Universidade Federal do Rio Grande do Norte
Centro de Ciências Exatas e da Terra
Departamento de Estatística
Programa de Pós-graduação em Matemática Aplicada e Estatística

Predição em Modelos de Tempo de Falha Acelerado com Efeito Aleatório para Avaliação de Riscos de Falha em Poços Petrolíferos

João Batista Carvalho

Orientadora: Dione Maria Valença

Co-orientador: Julio da Motta Singer

Natal, maio de 2010

João Batista Carvalho

Predição em Modelos de Tempo de Falha Acelerado com Efeito Aleatório para Avaliação de Riscos de Falha em Poços Petrolíferos

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

Orientadora: Profª Drª Dione Maria Valença

Natal
Maio/2010

Dedico esta dissertação aos meus pais, Severino e Maria; à minha família; ao meu tio Dedé (in memoriam) e sobrinho Jandi (in memoriam).

Agradecimentos

Por dever de justiça não poderia deixar de fazer meus sinceros agradecimentos a todos aqueles que contribuíram para realização desta dissertação.

Em particular, sou grato à professora Dione pela orientação, paciência e amizade.

Aos professores Julio Singer, Juvêncio Nobre e Idemauro Lara pelas magníficas sugestões.

Aos professores, sejam do PPGMAE ou não, pelos ensinamentos, dentre eles Arimatéia, Damião Nóbrega, Paulo Lúcio, André Gustavo, Pledson Guedes e André Pinho.

Aos "pousistas" da residência de Pós-Graduação da UFRN pela convivência durante este trabalho, compartilhando sonhos e frustrações.

Aos colegas de mestrado pelo companheirismo.

Aos meus amigos e familiares pela confiança.

À PETROBRAS pelo apoio financeiro no primeiro semestre do curso e por disponibilizar os dados analisados neste trabalho.

À CAPES pelo apoio financeiro.

Enfim, aos meus pais por me possibilitarem o dom da vida e pelo apoio em minhas decisões.

Quando a incerteza da realização de um bom trabalho surge, a autoestima, a autoconfiança e o estímulo das pessoas amadas são as fontes que nos conduzem à persistência, fator indispensável às grandes conquistas.

Resumo

Consideramos técnicas de predição baseadas em modelos de tempo de falha acelerado com efeito aleatório para dados de sobrevivência correlacionados. Além do enfoque bayesiano através do estimador de Bayes empírico (EBE), também discutimos sobre o uso de um preditor clássico, o melhor preditor linear não viciado empírico (*empirical best linear unbiased predictor* - EBLUP). Para ilustrar a utilização desses preditores, fazemos aplicações a um conjunto de dados reais envolvendo tempos entre falhas de equipamentos de poços de petróleo da Bacia Potiguar. Nesse contexto, o objetivo é predizer os riscos/probabilidades de falha com a finalidade de subsidiar programas de manutenção preventiva. Os resultados obtidos mostram que ambos os métodos são adequados para prever falhas futuras, proporcionando boas decisões em relação ao emprego e economia de recursos para manutenção preventiva.

Palavras-chave: Predição de efeitos aleatórios, estimador de Bayes empírico, EBLUP, modelos para dados de sobrevivência correlacionados, falhas em poços petrolíferos.

Abstract

We considered prediction techniques based on models of accelerated failure time with random effects for correlated survival data. Besides the bayesian approach through empirical Bayes estimator, we also discussed about the use of a classical predictor, the Empirical Best Linear Unbiased Predictor (EBLUP). In order to illustrate the use of these predictors, we considered applications on a real data set coming from the oil industry. More specifically, the data set involves the mean time between failure of petroleum-well equipments of the Bacia Potiguar. The goal of this study is to predict the risk/probability of failure in order to help a preventive maintenance program. The results show that both methods are suitable to predict future failures, providing good decisions in relation to employment and economy of resources for preventive maintenance.

Keywords: Prediction of random effects, empirical Bayes estimator, EBLUP, models for correlated survival data, failures in petroleum-well.

Sumário

1 Introdução **1**

1.1 Motivação: Falhas em Poços de Petróleo 1

1.2 Referencial Teórico 4

1.3 Objetivos 5

2 Conceitos Básicos de Análise de Sobrevivência **6**

2.1 Distribuição do Tempo até a Falha 6

2.2 Censura 8

2.3 Distribuição Weibull 9

2.4 Dados de Sobrevivência Independentes 10

2.4.1 Modelos de Tempo de Falha Acelerado 11

2.5 Dados de Sobrevivência Correlacionados 12

3 Predição de Efeitos Aleatórios **14**

3.1 Modelo Linear Misto 15

3.2 Predição Clássica - BLUP e EBLUP 17

3.2.1 Exemplo: ANOVA com um único fator aleatório 19

3.2.2 BLUE e BLUP Empíricos 20

3.3 Predição Bayesiana - Bayes Empírico 23

3.3.1 Normalidade: relação entre o EBLUP e o EBE 24

4 Predição em Modelos de Tempo de Falha Acelerado com Efeito Aleatório 26

4.1 Predição Via Estimador de Bayes Empírico 27

4.1.1 Função de Verossimilhança Marginal 27

4.1.2 Predição dos Efeitos Aleatórios 29

4.2 Predição Via EBLUP 29

4.3 Predição de Riscos e Probabilidades de Falha 31

5 Aplicação e Discussão 33

5.1 Ajustes do Modelo 33

5.2 Riscos de Falha Preditos 36

5.3 Efeitos Aleatórios Preditos 39

5.4 Validação do Modelo 39

6 Considerações Finais 46

6.1 Conclusão 46

6.2 Trabalhos Futuros 47

Referências Bibliográficas 48

A Outros Métodos de Predição 51

B Tabelas 54

C Programas 57

Lista de Tabelas

5.1 Resultados dos ajustes 36
5.2 Percentual de acertos referente aos poços indicados para manutenção preventiva 44
5.3 Percentual de acertos referente aos poços-coluna não indicados para manutenção preventiva 45
5.4 Percentual de decisões corretas considerando MP e MNP 45

B.1 Número de poços indicados para manutenção preventiva e que de fato falharam 54
B.2 Número de poços indicados para manutenção preventiva 55
B.3 Número de poços não indicados para manutenção preventiva e que de fato não falharam 55
B.4 Número de poços não indicados para manutenção preventiva 55
B.5 Número de decisões corretas 56

Lista de Figuras

5.1 *Box plots* dos logaritmos dos riscos considerando o EBE e o EBLUP, respectivamente. 37
5.2 *Box plots* dos logaritmos dos riscos por ano - EBE. 38
5.3 *Box plots* dos logaritmos dos riscos por ano - EBLUP. 38
5.4 Histograma dos efeitos aleatórios (EA) preditos pelo EBE 40
5.5 Histograma dos efeitos aleatórios (EA) preditos pelo EBLUP 40
5.6 *Box plots* das probabilidades condicionais de falha - EBE. 42
5.7 *Box plots* das probabilidades condicionais de falha - EBLUP. 43

Capítulo 1

Introdução

Neste trabalho abordamos modelos estatísticos para tratar dados de sobrevivência correlacionados. Com base nesses modelos, realizamos um estudo com dados de tempo entre falhas de equipamentos de poços petrolíferos, cedidos pela PETROBRAS.

Neste capítulo, introduzimos o problema das falhas em poços petrolíferos e fornecemos informações detalhadas sobre o conjunto de dados de interesse. Em seguida, apresentamos referenciais teóricos e os objetivos do trabalho.

1.1 Motivação: Falhas em Poços de Petróleo

Em poços de petróleo, a elevação dos fluidos (óleos/gás) pode ser feita por meio de diversas técnicas. Na Unidade de Negócio de Exploração e Produção do Rio Grande do Norte e Ceará (Bacia Potiguar), explorada pela PETROBRAS, as técnicas mais utilizadas são o *bombeio mecânico* (BM), também conhecido como "cavalo de pau" e o *bombeio por cavidades progressivas* (BCP). Essas técnicas exigem a instalação de equipamentos de sub-superfície que podem apresentar defeitos ao longo do tempo devido à entrada de areia, corrosão, pressão interna etc, causando interrupções (falhas) no funcionamento do poço.

Por isso, há necessidade de serviços de manutenção corretiva, que são medidas tomadas no sentido de corrigir/substituir os equipamentos que provocaram a falha. Contudo, além da inconveniência de suspender a produção, tais serviços são realizados por sondas de produção terrestre móveis terceirizadas e com custos muito elevados.

Para minimizar essas falhas há interesse em programas de manutenção preventiva. Em geral, esses serviços são economicamente mais viáveis em relação aos de manutenção corretiva. Por exemplo, é mais barato e rápido fazer a limpeza de um equipamento de subsuperfície, evitando seu desgaste, do que substituí-lo por um novo futuramente, caso a limpeza não seja realizada. Contudo, devido à grande quantidade de poços da Bacia Potiguar (aproximadamente 5.000) é necessário identificar em quais poços devem-se executar os serviços de prevenção. Para tratar essa questão pode-se considerar a modelagem estatística dos tempos de funcionamento dos poços até a ocorrência das falhas. Com isso, é possível identificar os poços com maiores riscos de falha em um determinado instante. Convenientemente, esses poços seriam aqueles selecionados para a execução dos serviços de prevenção.

Este trabalho tem como motivação prática um estudo retrospectivo sobre tempos entre falhas de equipamentos de subsuperfície de uma amostra de poços de petróleo terrestres da Bacia Potiguar, no período de janeiro de 2000 a dezembro de 2006. Detalhes sobre o processo de amostragem podem ser vistos em Dantas (2008). Além dos tempos entre falhas, foram registradas diversas informações concomitantes para cada poço, tais como produção, profundidade, idade, localização etc.

