretações obtidas.

Quanto às estimativas do parâmetro θ, observamos que quando não há correlação entre os tempos de vida, são próximas do verdadeiro valor, independente da quantidade de grupos e do percentual de censura. Entretanto, com o aumento da variância do efeito aleatório, notamos que o método é extremamente sensível ao número de grupos, em geral isso é mais acentuado quando o percentual de censura cresce. Através de um estudo de simulação em menor escala notamos que, dimunuindo o número de grupos mas conservando o tamanho da amostra, os resultados são ainda mais distorci-

dos. Na ausência de censura as estimativas obtidas se distanciam muito pouco com o crescimento do valor do parâmetro θ. A Figura 4.4 nos auxilia nessas conclusões.

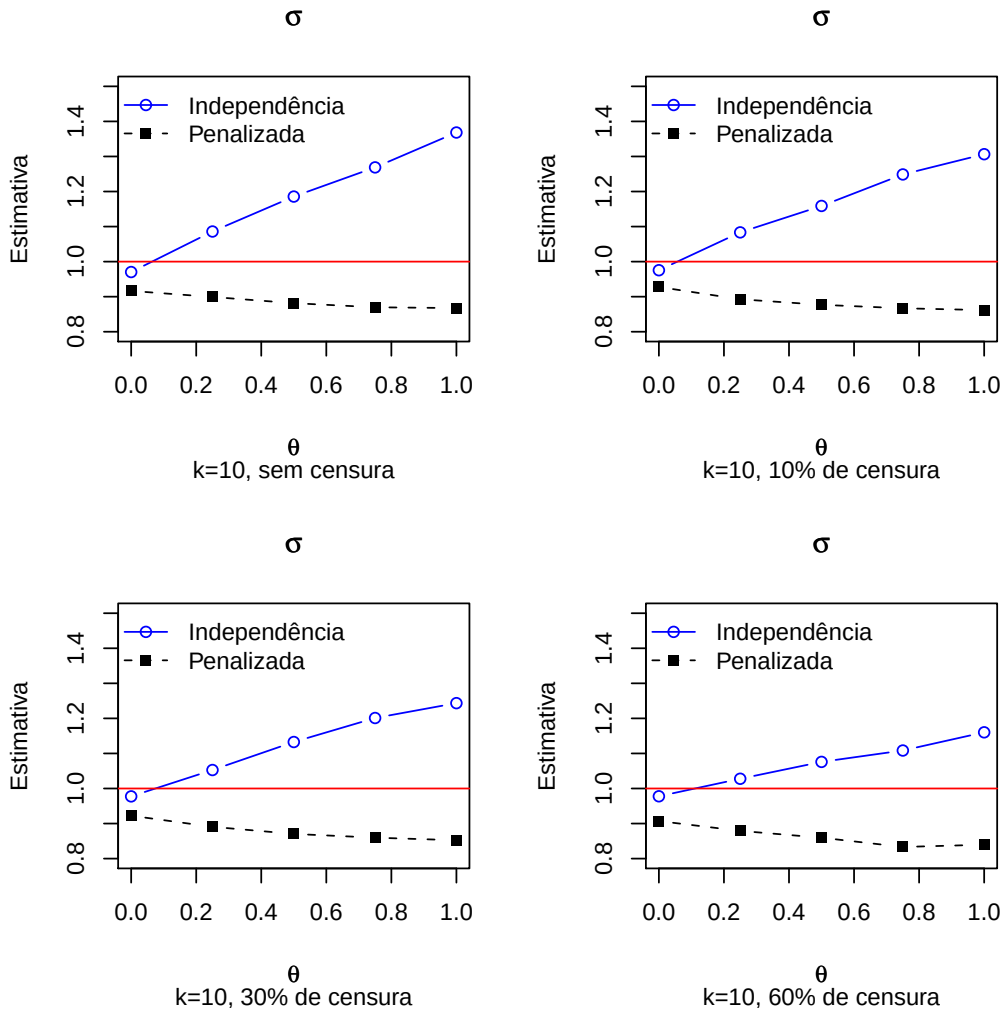

Figura 4.1: Estimativas do parâmetro σ para o método que assume independência e para o método baseado na verossimilhança penalizada para 0%, 10%, 30% e 60% de censura, com 10 grupos.