Nesse contexto, quando um poço volta a funcionar normalmente após uma manutenção, seu tempo de funcionamento até a próxima falha é registrado. Portanto, diversos tempos entre falhas consecutivas para cada poço são registrados. Por se tratar de dados em que são feitas diversas observações da variável resposta (tempo entre falhas) para cada unidade

amostral (poço), esperamos que os tempos entre falhas de um mesmo poço estejam correlacionados (correlação intra-unidades amostrais).

Devido à presença de censuras nos dados (ver Seção 2.2), as técnicas estatísticas usuais de regressão ou de modelos lineares generalizados não são apropriadas para modelar a associação existente entre os tempos de falha e as informações concomitantes (covariáveis). Para contornar o problema, é necessário o uso de técnicas próprias de *Análise de Sobrevivência*.

Dantas (2008) ajustou um modelo de tempo de falha acelerado a um subconjunto desses dados, considerando apenas o tempo até a primeira falha. Mais recentemente, Santos (2009) ajustou um modelo de tempo de falha acelerado estendido aos dados completos, que corresponde aos tempos entre falhas em todo o período (2000 a 2006). Nesses trabalhos a ênfase foi a identificação das covariáveis (fatores) que influenciavam o tempo até a falha dos poços.

Para atingir o principal interesse da PETROBRAS, ou seja, a identificação de poços que necessitam de manutenção preventiva, é necessário predizer os riscos de falha dos poços além do ajuste do modelo estendido. A especificação de métodos de predição de funções apropriadas relacionadas ao risco ou à probabilidade de falha nos poços e a utilização dessas funções para orientar a tomada de decisões com respeito à realização ou não da manutenção preventiva são os objetivos deste trabalho.

Um importante aspecto a ser apontado é o fato de o modelo estendido adotado em Santos (2009) considerar um efeito aleatório para tratar da possível estrutura de correlação existente entre os tempos observados de cada poço. Todavia, a predição de funções de um efeito aleatório em Análise de Sobrevivência corresponde à uma abordagem teórica ainda emergente na literatura.

1.2 Referencial Teórico

Uma maneira de modelar a estrutura de correlação intra-unidades amostrais é assumir que o efeito conjunto de covariáveis importantes que não foram observadas (covariáveis omissas) pode ser descrito por meio da inclusão de um ou mais efeitos aleatórios no modelo. Foi esta ideia que originou os principais modelos que tratam do assunto.

De modo geral, o termo efeito aleatório designa uma variável aleatória não observável. Robinson (1991) faz uma ampla discussão sobre efeitos aleatórios e apresenta uma série de exemplos de sua utilização: estimação de méritos genéticos em criação de animais, teoria da credibilidade para elaboração de prêmios de seguro etc.

Vaupel et al. (1979) estendem o modelo de riscos proporcionais (Cox, 1972) incluindo um efeito aleatório. O modelo estendido ficou conhecido como *modelo de fragilidade* porque nos estudos com pacientes acometidos por uma certa doença, grandes valores do efeito aleatório estão associados a altos riscos de morte, caracterizando indivíduos mais "frágeis". Posteriormente, esse modelo foi generalizado por diversos autores, dentre os quais mencionamos Hougaard (1986), Oakes (1989) e Klein (1992). Em particular, Anderson e Louis (1995) usam essa classe de modelos na análise de dados de sobrevivência bivariados considerando distribuições paramétricas e não paramétricas para a fragilidade (efeito aleatório); Klein et al. (1999), consideram um modelo de sobrevivência baseado na distribuição lognormal e Pan (2001) trata de modelos semi-paramétricos de tempo de falha acelerado com distribuição gama para a fragilidade.

Com base no modelo de fragilidade, Valença (2003) e Lambert et al. (2004) propõem uma nova abordagem para o problema, estendendo o modelo de tempo de falha acelerado por meio da adição de um efeito aleatório associado a cada unidade amostral. Lambert et al. (2004) usam o modelo para avaliar o efeito de covariáveis nos tempos de sobrevivência após o transplante de rins de pacientes de 31 clínicas do Reino Unido e propõem a predição

dos efeitos aleatórios baseada no enfoque bayesiano empírico (Carlin e Louis, 1998). Ao invés do termo *fragilidade*, Lambert et al. (2004) usam o termo *vigor* porque nesse estudo, grandes (pequenos) valores do efeito aleatório implicam grandes (pequenos) tempos de sobrevivência, caracterizando indivíduos mais "resistentes". Eles também consideram diferentes combinações da distribuição dos efeitos aleatórios e da função de risco base na análise dos dados.

1.3 Objetivos

Este trabalho tem como objetivo principal realizar um estudo sobre predição em modelos de tempo de falha acelerado com efeito aleatório. Especificamente desejamos:

(i) Descrever e estudar o uso de preditores bayesiano, estimador de Bayes empírico (EBE), e clássico, melhor preditor linear não viciado empírico (*empirical best linear unbiased predictor* - EBLUP), para predição de efeitos aleatórios nesses modelos;

(ii) ajustar os modelos aos dados sobre tempos entre falhas em poços petrolíferos com base nos preditores EBE e EBLUP para obter probabilidades/riscos de falha;

(iii) avaliar através da técnica de validação externa a eficácia dos modelos ajustados na indicação de ações de manutenção preventiva.

No Capítulo 2, apresentamos alguns conceitos básicos de análise sobrevivência. Uma definição de efeitos aleatórios, bem como uma apresentação dos principais métodos de predição em modelos com efeitos aleatórios são tratadas no Capítulo 3. No Capítulo 4, discutimos os modelos de tempo de falha acelerado com efeito aleatório. No Capítulo 5, essa classe de modelos é aplicada aos dados de poços petrolíferos da Bacia Potiguar. Por fim, conclusões e considerações sobre trabalhos futuros são apresentadas no Capítulo 6.

Capítulo 2

Conceitos Básicos de Análise de Sobrevivência

A Análise de Sobrevivência se constitui de um conjunto de técnicas estatísticas que tem como finalidade modelar o tempo até a ocorrência de um determinado evento de interesse. Os primeiros estudos nessa área foram baseados em estudos de mortalidade, o que justifica, portanto, o termo *sobrevivência* em sua denominação. A variável aleatória em estudo é não negativa e pode representar o tempo até a falha de um determinado equipamento eletrônico, tempo de vida de pacientes após uma cirurgia etc. Na literatura, são conhecidos como tempo de vida, tempo de sobrevivência ou ainda tempo até a falha. A seguir, apresentamos alguns conceitos úteis para nosso objetivo. Mais detalhes podem ser obtidos em Lawless (2003) ou Kalbfleisch e Prentice (2002), por exemplo.

2.1 Distribuição do Tempo até a Falha

Seja T uma variável aleatória contínua não negativa, com função densidade de probabilidade $f(t)$, representando o tempo até a falha de uma determinada unidade amostral

proveniente de uma população em estudo. A distribuição de T é caracterizada pelas funções de sobrevivência e de risco, definidas a seguir.

A função de sobrevivência é definida como

$$S(t) = P(T > t) = \int_t^{\infty} f(t)dt = 1 - F(t), \quad t > 0, \tag{2.1}$$

sendo $F(t)$ a função de distribuição acumulada de T.

A função de sobrevivência é interpretada como a probabilidade de uma observação não falhar (sobreviver) durante um certo tempo t e tem as seguintes propriedades:

(i) monótona decrescente;

(ii) contínua à esquerda;

(iii) S(0)=1;

(iv) $\lim_{t\to\infty} S(t) = 0$.

A propriedade (iv) indica que a probabilidade de um indivíduo sobreviver por um período muito grande é zero.

Outra função de particular relevância na análise de dados de sobrevivência é a função de risco ou função de taxa de falha. Ela representa a taxa de falha instantânea de uma observação no tempo t dado que não falhou até esse instante. A função de risco é definida como

$$h(t) = \lim_{\Delta t\to 0^+} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{2.2}$$

Os riscos de falha são números positivos, mas sem limite superior e que, dependendo da característica da falha, as taxas de falha podem sofrer aumento, diminuição, oscilação ou mesmo permanecerem constantes ao longo do tempo. Em alguns casos, a função de risco é mais informativa do que a própria função de sobrevivência, pois diferentes funções de so-

brevivência com formas semelhantes podem corresponder à funções de risco drasticamente diferentes. A relação entre essas funções é dada por

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln S(t)), \tag{2.3}$$

dado que $f(t) = -(d/dt)S(t)$.

2.2 Censura

Em estudos de sobrevivência algumas unidades amostrais podem deixar de ser acompanhadas por motivos de natureza aleatória ou por não terem experimentado o evento de interesse até o final do estudo. Isso gera observações incompletas ou parciais dos tempos até a falha, denominadas *censuras*. É necessário incorporá-las na análise estatística por duas razões principais. Primeiramente, porque indicam que o tempo até a ocorrência da falha é superior ao tempo de acompanhamento, ou seja, elas trazem informação sobre a "durabilidade" do item/indivíduo. Por outro lado, negligenciá-las no cálculo das estatísticas pode acarretar em estimativas viciadas (Colosimo e Giolo, 2006). Quando a distribuição da censura não depende de parâmetros de interesse, ela é denominada de censura não informativa.

Consideremos o exemplo da Introdução em que uma amostra aleatória de poços foi observada entre 2000 e 2006. A ocorrência de censuras se deve ao fato de alguns desses poços terem sido inativados durante o período de estudo e por isso deixaram de ser acompanhados e outros após uma manutenção preventiva, não terem voltado a falhar até o final do estudo.

Na análise de dados sobrevivência, em geral, não se pode assumir distribuição normal para os tempos até a falha, pois estes geralmente apresentam uma distribuição assimétrica positiva. Dentre as distribuições mais comuns consideradas para o tempo até a falha está a

distribuição Weibull. Tendo em vista sua utilidade, a seguir fazemos uma breve descrição de suas principais características.

2.3 Distribuição Weibull

Se T possui distribuição Weibull com parâmetros $\alpha > 0$ e $\gamma > 0$, então as respectivas expressões de sua função densidade, função de sobrevivência e função de risco são

$$f(t) = \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right] I_{(0,\infty)}(t), \tag{2.4}$$

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^{\gamma}\right] I_{(0,\infty)}(t) \tag{2.5}$$

e

$$h(t) = \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1} I_{(0,\infty)}(t). \tag{2.6}$$

Notemos que quando $\gamma = 1$, a distribuição Weibull se reduz a distribuição exponencial com parâmetro α. O r-ésimo momento centrado em zero da Weibull é $\alpha^r\Gamma(1+r/\gamma)$, sendo Γ a função gama. Características importantes da função de risco acima são: monotonicidade decrescente para $\gamma < 1$, crescente para $\gamma > 1$ e constante para $\gamma = 1$.

Alguns autores, entre eles Hougaard (1986) e Morris e Christiansen (1995), consideram uma outra parametrização para a distribuição Weibull, em que $\eta = \alpha^{-\gamma}$. Consequentemente, as funções de sobrevivência e de risco reparametrizadas adquirem as seguintes formas:

$$S(t) = \exp\left[-\eta t^{\gamma}\right] \quad \text{e} \quad h(t) = \eta\gamma t^{\gamma-1}. \tag{2.7}$$

Outra distribuição importante para análise de dados de sobrevivência que pode ser obtida a partir da distribuição Weibull, por meio da relação $Y = \ln T$, é a distribuição valor

extremo, também conhecida como Gumbel, cujas funções densidade e de sobrevivência são, respectivamente,

$$f(y) = \frac{1}{\sigma} exp\left[\frac{y-\mu}{\sigma} - exp\left(\frac{y-\mu}{\sigma}\right)\right] I_{\mathbb{R}}(y) \tag{2.8}$$

e

$$S(y) = exp\left[-exp\left(\frac{y-\mu}{\sigma}\right)\right] I_{\mathbb{R}}(y), \tag{2.9}$$

As seguintes relações entre os parâmetros das referidas distribuições são $\mu = \ln \alpha$ e $\sigma = \gamma^{-1}$. A média e a variância da distribuição valor extremo são, respectivamente, $\mu - c\sigma$ e $\sigma^2\pi^2/6$, com $c = 0,5772$ (constante de Euler). Também é válida a relação $Y = \mu + \sigma\epsilon$, em que ϵ é uma variável aleatória com distribuição valor extremo padrão cujas média e variância são conhecidas e cuja função densidade é

$$f(z) = exp[z - exp(z)] I_{\mathbb{R}}(z), \tag{2.10}$$

2.4 Dados de Sobrevivência Independentes

Para modelar o mecanismo de censura aleatória consideremos T e C variáveis aleatórias independentes representando os tempos até a falha e até a censura, respectivamente. O que se observa para uma unidade amostral em estudo é, portanto, o mínimo entre T e C, que indicamos por $T^* = \min(T, C)$. Definamos também $\delta = I(T \leq C)$, em que I é a função indicadora.

A resposta para a i-ésima unidade amostral ($i = 1, ..., n$), sendo n o número de indivíduos em uma amostra aleatória, é representada pelo par de variáveis aleatórias (T_i^*, δ_i), em que T_i^* é o tempo até a falha ou até a censura e δ_i, a variável indicadora do evento. Notemos que a resposta é uma variável aleatória mista, pois possui uma componente contínua, T_i^*, e uma componente discreta, δ_i.

Em estudos que consideram covariáveis que se relacionam com a resposta, os dados de sobrevivência são representados por $(T_i^*, \delta_i, \mathbf{x}_i)$, em que $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^\top$ representa um vetor $p \times 1$ de covariáveis para a i-ésima unidade amostral. O efeito que $\mathbf{x}_i$ exerce sobre (T_i^*, δ_i) pode ser expresso por meio de um modelo de regressão. Uma classe importante de modelos de regressão usados em análise de sobrevivência quando há independência nos dados são os modelos de tempo de falha acelerado. A seguir, fazemos uma breve abordagem sobre essa classe.

2.4.1 Modelos de Tempo de Falha Acelerado

Esta classe de modelos, que pertence à classe de modelos de localização e escala, caracteriza-se pelo fato de $Y = \ln T$ ter uma distribuição com parâmetro de localização $\mu(\mathbf{x})$, dependendo do vetor de covariáveis $\mathbf{x}$, e um parâmetro de escala $\sigma > 0$. Em geral, consideramos $\mu(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$, com $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$ e em que a primeira componente do vetor $\mathbf{x}$ é sempre igual a 1. Uma maneira de representar um modelo de tempo de falha acelerado é

$$Y = \boldsymbol{\beta}^\top \mathbf{x} + \sigma\epsilon, \tag{2.11}$$

em que ϵ é um erro aleatório cuja distribuição não depende de $\mathbf{x}$ e determina a forma da distribuição de Y. Isto corresponde a um modelo de regressão linear para o logaritmo dos tempos, Y. Se considerarmos o modelo na escala original dos tempos, isto é, $T = \exp(\boldsymbol{\beta}^\top \mathbf{x}) \exp(\sigma\epsilon)$, notamos que as covariáveis atuam multiplicativamente sobre T. A interpretação desse fato é que as covariáveis agem no sentido de acelerar ou desacelerar o tempo de falha, o que justifica o termo *acelerado* na denominação do modelo.

Um importante representante dessa classe de modelos é o modelo de regressão Weibull. Este é obtido quando supomos uma distribuição valor extremo padrão para ϵ, cuja densidade é dada em (2.10). Portanto, $Y = \ln T$ tem distribuição valor extremo e, consequente-

mente, T tem distribuição Weibull. As relações entre os parâmetros dessas distribuições são $\alpha(\mathbf{x}) = \exp(\mu(\mathbf{x})) = \exp(\boldsymbol{\beta}^\top \mathbf{x})$ e $\sigma = \gamma^{-1}$. Com a inclusão de covariáveis, as respectivas funções de sobrevivência e de risco para T são

$$S(t) = \exp\left[-\exp(-\gamma \boldsymbol{\beta}^\top \mathbf{x}) t^\gamma\right] \tag{2.12}$$

e

$$h(t) = \gamma \exp(-\gamma \boldsymbol{\beta}^\top \mathbf{x}) t^{\gamma-1}. \tag{2.13}$$

2.5 Dados de Sobrevivência Correlacionados

Quando consideramos tempos de sobrevivência de gêmeos, de pacientes submetidos ao mesmo processo cirúrgico ou aqueles que foram observados repetidas vezes para cada unidade amostral, esperamos que as respectivas observações sejam correlacionadas. Dados dessa natureza são denominados *dados de sobrevivência correlacionados* (Hougaard, 2000). Nesse contexto, modelos de Análise de Sobrevivência tradicionais, como os *modelos de riscos proporcionais de Cox* (Cox, 1972) e os *modelos de tempo de falha acelerado*, descritos na seção anterior, não podem ser diretamente aplicados, pois supõem independência nos dados. Consequentemente, são necessários outros tipos de modelos, como aqueles que envolvem efeitos aleatórios para tratar a correlação dos dados (Valença, 2003; Lambert et al., 2004).

A ideia básica que norteia esses modelos é que existem fatores de risco não observados que são comuns entre os indivíduos de um mesmo grupo e constantes ao longo do tempo para cada indivíduo. Portanto, assume-se que esses fatores são responsáveis pela dependência existente entre os tempos de vida. Outro ponto chave é a suposição de independência condicional entre os tempos de vida dado o conhecimento desses fatores.

Um subconjunto dessa classe de modelos, que será estudado com mais detalhes no Capítulo 4, é a classe dos modelos de tempo de falha acelerado com efeito aleatório. Nestes modelos, os fatores de risco comuns não observados são vistos como covariáveis omissas. Assumimos que essa omissão de covariáveis é a causa da dependência entre as observações de uma mesma unidade amostral. Portanto, os modelos incorporam um efeito aleatório para representar as covariáveis omissas e, consequentemente, tratar a dependência. Kleiding et al. (1997) discutem sobre o papel dos modelos de tempo de falha acelerado em descrever a dependência devida às covariáveis omissas. No capítulo seguinte, apresentamos alguns métodos de predição de efeitos aleatórios em modelos para dados não censurados.

Capítulo 3

Predição de Efeitos Aleatórios

Os efeitos de fatores na variável resposta de interesse podem ser fixos ou aleatórios. Os efeitos são fixos quando todos os níveis do fator no qual estamos interessados são avaliados. Por exemplo, a Bacia Potiguar é dividida em quatro unidades administrativas de acordo com a localização geográfica dos poços de petróleo. São elas: Canto do Amaro, Alto do Rodrigues, Campo do Estreito e Fazenda Riacho da Forquilha, que constituem os níveis do fator Unidade Administrativa. Portanto, se considerarmos um experimento em que se deseja verificar o efeito da localização do poço no tempo de vida útil dos poços de petróleo pertencentes à Bacia Potiguar, teremos um caso de efeito fixo. De acordo com Searle et al. (1992), os efeitos são aleatórios quando apenas uma amostra aleatória dos níveis de um fator é avaliada.