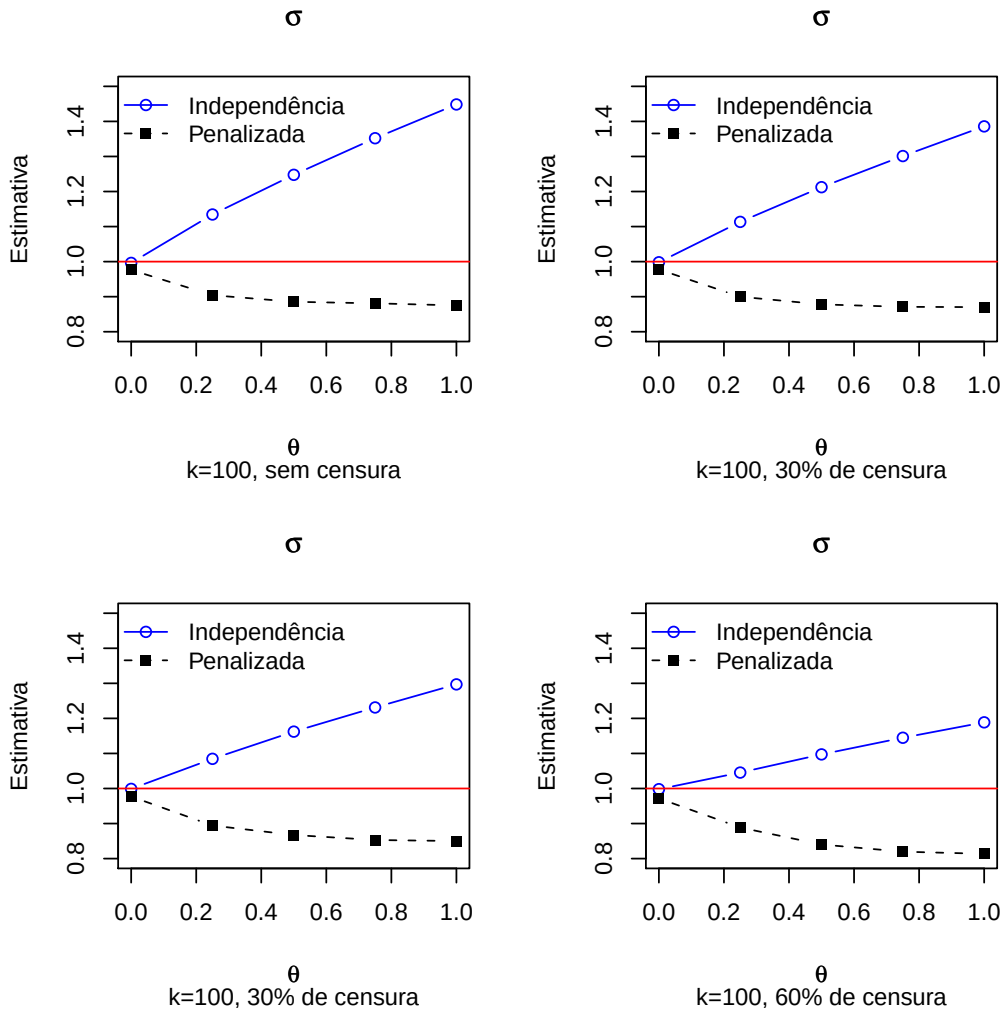

Figura 4.2: Estimativas do parâmetro σ para o método que assume independência e para o método baseado na verossimilhança penalizada para 0%, 10%, 30% e 60% de censura, com 100 grupos.