Modelos em que todos os fatores têm efeitos fixos são denominados modelos de efeitos fixos ou simplesmente modelos fixos. Modelos em que todos os fatores apresentam efeitos aleatórios são denominados modelos de efeitos aleatórios ou modelos aleatórios. Modelos que apresentam alguns fatores com efeitos fixos e outros com efeitos aleatórios são denominados modelos mistos. Em modelos mistos, efeitos fixos correspondem aos parâmetros e efeitos aleatórios às variáveis aleatórias não observáveis. Nos ajustes desses modelos,

dizemos que estimamos a parte fixa e predizemos a aleatória. Os modelos mistos são úteis para modelar a estrutura de covariância intra-unidades amostrais. A seguir, tratamos do processo de predição em modelos lineares mistos.

3.1 Modelo Linear Misto

Consideramos o modelo linear misto com um único fator aleatório definido a seguir. O caso geral, que engloba vários fatores, pode ser visto em Demidenko (2004) e Robinson (1991), por exemplos. Desse modo, o referido modelo é expresso por

$$Y_{ij} = b_i + \mathbf{X}_{ij}^{\top}\boldsymbol{\beta} + e_{ij}, \tag{3.1}$$

em que Y_{ij}, $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$, é a j-ésima observação da i-ésima unidade amostral, k é o total de unidades amostrais, n_i é o total de observações associadas à i-ésima unidade amostral, b_i's são efeitos aleatórios não correlacionados, com médias zero e variâncias iguais (σ_b^2), $\mathbf{X}_{ij}$'s são vetores $p \times 1$ de covariáveis, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^{\top}$ é um vetor $p \times 1$ de parâmetros (efeitos fixos) e e_{ij}'s são erros aleatórios independentes, com médias zero, variâncias iguais (σ_e^2) e não correlacionados com os b_i's.

A presença de b_i na equação do modelo serve para modelar a variância de Y_{ij}, que é dada por

$$\mathrm{Var}(Y_{ij}) = \sigma_b^2 + \sigma_e^2, \tag{3.2}$$

sendo σ_b^2 e σ_e^2 conhecidas como componentes de variância.

Um dos principais pressupostos do modelo é a suposição de independência condicional entre as observações da i-ésima unidade amostral dado o efeito aleatório b_i. Em resumo, esse modelo é um modelo de independência condicional homocedástico com efeitos aleatórios não correlacionados.

Escrevendo o modelo em forma matricial temos

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \tag{3.3}$$

em que $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \ldots, \mathbf{Y}_k^\top)^\top$, com $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^\top$, $i = 1, 2, \ldots, k$, $n = \sum_{i=1}^{k} n_i$, e $\mathbf{b} = (b_1, b_2, \ldots, b_k)^\top$ são vetores de variáveis aleatórias, tais que $\mathbf{Y}$ é um vetor $n \times 1$ observável de respostas e $\mathbf{b}$, um vetor $k \times 1$ não observável de efeitos aleatórios, com $E(\mathbf{b}) = \mathbf{0}$ e $\mathrm{Var}(\mathbf{b}) = \sigma_b^2 \mathbf{I}_k$, em que $\mathbf{I}_k$ é a matriz identidade de ordem k. Ademais, $\mathbf{e}$ é um vetor $n \times 1$ de erros aleatórios, com $\mathrm{E}(\mathbf{e}) = \mathbf{0}$, $\mathrm{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}_n$ e não correlacionado com $\mathbf{b}$, $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_k^\top)^\top$, $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})^\top$ é uma matriz $n \times p$ de covariáveis e $\mathbf{Z}$, uma matriz $n \times k$ de especificação, ambas conhecidas e de posto completo, tal que

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1}_{n_k} \end{bmatrix} = \oplus_{i=1}^{k} \mathbf{1}_{n_i}, \tag{3.4}$$

em que $\mathbf{1}_{n_i}$, $i = 1, \ldots, k$, denota um vetor $n_i \times 1$ com todos os elementos iguais a 1 e, portanto, $\mathbf{Z}$ é uma matriz diagonal em blocos $n \times k$, cujo i-ésimo elemento da diagonal principal é formado por um vetor coluna de dimensão n_i de uns e $\oplus$ representa a soma direta (Searle at al., 1992).

Com base na definição desse modelo, temos

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{Y}} &= E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \\ \mathbf{V} &= \mathrm{Var}(\mathbf{Y}) = \sigma_b^2 \mathbf{Z}\mathbf{Z}^\top + \sigma_e^2 \mathbf{I}_n \\ \mathbf{C} &= \mathrm{Cov}(\mathbf{b}, \mathbf{Y}^\top) = \sigma_b^2 \mathbf{Z}^\top \end{aligned} \tag{3.5}$$

3.2 Predição Clássica - BLUP e EBLUP

Os três métodos mais comuns de predição de efeitos aleatórios em inferência clássica (Searle et al., 1992; Demidenko, 2004) são: *Melhor Predição*, *Melhor Predição Linear* e *Melhor Predição Linear Não Viciada*. Os respectivos preditores são comumente indicados na literatura por meio das siglas que se referem aos termos em inglês, isto é, BP (*Best Predictor*), BLP (*Best Linear Predictor*) e BLUP (*Best Linear Unbiased Predictor*). Para obter o BP é necessário o conhecimento de todos os parâmetros da distribuição conjunta de $\mathbf{b}$ e $\mathbf{Y}$, denotada por $f(\mathbf{b}, \mathbf{y})$. Para obter o BLP é necessário assumir apenas o conhecimento do primeiro e segundo momentos de $f(\mathbf{b}, \mathbf{y})$ e para obter o BLUP apenas o segundo momento é admitido conhecido (Searle et al., 1992). Vale salientar que tanto o BLP quanto o BLUP são preditores essencialmente não paramétricos, pois para desenvolvê-los não precisamos assumir a forma da distribuição.

Um critério para obter um preditor para $\mathbf{b}$, denotado por $\tilde{\mathbf{b}}$, é minimizar o erro quadrático médio de predição, definido como

$$E[(\tilde{\mathbf{b}} - \mathbf{b})^\top \mathbf{A}(\tilde{\mathbf{b}} - \mathbf{b})] = \int\int (\tilde{\mathbf{b}} - \mathbf{b})^\top \mathbf{A}(\tilde{\mathbf{b}} - \mathbf{b}) f(\mathbf{b}, \mathbf{y}) d\mathbf{y} d\mathbf{b}, \tag{3.6}$$

sendo $\mathbf{A}$ qualquer matriz simétrica positiva definida.

Para fins deste estudo, abordamos apenas a predição via BLUP, pois o termo $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ é desconhecido no modelo (3.3). Como acréscimo de informação, disponibilizamos uma breve descrição dos outros dois métodos no Apêndice A.

O BLUP desenvolvido em Henderson (1975) satisfaz as seguintes propriedades:

i) Melhor, no sentido que minimiza $E[(\tilde{\mathbf{b}} - \mathbf{b})^\top \mathbf{A}(\tilde{\mathbf{b}} - \mathbf{b})]$;

ii) linear em $\mathbf{Y}$, pois é da forma $\tilde{\mathbf{b}} = \mathbf{a} + \mathbf{B}\mathbf{Y}$, com $\mathbf{a}$ e $\mathbf{B}$ não envolvendo $\boldsymbol{\beta}$ e

iii) não viciado, no sentido que $E(\tilde{\mathbf{b}}) = E(\mathbf{b})$.

Condicionalmente ao conhecimento de $\mathbf{C}$ e $\mathbf{V}$ definidas em (3.5), o melhor estimador linear não viciado (*best linear unbiased estimator* - BLUE) de $\boldsymbol{\beta}$ e o BLUP de $\mathbf{b}$ são dados, respectivamente, por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y} \quad \text{e} \quad \tilde{\mathbf{b}} = \mathbf{C}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{3.7}$$

que são soluções das equações de Henderson dadas por

$$\begin{bmatrix} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z} \\ \mathbf{Z}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{Z}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Y} \\ \mathbf{Z}^\top \boldsymbol{\Sigma}^{-1} \mathbf{Y} \end{bmatrix}, \tag{3.8}$$

sendo $\boldsymbol{\Sigma} = \text{Var}(\mathbf{e})$ e $\mathbf{D} = \text{Var}(\mathbf{b})$.

Como no modelo definido em (3.3), $\boldsymbol{\Sigma} = \sigma_e^2 \mathbf{I}_n$ e $\mathbf{D} = \sigma_b^2 \mathbf{I}_k$, podemos simplificar essas equações, ou seja,

$$\begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{Z} \\ \mathbf{Z}^\top \mathbf{X} & \mathbf{Z}^\top \mathbf{Z} + \frac{\sigma_e^2}{\sigma_b^2} \mathbf{I}_k \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \tilde{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top \mathbf{Y} \\ \mathbf{Z}^\top \mathbf{Y} \end{bmatrix}. \tag{3.9}$$

Consequentemente, $\widehat{\boldsymbol{\beta}}$ e $\tilde{\mathbf{b}}$ têm as expressões (3.7) com $\mathbf{C}$ e $\mathbf{V}$ dadas em (3.5).

Robinson (1991) fornece uma ampla discussão sobre o uso do BLUP em modelos lineares mistos gerais para resolver problemas de diversas áreas. A seguir, ilustramos o desenvolvimento da expressão do BLUP de $\mathbf{b}$ em um caso particular do modelo (3.3) sem presença de covariáveis.

3.2.1 Exemplo: ANOVA com um único fator aleatório

Consideremos a equação (3.3) com $\mathbf{X} = \mathbf{1}_n$, em que $\mathbf{1}_n$ denota um vetor $n \times 1$ com todos os elementos iguais a 1, $\boldsymbol{\beta} = \mu$, $\mathbf{Z} = \oplus_{i=1}^{k} \mathbf{1}_{n_i}$ e $\mathbf{b} = (b_1, \ldots, k)$. Portanto, (3.3) se reduz à seguinte forma:

$$\mathbf{Y} = \mathbf{1}_n \mu + \{\oplus_{i=1}^{k} \mathbf{1}_{n_i}\}\mathbf{b} + \mathbf{e}. \tag{3.10}$$

Consequentemente, $\boldsymbol{\mu}_{\mathbf{Y}} = \mu \mathbf{1}_n$, $\mathbf{D} = \sigma_b^2 \mathbf{I}_k$ e $\boldsymbol{\Sigma} = \sigma_e^2 \mathbf{I}_n$. Logo, de acordo com as expressões (3.5), Searle et al. (1992) mostram que

$$\begin{aligned} \mathbf{C} &= \sigma_b^2 \mathbf{I}_k \oplus_{i=1}^{k} \mathbf{1}_{n_i}^{\top} = \sigma_b^2 \oplus_{i=1}^{k} \mathbf{1}_{n_i}^{\top} \\ \mathbf{V} &= \oplus_{i=1}^{k} (\sigma_b^2 \mathbf{J}_{n_i} + \sigma_e^2 \mathbf{I}_{n_i}), \end{aligned} \tag{3.11}$$

em que $\mathbf{J}_{n_i} = \mathbf{1}_{n_i} \mathbf{1}_{n_i}^{\top}$. Consequentemente, temos que

$$\mathbf{V}^{-1} = \frac{1}{\sigma_e^2} \oplus_{i=1}^{k} \left(\mathbf{I}_{n_i} - \frac{\sigma_b^2}{\sigma_e^2 + n_i \sigma_b^2} \mathbf{J}_{n_i} \right). \tag{3.12}$$

Substituindo esses resultados nas equações (3.7), o BLUE de μ e o BLUP de b_i são dados, respectivamente, por

$$\widehat{\mu} = \frac{\displaystyle\sum_{i=1}^{k} \frac{n_i y_{i.}}{\sigma_e^2 + n_i \sigma_b^2}}{\displaystyle\sum_{i=1}^{k} \frac{n_i}{\sigma_e^2 + n_i \sigma_b^2}} \tag{3.13}$$

e

$$\tilde{b}_i = \frac{n_i \sigma_b^2}{\sigma_e^2 + n_i \sigma_b^2} (\bar{Y}_{i.} - \widehat{\mu}), \forall\, i = 1, \ldots, k. \tag{3.14}$$

em que $\bar{Y}_{i.}$ é a média das observações da i-ésima unidade amostral.

3.2.2 BLUE e BLUP Empíricos

De acordo com o modelo (3.3), a existência do BLUP está condicionada ao conhecimento de $\mathbf{C}$ e $\mathbf{V}$ ou, mais especificamente, dos componentes de variância σ_b^2 e σ_e^2. Em situações práticas, dificilmente conhecemos essas quantidades e, portanto, devemos substituí-las por estimativas convenientes na tentativa de encontrar uma aproximação razoável para o BLUE de $\boldsymbol{\beta}$ e para o BLUP de $\mathbf{b}$. Desse modo, geramos o que chamamos de *melhor estimador linear não viciado empírico* (EBLUE) e *melhor preditor linear não viciado empírico* (EBLUP).

Os métodos de estimação mais comumente usados para componentes de variância são máxima verossimilhança e máxima verossimilhança restrita. Uma exigência básica desses métodos é o conhecimento da forma da distribuição dos dados. A suposição de normalidade é mais comum porque torna os procedimentos de obtenção de estimadores e preditores matematicamente tratáveis. Propriedades assintóticas do BLUP considerando os estimadores de máxima verossimilhança restrita (REML) para os componentes de variância são discutidas em Jiang (1997). Além desses métodos, existem na literatura métodos não paramétricos para estimar componentes de variância que são baseados em funções quadráticas dos dados, a exemplo do MINQUE (*minimum norm quadratic unbiased estimator*) e MIVQUE (*minimum variance quadratic unbiased estimator*) (Rao, 1970 e 1971a,b), que são não viciados e têm norma e variância mínima, respectivamente. Outros métodos de estimação são apresentados em Searle et al. (1992) e Demidenko (2004).

Estimação dos Componentes de Variância

Buscando um procedimento não paramétrico de estimação dos componentes de variância, consideramos aqui um processo de estimação iterativo conhecido como I-MINQUE (*iterative minimum norm quadratic unbiased estimator*), que é uma extensão do MINQUE (Searle et al., 1992).

Consideremos o modelo linear misto com r fatores aleatórios, cuja equação é uma extensão de (3.3), ou seja,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{r} \mathbf{Z}_l \mathbf{b}_l + \mathbf{e}, \tag{3.15}$$

em que cada $\mathbf{b}_l$ é um vetor de efeitos aleatórios.

Fazendo $\mathbf{b}_0 = \mathbf{e}$ e $\mathbf{Z}_0 = \mathbf{I}_n$, podemos reescrever (3.15) da seguinte forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=0}^{r} \mathbf{Z}_l \mathbf{b}_l. \tag{3.16}$$

Considerando (3.16), Searle et al. (1992) apresentam as equações para determinar um MINQUE em modelos lineares mistos com r fatores aleatórios e simplificam essas equações para o modelo com um fator aleatório apresentado na Seção 3.2.1, em que $r = 1$ e $\mathbf{Z}_1 = \mathbf{Z}$. Para determinar um MINQUE de (σ_e^2, σ_b^2) no modelo linear misto (3.3), precisamos resolver o seguinte sistema de equações:

$$\begin{bmatrix} s_{00} & s_{01} \\ s_{10} & s_{11} \end{bmatrix} \begin{bmatrix} \sigma_e^2 \\ \sigma_b^2 \end{bmatrix} = \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}, \tag{3.17}$$

ou seja,

$$\begin{aligned} \sigma_e^2 &= (u_0 - s_{10}\sigma_b^2)/s_{00} \\ \sigma_b^2 &= (s_{00}u_1 - s_{10}u_0)/(s_{00}s_{11} - s_{10}^2), \end{aligned} \tag{3.18}$$

em que $s_{00} = SESQ(\mathbf{P}_0)$, $s_{01} = s_{10} = SESQ(\mathbf{P}_0\mathbf{Z})$, $s_{11} = SESQ(\mathbf{Z}^\top\mathbf{P}_0\mathbf{Z})$, $u_0 = SESQ(\mathbf{P}_0\mathbf{Y})$ e $u_1 = SESQ(\mathbf{Z}^\top\mathbf{P}_0\mathbf{Y})$), com SESQ (*sum elements square*) representando a soma dos quadrados dos elementos das matrizes correspondentes e

$$\mathbf{P}_0 = \mathbf{V}_0^{-1} - \mathbf{V}_0^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{V}_0^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}_0^{-1}, \tag{3.19}$$

sendo $\mathbf{V}_0$ uma estimativa de $\mathbf{V}$, expressa em (3.11).

O MINQUE utiliza valores estimados a priori para os componentes de variância em $\mathbf{V}$, obtendo $\mathbf{V}_0$. A partir disso, podemos obter uma solução para o sistema de equações (3.18). Em vez de usar essa primeira solução como estimativas dos componentes de variância, podemos optar por uma solução obtida através de um processo iterativo, que consiste em usar as estimativas MINQUE obtidas em cada iteração como valores a priori na etapa seguinte. Desse modo, o processo segue iterativamente até que a diferença entre duas soluções consecutivas não ultrapasse um valor pequeno, estipulado arbitrariamente. Esse procedimento é denominada *MINQUE iterativo* (I-MINQUE). A vantagem de usar o I-MINQUE é que o mesmo não requer muito esforço computacional, pois se baseia na expressão explícita do MINQUE, além de não requerer o conhecimento da forma das distribuições das quantidades aleatórias envolvidas no modelo. Somemos a isso o fato de as estimativas I-MINQUE, sob a suposição de normalidade, serem iguais às REML quando fornecemos os mesmos valores iniciais e quando há convergência no processo de estimação (Searle, 1992).

Erro Padrão do EBLUE

Sob o modelo (3.3), seja $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\hat{\sigma}_e^2, \hat{\sigma}_b^2)$, o EBLUE de $\widehat{\boldsymbol{\beta}}$ expresso em (3.7), em que $\hat{\sigma}_e^2$ e $\hat{\sigma}_b^2$ são os respectivos estimadores de σ_e^2 e σ_b^2.

Com base em (3.7), é fácil verificar que um estimador simples da matriz de covariâncias do EBLUE, que ignora a variabilidade de $\hat{\sigma}_e^2$ e $\hat{\sigma}_b^2$, é dado por

$$\widehat{\mathrm{Var}}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^\top \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}, \tag{3.20}$$

em que $\hat{\mathbf{V}}$ depende de $\hat{\sigma}_e^2$ e $\hat{\sigma}_b^2$. Assim, um estimador do erro padrão de $\tilde{\beta}_l$, $l = 0, \ldots, p-1$ é dado pela raiz quadrada do l-ésimo elemento da diagonal principal de (3.20).

Alternativas para determinar estimadores do erro padrão do EBLUE foram propostas por Das et al. (2004), que fazem um estudo sobre erro quadrático médio do EBLUP e citam alguns trabalhos sobre o assunto.