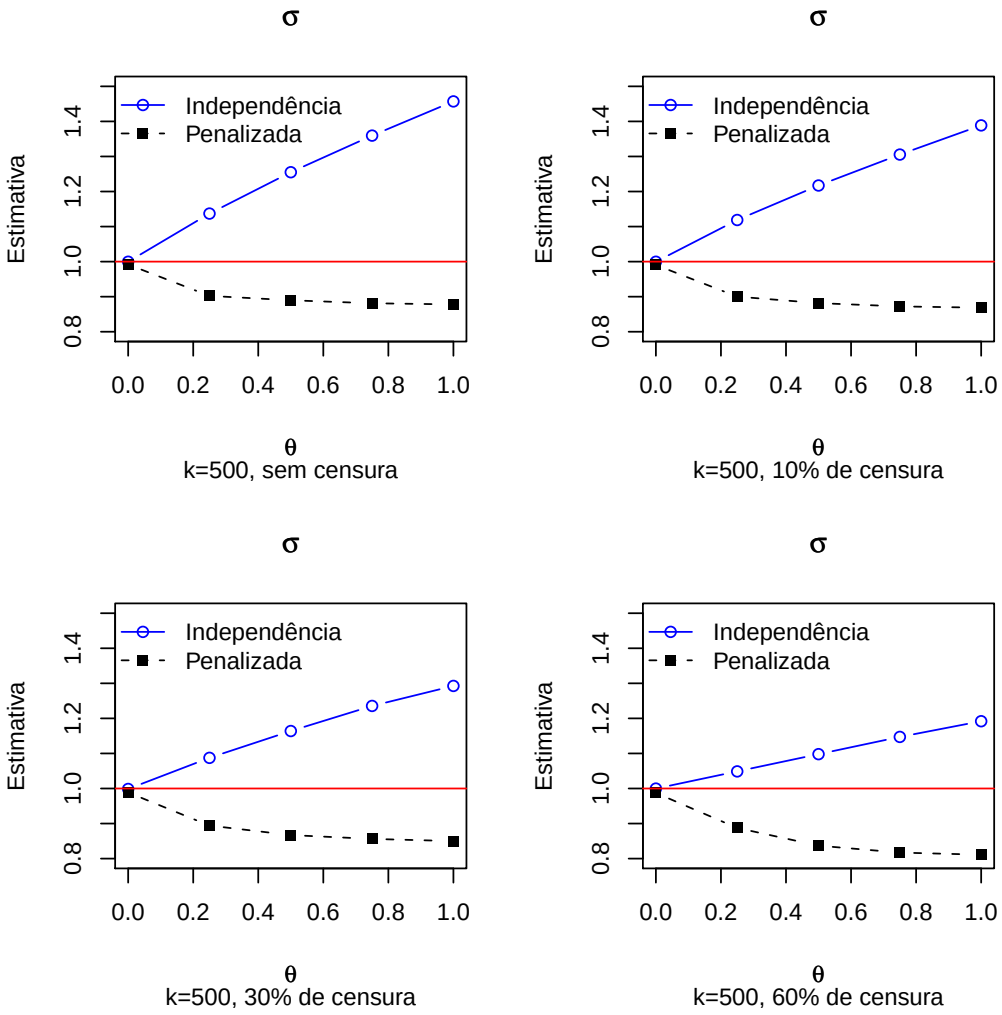

Figura 4.3: Estimativas do parâmetro σ para o método que assume independência e para o método baseado na verossimilhança penalizada para 0%, 10%, 30% e 60% de censura, com 500 grupos.

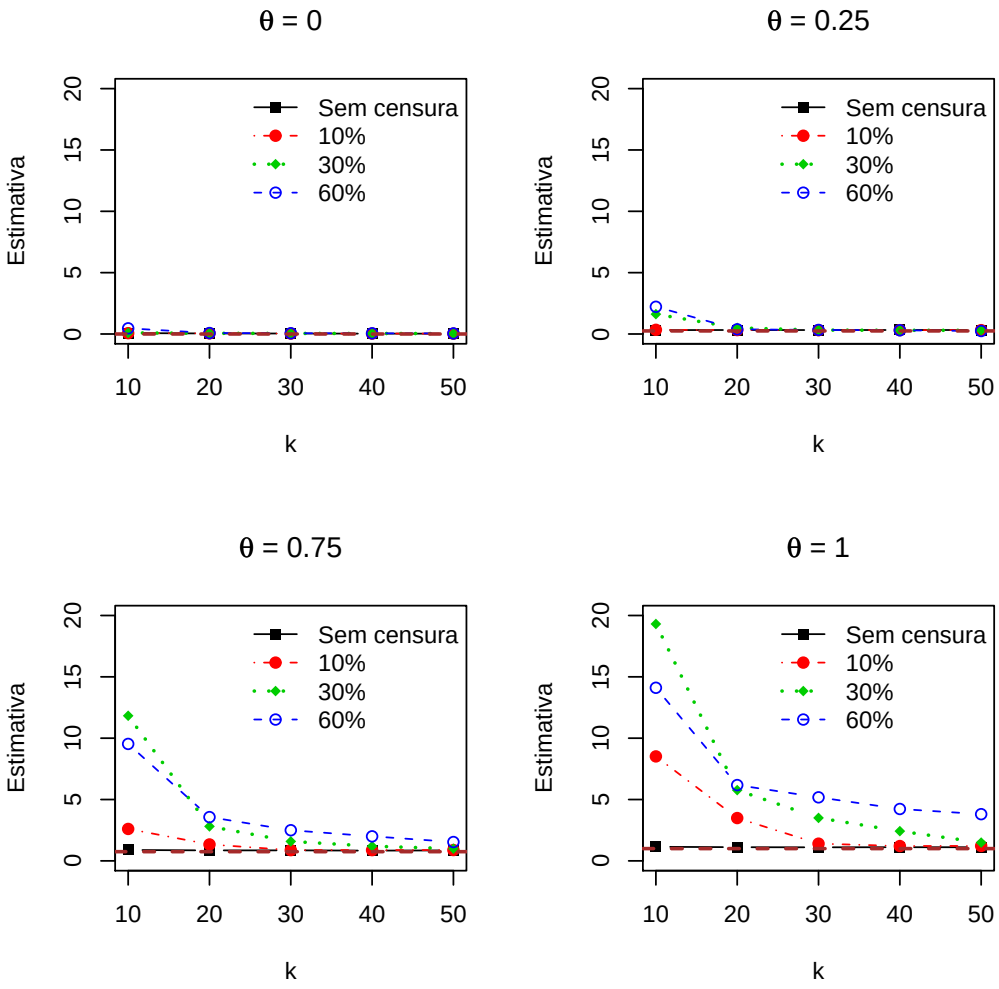

Figura 4.4: Estimativas do parâmetro θ para 0%, 10%, 30% e 60% de censura, variando o número de grupos.

Capítulo 5

Aplicação

Como vimos no Capítulo 1, um dos objetivos deste trabalho é avaliar o relacionamento entre o tempo de funcionamento de poços de petróleo com características relacionadas a eles. Usaremos, para isso, os métodos apresentados para dados correlacionados que vimos anteriormente. Esse é o assunto deste capítulo.

5.1 Análise Exploratória

Antes de obtermos o modelo ajustado, faremos uma exploração gráfica dos dados, sob a suposição de independência, estimando a função de sobrevivência através do estimador não-paramétrico de Kaplan-Meier, para cada covariável selecionada. São elas:

- *Produção base do poço-coluna*, medida em m^3/dia;
- *Método de elevação.* Embora a Bacia Potiguar apresente poços com energia suficiente para a elevação dos fluidos (poços surgentes), na sua grande maioria é necessário a instalação de equipamentos que exerçam pressão no fundo do poço para a elevação do fluido até a superfície. Neste estudo os dois métodos artificiais considerados foram BM - Bombeio mecânico e BCP - Bombeio por Cavidade Progressiva.
- *Idade do poço-coluna*, medida no momento da falha, em anos;

- *Unidade que administra os poços-coluna* de acordo com a sua localização geográfica. São quatro unidades:
  - ARG: Unidade Operacional Alto do Rodrigues
  - CAM: Unidade Operacional Canto do Amaro
  - ET: Unidade Operacional Campo de Estreito
  - RFQ: Unidade Operacional Fazenda Riacho da Forquilha
- *Profundidade* onde se encontra instalada a bomba de produção do poço-coluna, em metros.