3.3 Predição Bayesiana - Bayes Empírico

Na abordagem bayesiana, além da especificação do modelo para o vetor $n \times 1$ de respostas $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, \ldots, \mathbf{Y}_k^\top)^\top$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, $n = \sum_{i=1}^k n_i$, dado o vetor de parâmetros $\mathbf{b} = (b_1, b_2, \ldots, b_k)^\top$, usualmente na forma de uma distribuição de probabilidade condicional $f(\mathbf{y}|\mathbf{b})$, supomos que $\mathbf{b}$ é uma quantidade aleatória tendo uma distribuição *a priori* $g(\mathbf{b}|\boldsymbol{\theta})$, sendo $\boldsymbol{\theta}$ o vetor de hiperparâmetros (Carlin e Louis, 1998). A inferência sobre $\mathbf{b}$ é baseada em sua distribuição *a posteriori*, dada por

$$\pi(\mathbf{b}|\mathbf{y}, \boldsymbol{\theta}) = \frac{f(\mathbf{y}, \mathbf{b}|\boldsymbol{\theta})}{m(\mathbf{y}|\boldsymbol{\theta})} = \frac{f(\mathbf{y}, \mathbf{b}|\boldsymbol{\theta})}{\int f(\mathbf{y}, \mathbf{b}|\boldsymbol{\theta}) d\mathbf{b}} = \frac{f(\mathbf{y}|\mathbf{b}) g(\mathbf{b}|\boldsymbol{\theta})}{\int f(\mathbf{y}|\mathbf{b}) g(\mathbf{b}|\boldsymbol{\theta}) d\mathbf{b}}, \tag{3.21}$$

com $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\theta})$ representando a distribuição conjunta de $(\mathbf{Y}, \mathbf{b})$ dado $\boldsymbol{\theta}$ e $m(\mathbf{y}|\boldsymbol{\theta})$, a distribuição marginal de $\mathbf{Y}$ dado $\boldsymbol{\theta}$. É conveniente ressaltar que a distribuição a posteriori de $\mathbf{b}$ depende do vetor de parâmetros $\boldsymbol{\theta}$. Isto implica uma hierarquia, pois para que $\mathbf{b}$ possa ser estimado, se faz necessário estimar primeiro $\boldsymbol{\theta}$. Se desconhecermos o valor $\boldsymbol{\theta}$, a solução bayesiana pode ser usada para quantificar esta incerteza considerando uma distribuição a priori para $\boldsymbol{\theta}$, chamada de *hiperpriori*. Denotando esta priori por $h(\boldsymbol{\theta})$, a posteriori desejada para $\mathbf{b}$ é obtida pela marginalização sobre $\boldsymbol{\theta}$, isto é,

$$\pi(\mathbf{b}|\mathbf{y}) = \frac{f(\mathbf{y}, \mathbf{b})}{m(\mathbf{y})} = \frac{\int f(\mathbf{y}, \mathbf{b}, \boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \int f(\mathbf{y}, \mathbf{b}, \boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{b}} = \frac{\int f(\mathbf{y}|\mathbf{b}) g(\mathbf{b}|\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \int f(\mathbf{y}|\mathbf{b}) g(\mathbf{b}|\boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{b}}. \tag{3.22}$$

Alternativamente ao uso de uma hiperpriori para $\boldsymbol{\theta}$, podemos simplesmente substituir $\boldsymbol{\theta}$ por uma estimativa $\hat{\boldsymbol{\theta}}$ obtida como o valor que maximiza a distribuição marginal $m(\mathbf{y}|\boldsymbol{\theta})$

na equação (3.21), vista como função de $\boldsymbol{\theta}$. Desse modo, a inferência sobre $\mathbf{b}$ pode ser baseada em uma distribuição a posteriori estimada, $\pi(\mathbf{b}|\mathbf{y};\hat{\boldsymbol{\theta}})$, que é obtida da equação (3.21) substituindo-se $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$. Tal abordagem é referida como *bayesiana empírica*, pois a estimação é feita com base nos dados amostrais ao invés de supormos uma hiperpriori para $\boldsymbol{\theta}$.

Uma vez que a distribuição a posteriori estimada tenha sido determinada, ela pode ser usada para se fazer a predição de $\mathbf{b}$. O estimador usado é então chamado de *estimador de Bayes empírico* (EBE) que, dependendo da função de perda considerada, pode ser a média, a mediana ou a moda de $\pi(\mathbf{b}|\mathbf{y};\hat{\boldsymbol{\theta}})$.

No caso do modelo linear misto (3.3), a abordagem bayesiana empírica pode ser empregada para predizer um valor para o vetor de efeitos aleatórios $\mathbf{b}$. Todavia, vale salientar que no modelo (3.3), a distribuição conjunta de $\mathbf{Y}$ e $\mathbf{b}$ depende dos parâmetros $\boldsymbol{\beta}$ e σ_e^2, além do hiperparâmetro representado por σ_b^2. Assim, considerando $\boldsymbol{\lambda} = (\boldsymbol{\beta}^\top, \sigma_e^2, \sigma_b^2)^\top$ o vetor de parâmetros desconhecidos do modelo, a distribuição a posteriori estimada de $\mathbf{b}$, denotada por $\pi(\mathbf{b}|\mathbf{y};\hat{\boldsymbol{\lambda}})$, é

$$\pi(\mathbf{b}|\mathbf{y};\hat{\boldsymbol{\lambda}}) = \frac{f(\mathbf{y},\mathbf{b};\hat{\boldsymbol{\lambda}})}{m(\mathbf{y};\hat{\boldsymbol{\lambda}})} = \frac{f(\mathbf{y}|\mathbf{b};\hat{\boldsymbol{\beta}},\hat{\sigma}_e^2)g(\mathbf{b};\hat{\sigma}_b^2)}{\int f(\mathbf{y}|\mathbf{b};\hat{\boldsymbol{\beta}},\hat{\sigma}_e^2)g(\mathbf{b};\hat{\sigma}_b^2)d\mathbf{b}}, \tag{3.23}$$

em que $f(\mathbf{y}|\mathbf{b};\hat{\boldsymbol{\beta}},\hat{\sigma}_e^2)$ é a distribuição condicional de $\mathbf{Y}$ dado $\mathbf{b}$, com $\boldsymbol{\beta}$ e σ_e^2 substituídos por $\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}_e^2$, respectivamente e $g(\mathbf{b};\hat{\sigma}_b^2)$, a distribuição a priori de $\mathbf{b}$ substituindo σ_b^2 por $\hat{\sigma}_b^2$.

3.3.1 Normalidade: relação entre o EBLUP e o EBE

Sob suposição de normalidade, isto é,

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{Y} \end{bmatrix} \sim N_{k+n}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{V} \end{bmatrix} \right), \tag{3.24}$$

temos que $\mathbf{b}|\mathbf{Y} \sim N_k(\mathbf{C}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{D} - \mathbf{C}\mathbf{V}^{-1}\mathbf{C}^\top)$. Ademais, a expressão da média da distribuição condicional de $\mathbf{b}|\mathbf{Y}$ é equivalente ao BLUP de (3.7). Substituindo $\mathbf{C}$ e $\mathbf{V}$ por estimativas convenientes, determinadas pela maximização da distribuição marginal de $\mathbf{Y}$ e $\boldsymbol{\beta}$ por seu EBLUE, podemos obter o EBLUP de $\mathbf{b}$. Nesse caso, o EBLUP pode ser considerado um estimador de Bayes empírico, pois é a média da distribuição a posteriori de $\mathbf{b}$ estimada. Outras relações entre o EBLUP e o estimador de Bayes empírico podem ser encontradas em Karunamuni (2002).

Ao longo deste capítulo, abordamos procedimentos de predição em modelos lineares mistos com um único fator aleatório para dados não censurados. No capítulo seguinte, tratamos de procedimentos de predição em um tipo de modelo linear misto utilizado em Análise de Sobrevivência para analisar dados correlacionados e censurados. Estudamos a aplicação das metodologias do EBLUP e do estimador de Bayes empírico nessa classe de modelos.

Capítulo 4

Predição em Modelos de Tempo de Falha Acelerado com Efeito Aleatório

Consideremos uma amostra aleatória de k unidades amostrais em que se registram n_i observações para a i-ésima unidade amostral no decorrer do período de estudo. Sejam as respostas dadas por $Y_{ij} = \min(\ln T_{ij}, \ln C_{ij})$, $i = 1, 2, \ldots, k$ e $j = 1, 2, \ldots, n_i$, em que T_{ij} representa o tempo até a j-ésima falha da unidade amostral i e C_{ij}, o tempo até a j-ésima censura de i e seja $\delta_{ij} = I(T_{ij} \leq C_{ij})$ o indicador de falhas.O modelo de tempo de falha acelerado com efeito aleatório é definido como

$$\ln T_{ij} = b_i + \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \sigma\epsilon_{ij}, \tag{4.1}$$

em que $\mathbf{x}_{ij}$ denota um vetor $p \times 1$ de covariáveis com a primeira componente igual a 1, $\boldsymbol{\beta}$ um vetor $p \times 1$ de parâmetros referentes aos efeitos fixos, σ um parâmetro de escala, ϵ_{ij} são erros aleatórios independentes e identicamente distribuídos com média e variância conhecidas e b_i, $i = 1, 2, \ldots, k$, são variáveis aleatórias independentes e identicamente distribuídas com função densidade $g(b_i; \sigma_b^2)$, média zero e variância σ_b^2. Além disso, assumimos que:

(i) $\mathrm{Cov}(b_i, \epsilon_{ij}) = 0$;

(ii) condicionalmente ao efeito aleatório b_i, as respostas referentes à i-ésima unidade amostral são independentes e

(iii) os efeitos aleatórios b_i são independentes dos tempos de censura. Quando $\sigma_b^2 = 0$, esse modelo se torna o modelo de tempo de falha acelerado usual dado em (2.11).

O ajuste do modelo (4.1) aos dados consiste na estimação de $\boldsymbol{\beta}$ e σ e na predição de b_i.

4.1 Predição Via Estimador de Bayes Empírico

Seja $\boldsymbol{\lambda} = (\boldsymbol{\beta}^\top, \sigma, \sigma_b^2)^\top$ o vetor de parâmetros desconhecidos do modelo descrito em (4.1). Para que possamos prever cada efeito aleatório b_i, precisamos antes estimar o vetor de parâmetros $\boldsymbol{\lambda}$. Nos modelos com efeito aleatório, isso geralmente é feito usando métodos de máxima verossimilhança que se baseiam na distribuição marginal das respostas (Y_{ij}, δ_{ij}), $i = 1, 2, \ldots, k$ e $j = 1, 2, \ldots, n_i$, vista como função de $\boldsymbol{\lambda}$ e denominada função de verossimilhança marginal.

4.1.1 Função de Verossimilhança Marginal

Sejam f e S as respectivas funções densidade e de sobrevivência condicionais de $\ln T_{ij}$ dado b_i. Assumindo que o mecanismo de censura é aleatório, podemos representar a função de verossimilhança relativa à distribuição condicional de (Y_{ij}, δ_{ij}) dado o efeito aleatório b_i como

$$f(y_{ij}|b_i, \mathbf{x}_{ij}, \boldsymbol{\beta}, \sigma)^{\delta_{ij}} S(y_{ij}|b_i, \mathbf{x}_{ij}, \boldsymbol{\beta}, \sigma)^{(1-\delta_{ij})}. \tag{4.2}$$

As observações para a i-ésima unidade amostral são $(\mathbf{Y}_i, \boldsymbol{\delta}_i) = ((Y_{i1}, \delta_{i1}), \ldots, (Y_{in_i}, \delta_{in_i}))^\top$. Pela suposição de independência condicional de $(\mathbf{Y}_i, \boldsymbol{\delta}_i)$ dado b_i, temos que a função de verossimilhança relativa à distribuição condicional $(\mathbf{Y}_i, \boldsymbol{\delta}_i)$ dado b_i é

$$L_i(\boldsymbol{\beta}, \sigma | b_i) = \prod_{j=1}^{n_i} f(y_{ij}|b_i, \mathbf{x}_{ij})^{\delta_{ij}} S(y_{ij}|b_i, \mathbf{x}_{ij})^{(1-\delta_{ij})}, \tag{4.3}$$

para $i = 1, 2, ..., k$. Consequentemente, a função de verossimilhança relativa à distribuição marginal de $(\mathbf{Y}_i, \boldsymbol{\delta}_i)$ tem a seguinte expressão:

$$L_i(\boldsymbol{\lambda}) = \int L_i(\boldsymbol{\beta}, \sigma | b_i) g(b_i; \sigma_b^2) db_i. \tag{4.4}$$

Finalmente, assumindo independência entre os vetores de variáveis aleatórias $(\mathbf{Y}_1, \boldsymbol{\delta}_1)$, $(\mathbf{Y}_2, \boldsymbol{\delta}_2)$, ..., $(\mathbf{Y}_k, \boldsymbol{\delta}_k)$, concluímos que a função de verossimilhança marginal para toda a amostra é

$$L(\boldsymbol{\lambda}) = \prod_{i=1}^{k} L_i(\boldsymbol{\lambda}) = \prod_{i=1}^{k} \int L_i(\boldsymbol{\beta}, \sigma | b_i) g(b_i; \sigma_b^2) db_i. \tag{4.5}$$

Dependendo das distribuições postuladas para o efeito aleatório e/ou para os tempos entre falhas, a integral em (4.5) pode não ser analiticamente tratável. Um exemplo típico é quando assumimos uma distribuição normal para o efeito aleatório. Nesse caso, podemos adotar as ideias apresentadas em Valença (2003) e Lambert et al. (2004), que utilizam uma quadratura gaussiana adaptada (Liu e Pierce, 1994) para aproximar a integral em (4.5).

4.1.2 Predição dos Efeitos Aleatórios

Assim como Lambert et al. (2004), consideramos como preditor de $\mathbf{b}$, a moda de sua distribuição a posteriori estimada. Na classe de modelos considerada essa distribuição é dada por

$$\pi(\mathbf{b}|\mathbf{y}, \hat{\boldsymbol{\lambda}}) = \frac{\prod_{i=1}^{k} L_i(\hat{\boldsymbol{\beta}}, \hat{\sigma}|b_i) g(b_i; \hat{\sigma}_b^2)}{L(\hat{\boldsymbol{\lambda}})}, \tag{4.6}$$

em que $\hat{\boldsymbol{\lambda}}$ é a estimativa de máxima verossimilhança de $\boldsymbol{\lambda}$, obtida pela maximização de (4.5).

Notemos que $\pi(\mathbf{b}|\mathbf{y}, \hat{\boldsymbol{\lambda}})$ depende de $\mathbf{b}$ somente através de $L_i(\hat{\boldsymbol{\beta}}, \hat{\sigma}|b_i)$ e de $g(b_i; \hat{\sigma}_b^2)$ e, assim, o estimador de Bayes empírico de b_i é definido como o ponto que maximiza

$$L_i(\hat{\boldsymbol{\beta}}, \hat{\sigma}|b_i) g(b_i; \hat{\sigma}_b^2) \tag{4.7}$$

ou equivalentemente,

$$\ln[L_i(\hat{\boldsymbol{\beta}}, \hat{\sigma}|b_i) g(b_i; \hat{\sigma}_b^2)]. \tag{4.8}$$

Para ajustar o modelo (4.1) através da abordagem bayesiana empírica, podemos usar o procedimento NLMIXED (*nonlinear mixed*) do *Statistical Analysis Sistem* - SAS (SAS, 2009, versão 9.1), que utiliza o método numérico quase-Newton (Davison, 1991) como método padrão de maximização.

4.2 Predição Via EBLUP

Por se tratar de um método não paramétrico e, consequentemente, não assumir forma distribucional para o efeito aleatório nem para o erro aleatório, a predição de efeitos aleatórios em modelos de tempo de falha acelerado com efeito aleatório via EBLUP é com-

putacionalmente simples. Em outras palavras, o ajuste do modelo não envolve resolução de integrais complicadas como geralmente ocorre na predição via estimador de Bayes empírico.

Escrevendo o modelo (4.1) em forma matricial, temos

$$\ln \mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \sigma\boldsymbol{\epsilon}, \tag{4.9}$$

em que $\mathbf{T} = (\mathbf{T}_1^\top, \mathbf{T}_2^\top, \ldots, \mathbf{T}_k^\top)^\top$, $\mathbf{T}_i = (T_{i1}, T_{i2}, \ldots, T_{in_i})^\top$, $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top, \ldots, \mathbf{X}_k^\top)^\top$, $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})^\top$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$, $\mathbf{b} = (b_1, b_2, \ldots, b_k)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$ e $\mathbf{Z}$ é especificada como em (3.4).

Vale destacar que o EBLUP é obtido a partir do modelo linear misto (3.3), que difere daquele especificado em (4.9), pois $\mathrm{E}(\boldsymbol{\epsilon}) \neq 0$ e, consequentemente, $\mathrm{E}(\ln \mathbf{T}) \neq \mathbf{X}\boldsymbol{\beta}$.

Para contornar esse problema, fazemos $\mathbf{e} = \sigma[\boldsymbol{\epsilon} - \mathrm{E}(\boldsymbol{\epsilon})]$, o que implica $\mathrm{E}(\mathbf{e}) = \mathbf{0}$ e $\mathrm{Var}(\mathbf{e}) = \sigma^2 \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\sigma_\epsilon^2\mathbf{I}_n = \sigma_e^2\mathbf{I}_n$. Além disso, como $\sigma\boldsymbol{\epsilon} = \mathbf{e} + \sigma\,\mathrm{E}(\boldsymbol{\epsilon})$, podemos reescrever (4.9) da seguinte forma:

$$\ln \mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \tag{4.10}$$

em que $\boldsymbol{\beta} = (\beta_0 + \sigma\,\mathrm{E}(\epsilon_i), \beta_1, \ldots, \beta_{p-1})^\top$.

Notemos que o modelo (3.3) difere de (4.10) pelo fato deste ter como resposta $\mathbf{Y} = \min(\ln \mathbf{T}, \ln \mathbf{C})$, em que $\mathbf{C} = (\mathbf{C}_1^\top, \mathbf{C}_2^\top, \ldots, \mathbf{C}_k^\top)^\top$, $\mathbf{C}_i = (C_{i1}, C_{i2}, \ldots, C_{in_i})^\top$, é o vetor de censuras. De fato, podemos escrever

$$Y_{ij} = \delta_{ij} \ln T_{ij} + (1 - \delta_{ij}) \ln C_{ij}. \tag{4.11}$$

Dessa forma, se $\delta_{ij} = 1$, $\forall i, j$, temos que (4.10) representa um modelo linear misto com $\mathbf{Y} = \ln \mathbf{T}$ e satisfazendo as propriedades estabelecidas em (3.3) e, portanto, podemos aplicar diretamente as expressões do EBLUE e EBLUP dadas em (3.7) para ajustar o

modelo (4.10). Contudo, se $\delta_{ij} = 0$ para algum i, j, não observamos T_{ij}, mas C_{ij}, que pode ser considerado como uma "imputação da observação censurada", mas que de fato subestima o verdadeiro tempo até a falha. Quando temos um percentual de censura acima de 30%, Ageel (2002) propõe usar como imputação da observação censurada a média dos tempos até a falha maiores do que C_{ij} para melhorar as estimativas do verdadeiro tempo até a falha. Essa ideia parece uma alternativa razoável para não subestimar o EBLUE e o EBLUP quando o percentual de censura é relativamente alto.