Na Figura 5.1 observamos que poços-coluna com alta produção aparentam ter menor tempo de funcionamento do que poços-coluna com baixa produção. Em relação ao método de elevação, o bombeio mecânico parece proporcionar maior tempo de funcionamento dos poços-coluna do que o bombeio por cavidade progressiva.

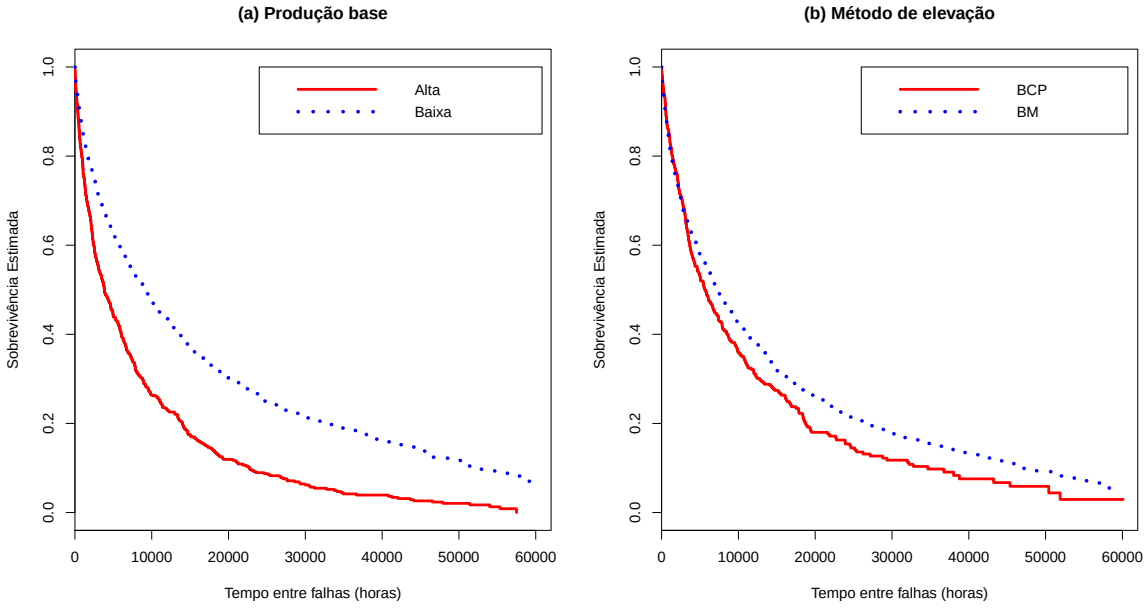

Figura 5.1: Curva de sobrevivência estimada pelo método de Kaplan-Meier por (a) produção base: baixa ($\leq 3,91\ m^3$/dia) e alta ($> 3,91\ m^3$/dia) e por (b) método de elevação.

Em relação a idade, os poços-coluna mais velhos aparentam maior tempo de sobrevivência se comparados com poços-coluna novos, conforme Figura 5.2. Quanto às unidades administrativas os poços-coluna localizados nas regiões Canto do Amaro e Fazenda Riacho da Forquilha apresentam maior sobrevivência do que os das regiões Alto do Rodrigues e Estreito.