O estudo detalhado e comparativo de procedimentos para tratar de modelos lineares mistos com observações censuradas no contexto de modelos de tempo de falha está fora do escopo deste trabalho. Todavia, citamos alguns trabalhos relacionados ao assunto. Em um contexto de modelos de regressão com dados censurados, existem diversas abordagens paramétricas e semiparamétricas. Taga (2008) cita vários trabalhos que abordam esse tema. O estudo de modelos lineares mistos com dados censurados é abordado por Hughes(1999) e Pettitt(1986), entre outros. Nestes casos, a abordagem é predominantemente paramétrica.

4.3 Predição de Riscos e Probabilidades de Falha

Um objetivo particular deste trabalho é a predição da função de risco da distribuição Weibull. Assumindo a distribuição valor extremo padrão (2.10) para ϵ_{ij} no modelo (4.1) e com base na expressão (2.13), temos que a função de risco de T_{ij} condicionada ao efeito aleatório b_i é dada por

$$h(t; b_i, \mathbf{x}_{ij}, \sigma, \boldsymbol{\beta}) = \sigma^{-1} t^{(\sigma^{-1}-1)} exp[-\sigma^{-1}(b_i + \mathbf{x}_{ij}^{\top}\boldsymbol{\beta})]. \tag{4.12}$$

Esta função de risco representa a taxa de falha instantânea no tempo t condicionada ao efeito aleatório b_i da i-ésima unidade amostral. Alternativamente, podemos estar inte-

ressados em obter a probabilidade condicional de falha em um intervalo de tempo, que é expressa por

$$P[t < T_{ij} \leq t + \Delta t | T_{ij} > t, b_i, \mathbf{x}_{ij}] = \frac{S(t|b_i, \mathbf{x}_{ij}) - S(t + \Delta t|b_i, \mathbf{x}_{ij})}{S(t|b_i, \mathbf{x}_{ij})}, \tag{4.13}$$

em que $S(t|b_i, \mathbf{x}_{ij})$ é a função de sobrevivência de T_{ij} dado b_i, ou seja,

$$S(t|b_i, \mathbf{x}_{ij}) = \exp\{-\exp\left[-\sigma^{-1}(b_i + \mathbf{x}_{ij}^{\top}\boldsymbol{\beta})\right] t^{\sigma^{-1}}\}. \tag{4.14}$$

Em relação ao nosso estudo, a expressão (4.13) é interpretada como a probabilidade do poço falhar nas próximas Δt unidades de tempo dado que vem funcionando sem falhar há t unidades de tempo.

Capítulo 5

Aplicação e Discussão

Neste capítulo, estudamos os tempos entre falhas de equipamentos de subsuperficie de poços petrolíferos da Bacia Potiguar, ocorridas no período de janeiro de 2000 a dezembro de 2006. Ajustamos o modelo de tempo de falha acelerado com efeito aleatório por ambos os métodos estudados e em seguida, por meio de uma abordagem particular, comparamos os desempenhos dos modelos ajustados na predição dos riscos de falha dos poços. Além disso, através de um método de validação externa, investigamos a adequação desses modelos em prever futuras falhas.

5.1 Ajustes do Modelo

Embora a maioria dos poços da Bacia Potiguar seja constituída por apenas uma coluna de produção - sistema de equipamentos responsável pela elevação do fluido até a superfície - existem poços com duas colunas de produção que funcionam de maneira independente. Para efeito deste estudo, as colunas de produção, denominadas de poços-coluna, são as unidades amostrais. Os dados analisados correspondem a uma amostra de 616 poços-coluna. Detalhes sobre o processo de amostragem podem ser vistos em Dantas

(2008). Para os poços-coluna amostrados foram registradas 2374 observações correspondentes aos tempos entre falhas, das quais 563 (23,7%) são censuras.

Seguindo a proposta de seleção de covariáveis sugerida por Collett (1994), o modelo selecionado em Santos (2009) considera as seguintes covariáveis:

- *Produção base do poço-coluna (PROD)*, medida em m^3/dia;
- *Idade do poço-coluna (ID)*, medida no momento da falha, em anos;
- *Método de elevação*: BM e BCP;
- *Profundidade (PROF)* onde se encontra instalada a bomba de produção do poço-coluna, medida em metros e
- *Unidade administrativa*, que se refere à localização geográfica dos poços. São elas: Unidade Operacional Alto do Rodrigues (ARG), Unidade Operacional Canto do Amaro (CAM), Unidade Operacional Campo de Estreito (ET) e Unidade Operacional Fazendo Riacho da Forquilha (RFQ).

Vale salientar que as interações da unidade administrativa com a produção e com a profundidade foram significativas. Portanto, a equação do modelo de tempo de falha acelerado com efeito aleatório proposto em Santos (2009) é expressa por

$$\begin{aligned}\ln T_{ij} = {} & b_i + \beta_{prod} PROD_{ij} + \beta_{bm} BM + \beta_{id} ID_{ij} + \beta_{cam} CAM + \beta_{et} ET + \beta_{rfq} RFQ + \\ & + \beta_{prof} PROF_i + \beta_{prod*cam} PROD_{ij} * CAM + \beta_{prod*et} PROD_{ij} * ET + \\ & + \beta_{prod*rfq} PROD_{ij} * RFQ + \beta_{prof*cam} PROF_i * CAM + \\ & + \beta_{prof*et} PROFB_i * ET + \beta_{prof*rfq} PROF_i * RFQ + \sigma \epsilon_{ij},\end{aligned} \tag{5.1}$$

em que assumiu-se distribuição normal para b_i, isto é, $b_i \sim N(0, \sigma_b^2)$, e distribuição condicional Weibull para T_{ij} dado b_i, $i = 1, \ldots, 616$ e $j = 1, \ldots, n_i$.

Considerando essas distribuições, o ajuste do modelo com base no estimador de Bayes empírico foi realizado usando o procedimento NLMIXED do SAS (SAS, 2009, versão 9.1). Para dar início ao processo iterativo, consideramos como valores iniciais para o vetor de coeficientes das covariáveis e para o parâmetro de escala σ as estimativas obtidas sob independência entre os tempos, ou seja, pelo modelo de tempo de falha acelerado usual e atribuímos 0,1 como valor inicial de σ_b^2 (ver Apêndice C). O NLMIXED além das estimativas dos parâmetros e seus erros padrão, fornece os efeitos aleatórios preditos.

Em relação ao EBLUP, o ajuste foi feito com auxílio do R (R, 2008, versão 2.8.1). Para a obtenção das estimativas dos componentes de variância através do método I-MINQUE, foram tomados como valores a priori, as estimativas obtidas pelo modelo de tempo de falha acelerado usual, ou seja, $\sigma_b^2 = 0$ e $\sigma = 1,3365$. Os erros padrão foram obtidos através da expressão (3.20). Os programas usados para determinar o I-MINQUE, EBLUE e erros padrão, assim como o EBLUP também encontram-se no Apêndice C.

A Tabela 5.1 mostra as estimativas do vetor de coeficientes das covariáveis obtidas através dos métodos Estimador de Bayes Empírico (EBE) e do EBLUP, além dos erros padrão das estimativas (E.P). Os primeiros resultados (colunas 2 a 5) tanto do EBE como do EBLUP se referem ao ajuste feito para o banco de dados completo (BDC), isto é, considerando os dados de todo o período (2000 a 2006), enquanto que os últimos (colunas 6 a 9) dizem respeito ao ajuste considerando apenas os dados de 2000 a 2005, que chamamos de banco de dados parcial (BDP). Usamos o BDP na validação do modelo, que discutimos mais adiante na Seção 5.4. Nesta e nas duas seguintes seções damos ênfase ao BDC.

De acordo com essa tabela, constatamos que o aumento da idade, bem como bombas instaladas mais profundamente contribui para aumentar o tempo de funcionamento (sem falhas) do poço-coluna. Com respeito à idade, uma justificativa técnica para esse fato é que à medida que o poço-coluna vai "envelhecendo" fica mais resistente, devido aos ajustes que são feitos em seus equipamentos. Por outro lado, quando a produção aumenta

Tabela 5.1: Resultados dos ajustes

	BDC (2000 a 2006)				BDP (2000 a 2005)			
	EBE		EBLUP		EBE		EBLUP	
Parâmetro	Estimativa	E.P	Estimativa	E.P	Estimativa	E.P	Estimativa	E.P
β_0	7,4698	0,2572	7,1274	0,2480	7,2389	0,2071	6,5580	0,2679
β_{prod}	-0,0498	0,0091	-0,0298	0,0115	-0,0228	0,0087	-0,0056	0,0121
β_{bm}	0,5271	0,1284	0,2804	0,1278	0,4970	0,1057	0,4048	0,1384
β_{id}	0,0787	0,0077	0,0514	0,0071	0,0465	0,0064	0,0425	0,0082
β_{cam}	1,3428	0,3071	0,8513	0,2994	1,3649	0,2825	1,4458	0,3486
β_{et}	0,9589	0,2275	0,8952	0,2182	0,9511	0,1744	1,2200	0,2260
β_{rfq}	1,8259	0,3279	1,3441	0,3188	1,1991	0,2685	1,0125	0,3454
β_{prof}	0,0021	0,0004	0,0014	0,0004	0,0019	0,0003	0,0018	0,0004
$\beta_{prod*cam}$	0,0323	0,0153	0,0347	0,0184	0,0171	0,0153	0,0008	0,0200