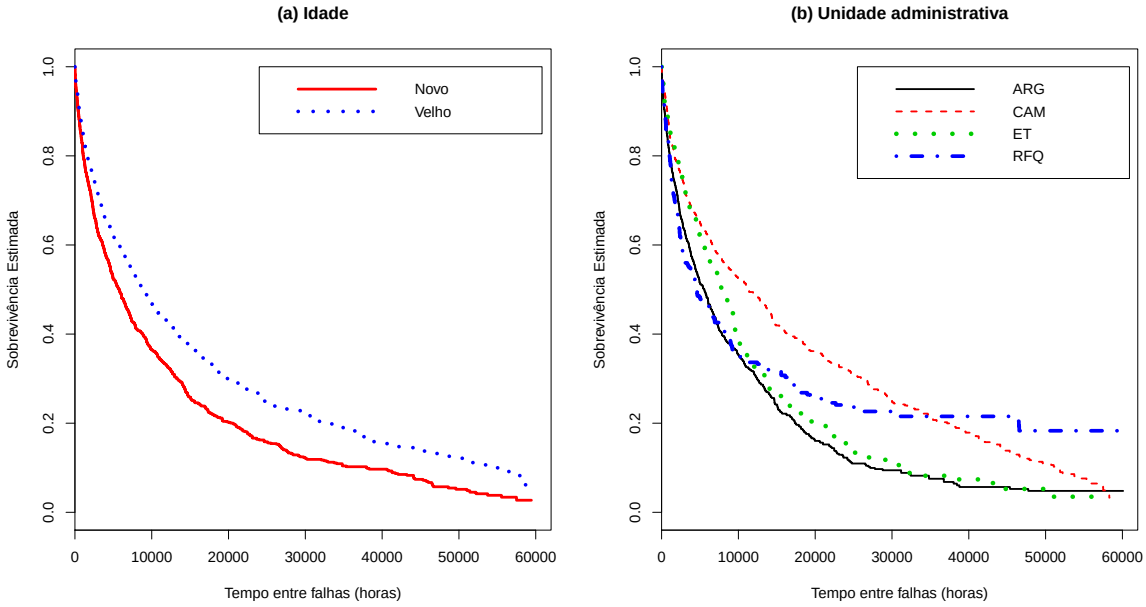

Figura 5.2: Curva de sobrevivência estimada pelo método de Kaplan-Meier por (a) idade : poço novo($\leq 9,22$ anos) e poço velho ($> 9,22$ anos) e por (b) unidade administrativa.

Finalmente, a Figura 5.3 indica que poços-coluna com instalação da bomba mais profunda parecem sobreviver mais.[1]

5.2 Ajuste dos dados

Seja y_{ij} o logaritmo do tempo decorrido até o poço-coluna i parar de funcionar ou ser censurado, na $j-$ésima recorrência. Em um estudo realizado em Dantas (2008), que utiliza esse conjunto de dados considerando apenas o tempo até a primeira falha,

[1] As covariáveis produção, idade e profundidade da bomba foram divididas em categorias que estão relacionadas com o valor médio medido para cada covariável.

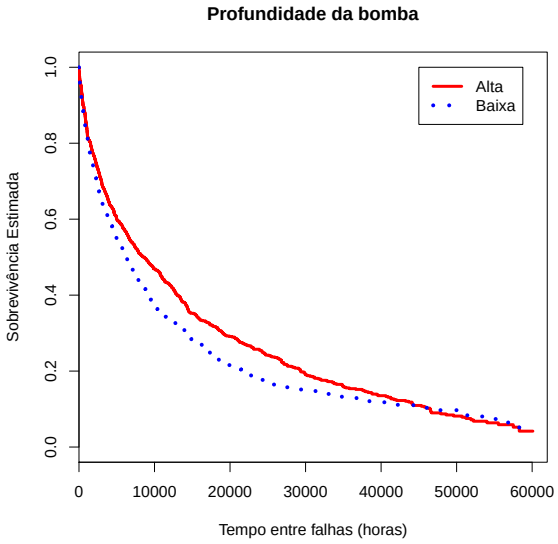

Figura 5.3: Curva de sobrevivência estimada pelo método de Kaplan-Meier por profundidade da bomba: baixa($\leq 507,34\ m$) e alta ($> 507,34\ m$).

verificou-se através da análise de resíduos, que a distribuição Weibull se ajustava bem aos dados. Desta forma, com base nesse resultado e considerando que observações de um mesmo poço-coluna podem estar relacionadas, assumimos um modelo Weibull com efeito aleatório para os dados.

Seguindo a proposta de seleção de covariáveis sugerida por Collett (Collett, 1994), optamos pelo seguinte modelo:

$$\begin{aligned}
\log \quad & T_{ij} = \alpha + U_i + \beta_{prod} PROD_{ij} + \beta_{bm} BM_i + \beta_{id} IDADE_{ij} + \beta_{cam} CAM_i + \beta_{et} ET_i \\
+ \quad & \beta_{rfq} RFQ_i + \beta_{profb} PROFB_i + \beta_{prod*cam} PROD_{ij} * CAM_i + \beta_{prod*et} PROD_{ij} * ET_i \\
+ \quad & \beta_{prod*rfq} PROD_{ij} * RFQ_i + \beta_{cam*profb} CAM_i * PROFB_i + \beta_{et*profb} ET_i * PROFB_i \\
+ \quad & \beta_{rfq*profb} RFQ_i * PROFB_i + \sigma\varepsilon_{ij}, \qquad (5.1)
\end{aligned}$$

em que $i = 1, 2, \ldots, 616$; $j = 1, 2, \ldots, 2374$, com $U_i \sim N(0, \theta)$, representando o efeito aleatório do poço-coluna i e com ε_{ij} representando o erro aleatório do modelo, com distribuição valor extremo padrão.

Na Tabela 5.1 listamos os resultados do ajuste do modelo com efeito aleatório.

Usando o procedimento NLMIXED do SAS adotamos o algoritmo quase-Newton para obtenção dos estimadores de máxima verossimilhança. Os valores iniciais das estimativas foram obtidos sob a hipótese de independência, usando o *software* R. Apresentamos, também, o ajuste do modelo considerando a penalização da verossimilhança.

Tabela 5.1: Estimação de máxima verossimilhança dos parâmetros para dados sobre poços-coluna de petróleo utilizando a distribuição Weibull.

	Com Efeito Aleatório (SAS)			Com Efeito Aleatório (R)		
Parâmetro	Estimativa	EP	P-valor	Estimativa	EP	P-valor
α	7.4689	0,2588	<0,0001	7,1757	0,2554	<0,0001
β_{prod}	-0.0499	0,0091	<0,0001	-0,0451	0,0085	<0,0001
β_{bm}	0,5291	0,1291	<0,0001	0,6898	0,1258	<0,0001
β_{id}	0,0792	0,0078	<0,0001	0,0892	0,0071	<0,0001
β_{cam}	1,3430	0,3089	<0,0001	1,4385	0,3069	<0,0001
β_{et}	0,9575	0,2275	<0,0001	0,8018	0,2285	<0,0001
β_{rfq}	1,8259	0,3305	<0,0001	1,7864	0,3296	<0,0001
β_{profb}	0,0021	0,0004	<0,0001	0,0021	0,0004	<0,0001
$\beta_{prod*cam}$	0,0394	0,0157	0,0125	0,0350	0,0146	0,0162
$\beta_{prod*et}$	0,0175	0,0138	0,2073	0,0157	0,0129	0,2220
$\beta_{prod*rfq}$	-0,0459	0,0187	0,0142	-0,0351	0,0172	0,0412
$\beta_{cam*profb}$	-0,0020	0,0005	0,0002	-0,0022	0,0005	0,0001
$\beta_{et*profb}$	-0,0028	0,0006	<0,0001	-0,0027	0,0006	<0,0001
$\beta_{rfq*profb}$	-0,0027	0,0006	<0,0001	-0,0028	0,0006	<0,0001
σ	1,1954	0,0245	-	1,0700	-	-
θ	0,4526	0,0597	-	0.5500	-	-

5.3 Conclusões

Após tratar da correlação dos dados a estimativa da variância do efeito aleatório é significativa ($\hat{\theta} = 0,4526$ para a proposta implementada no SAS e $\hat{\theta} = 0,5500$ para a proposta via R).

Com respeito aos objetivos do estudo, notamos que os poços-coluna nos quais o método de elevação utilizado foi o Bombeio Mecânico apresentaram maior tempo de funcionamento do que aqueles cujo método de elevação utilizado foi o Bombeio por Cavidade Progressiva ($\hat{\beta}_{bm} = 0,5291$ e $\hat{\beta}_{bm} = 0,6898$, proposta SAS e R, respectivamente, ambas com $p - valor$ menor que $0,0001$) e que o tempo de funcionamento dos poços-coluna localizados nas unidades operacionais ET, CAM e RFQ apresentaram maior tempo de funcionamento do que na unidade ARG.

Verificamos também que à medida que a produção aumenta, o funcionamento do poço-coluna diminui. Contudo, considerando a interação entre a produção do poço-coluna e a unidade que o administra, é possível concluir que os poços-coluna com produção elevada têm menor tempo de funcionamento se estiverem localizados nas unidade CAM, RFQ e ARG. No entanto, não é possível perceber que a produção afeta significativamente o tempo de funcionamento dos poços-coluna localizados na unidade ET.

Para a profundidade da bomba, porém, quanto mais profunda for a instalação da bomba, maior é o tempo livre de falha dos poços-coluna. Além disso constatamos que poços-coluna mais profundos tendem a ter maior tempo de funcionamento se estiverem nas unidades CAM e ARG. Nas unidades ET e RFQ o contrário acontece, de modo que o aumento da profundidade da bomba diminui o tempo de vida dos poços-coluna. O mesmo acontece com a idade, ou seja, poços-coluna mais velhos, em geral, tendem a funcionar por mais tempo de acordo com o aumento da idade. Contudo essa informação deve ser vista com cautela, pois apesar de mostrar a vantagem de funcionamento de poços-coluna após certos ajustes, com relação a poços-coluna muito novos, não deve ser extrapolada para um período de envelhecimento mais longo. Além do mais, acreditamos que poços-coluna muito velhos e pouco produtivos podem eventualmente apresentar longevidade devido ao pouco uso.

Capítulo 6

Considerações Finais

Neste trabalho estudamos dois procedimentos para ajustar um MTFA com efeito aleatório a dados de sobrevivência correlacionados. Aplicamos ambos os métodos a um conjunto de dados relativos a tempos entre falhas de equipamentos de sub-superfície de poços de petróleo e observamos que apesar de não haver descrição detalhada na literatura a implementação usada no *software* livre R apresenta resultados próximos daqueles obtidos pelo *software* SAS.

Em relação à estimativa do parâmetro θ, observamos através da simulação, que o método que utiliza a verossimilhança penalizada pode fornecer estimativas extremamente distorcidas quando temos poucos grupos.

Quanto às pesquisas futuras, pode ser de interesse:

- Realizar estudos de simulação para comparar simultaneamente os procedimentos de estimação abordados aqui (aproximação da verossimilhança marginal via QGHA e verossimilhança penalizada).

- Fazer um estudo detalhado do procedimento que usa a penalização da verossimilhança.

- Implementar computacionalmente esse procedimento de estimação utilizando como aproximação da verossimilhança marginal perfilada a QGHA ao invés da aproximação de Laplace que aparentemete é usada no R.

- Extender o estudo para diferentes suposições para as distribuições do tempo de vida e do efeito aleatório.

- Com respeito à aplicação e com o objetivo de utilizar o modelo para se fazer manutenção preventiva nos poços, é importante a prediçao da função de risco, assim como a validação do modelo através do estudo da sua real capacidade de predição do risco de falha em períodos futuros.

- Ainda com relação à aplicação, desenvolver ou buscar na literatura métodos para analisar resíduos em um MTFA com efeito aleatório para verificar se a distribuição Weibull se ajusta aos dados analisados.

Referências

ABRAMOWITZ, M.; STEGUN, I. *Handbook of mathematical functions.* New York: Dover, 1972.

ATKINSON, K.E. *An introduction to numerical analysis.* New York: Wiley, 1978.

BARROSO, L.C. et al. *Cálculo numérico com aplicações.* 2. ed. São Paulo: Harbra, 1987.

CARVALHO, M. S. et al. *Análise de sobrevida:* teoria e aplicações em saúde. Rio de Janeiro: FIOCRUZ, 2005.

CLAYTON, D.; CUZICK, J. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, v. 148, p. 82-117, 1985.

COLLETT, D. *Modelling survival data in medical research.* London: Chapman Hall, 1994.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada.* São Paulo: Edgard Blücher, 2006.

COX, D. Regression models and life tables. *Journal of the Royal Statistical Society* , 34, p. 187-220, 1972.

DANTAS, M. A. *Modelos de dados de falhas de equipamentos de sub-superfície*

em poços de petróleo da bacia potiguar. 2008. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal do Rio Grande do Norte, Natal, 2008.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.* Series B, v. 39, n. 1, p. 1-38, 1977.

HA, I. D; LEE, Y.; PAWITAN, Y. Genetic mixed linear models for twin survival data. *Behavior Genetics*, v. 37, n.4, p. 621-630, 2007.

HA, I.; LEE, Y.; SONG, J. K. Hierarchical-likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, v. 8, n.2, p. 163-176, 2002.

HOUGAARD, P. *Analysis of multivariate survival data.* New York: Springer, 2000.

HILDEBRAND, F. B. *Introduction to numerical analysis.* 2. ed. New Delhi: McGraw-Hill, 1974.

KAPLAN, E.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, p. 457-481, 1958.

KELLY, P. J. A review of software packages for analyzing correlated survival data. *The American Statistician*, v. 58, n. 4, p. 337-342, 2004.

KLEIN, J. Semiparametric estimation of random eects using the Cox model based on the EM algorithm. *Biometrics*, p. 795-806, 1992.

LAMBERT, P. et al. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*, v. 23, n.20, p. 3177-3192, 2004.

LAWLESS, J. F. *Statistical models and methods for life time data.* 2. ed. New York: Wiley-Interscience & Sons, 2003 .

LEE, Y.; NELDER, J. Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, p. 619-678, 1996.

LIU, Q.; PIERCE, D. A. A note on Gauss-Hermite quadrature. *Biometrika*, v. 81, n. 3, p. 215-224, 1994.

PAN, W. Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, v. 7, n. 1, p. 55-64, 2001.

RIPATTI, S.; PALMGREN, J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, v. 56, p. 1016-1022, 2000.

RITOV, Y. Estimation in a linear regression model with censored data. *The Annals of Statistics*, v. 18, n. 1, p. 303-328, 1990.

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

THERNEAU, T. M.; GRAMBSCH, P. M.; PANKRATZ, V.S. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, v. 12, n. 1, p. 156-175, 2003.

VALENÇA, D. M. *Teste de homogeneidade e estimação para dados de sobrevivência agrupados e com erros de medida*, 2003. Tese (doutorado em estatística) - Instituto de Matemática e Estatística/Universidade de São Paulo, 2003.

ZHANG, J.; PENG, Y. An alternative estimation method for the accelerated failure time frailty model. *Computational Statistics and Data Analysis*, v. 51, n. 9, p. 4413-4423, 2007.

Apêndice A

Programa para Simulações

Neste apêndice apresentamos o programa utilizado nas simulações, em que usamos a versão 2.7.1 do *software* R.

```
#----------------------------------
require(survival)
#----------------------------------
k=100 # número de grupos
n=5  # número de indivíduos no grupo
a=k*n
r=1000  # número de réplicas
teta=0.25  # variância do efeito aleatório
bet=0.5
sigma=1
alfa=5  # média do efeito aleatório
#----------------------------------
u=numeric()
x=numeric()
cod=sort(rep(1:k,n))
tc=numeric() # tempo de censura
tf=numeric() # tempo até a falha
tempo=numeric() # mínimo entre o tempo de censura e o tempo até a falha
logt=numeric()
delta=numeric()
al1=numeric() #alfa estimado pelo método que ignora a dependência
```

```
b1=numeric()  #beta estimado pelo método que ignora a dependência
s1=numeric()  #sigma estimado pelo método que ignora a dependência
al2=numeric() #alfa estimado pelo método que trata a dependência
b2=numeric()  #beta estimado pelo método que trata a dependência
t2=numeric()  #teta estimado pelo método que ignora a dependência
s2=numeric()  #sigma estimado pelo método que ignora a dependência
percens=numeric() #percentual de censura
%#---------------------------------
x=rnorm(a) #default (0,1)
for (j in 1:r){
   u=rnorm(k,alfa,sqrt(teta))# efeito aleatório
 erro=log(rexp(a,1)) #erro aleatório com distribuição valor extremo padrão
 tc=runif(a,0,160)
 tf=exp(u[cod]+bet*x+sigma*erro)
 delta=ifelse(tf<=tc,1,0)

   for (i in 1:a){
 tempo[i]=min(tf[i],tc[i])
 logt[i]=log(tempo[i])
 }

 percens[j]=1-mean(delta)
 percens[j]

# Ajuste do modelo ignorando a dependência
 ajuste1<-survreg(Surv(tempo, delta) ~ x, dist="weibull")
 summary(ajuste1)

# Ajuste do modelo tratando a dependência
 ajuste2<-survreg(Surv(tempo, delta)~x+ frailty.gaussian(cod, method=‘‘aic’’))
 summary(ajuste2)

{
 if (!is.null(ajuste2$history[[1]]$history))
 theta <- ajuste2$history[[1]]$history[nrow(ajuste2$history[[1]]$history), 1]
 else theta<-ajuste2$history[[1]]$theta
```

```
}
theta

#Sob independência
al1[j]=coef(ajuste1)[1]
b1[j]=coef(ajuste1)[2]
  s1[j]=ajuste1$scale

#Tratanto a dependência
al2[j]=coef(ajuste2)[1]
t2[j]=theta
b2[j]=coef(ajuste2)[2]
s2[j]=ajuste2$scale
}

mean(percens)


# Sob independência

#Média
(alhat1=mean(al1))
(bhat1=mean(b1))
(shat1=mean(s1))

#Erro padrão
(epal1=sqrt((r-1)*var(al1)/r))
(epb1=sqrt((r-1)*var(b1)/r))
(eps1=sqrt((r-1)*var(s1)/r))

#REQM
(eqmal1=sqrt(mean((al1-alfa)^2)))
(eqmb1=sqrt(mean((b1-bet)^2)))
(eqms1=sqrt(mean((s1-sigma)^2)))
```

```
# Tratanto a dependência
#Média
(that2=mean(t2))
(alhat2=mean(al2))
(bhat2=mean(b2))
(shat2=mean(s2))

#Erro padrão
(ept2=sqrt((r-1)*var(t2)/r))
(epal2=sqrt((r-1)*var(al2)/r))
(epb2=sqrt((r-1)*var(b2)/r))
(eps2=sqrt((r-1)*var(s2)/r))

#REQM
(eqmt2=sqrt(mean((t2-teta)^2)))
(eqmal2=sqrt(mean((al2-alfa)^2)))
(eqmb2=sqrt(mean((b2-bet)^2)))
(eqms2=sqrt(mean((s2-sigma)^2)))
```

Apêndice B

Programas para Aplicação

Neste apêndice serão apresentados os programas utilizados para obtenção das estimativas dos parâmetros.

B.1 Ajuste de Dados Correlacionados com o NLMIXED do SAS

Dados sobre Poços de Petróleo - (não ordenados)

```
data petro;
input  cod   tempo delta prod met   uni  profb  idade;
y=log(tempo);
datalines;
     10005 U  669   1  6,59 BCP  OP-ARG  510   14,15
     10005 U  436   1  8,50 BCP  OP-ARG  510   14,35

    ....

      9966 U  14950 0  90,3  BM  OP-ET   199   20,77
run;
proc sort data = petro;
    by  cod;
    run;
proc nlmixed  cov
    data=petro;
parms b0=7.46882  b1=-0.05021 b2=0.51968 b3=0.05872 b4=1.34207
 b5=0.96125 b6=1.82449 b7=0.00216 b8=0.03204 b9=0.01858 b10=-0.07772
```

```
 b11=-0.00209  b12=-0.00268 b13=-0.00271 sigma=1.34 teta=0.1;
    bounds sigma, teta>= 0;
    if(met="BM")then bm=1;
    if(met="BCP")then  bm=0;
    if(unid="OP-CAM")then cam=1;
    if(unid="OP-CAM")then et=0;
    if(unid="OP-CAM")then rfq=0;
    if(unid="OP-RFQ")then cam=0;
    if(unid="OP-RFQ")then et=0;
    if(unid="OP-RFQ")then rfq=1;
    if(unid="OP-ARG")then cam=0;
    if(unid="OP-ARG")then et=0;
    if(unid="OP-ARG")then rfq=0;
    if(unid="OP-ET")then cam=0;
    if(unid="OP-ET")then et=1;
    if(unid="OP-ET")then rfq=0;
    eta = b1*prod + b3*bm + b4*idade + b5*cam + b6*et + b7*rfq
     + b8*profb + b9*prod*cam + b10*prod*et + b11*prod*rfq
     + b13*cam*profb + b14*et*profb + b15*rfq*profb + u;
    s = (y-eta)/sigma;
    logp = delta*(s-log(sigma))- exp(s);
    model y ~ general(logp);
    random u ~ normal(0,teta) subject=cod;
    run;
```

B.2 Ajuste de Dados Correlacionados com o R

```
require(survival)

ajuste <-survreg(formula=Surv(tempo, delta) ~ prod + met + idade
+ unid + profb + prod*unid + met*profb + unid*profb
+ frailty.gaussian(cod, sparse=TRUE), petro)
summary(ajuste)
```

Para a obtenção do valor de teta, usamos o comando:

```
{
  if (!is.null(ajuste$history[[1]]$history))
  theta <- ajuste$history[[1]]$history[nrow(ajuste$history[[1]]$history), 1]
  else theta <- ajuste$history[[1]]$theta
}
theta
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo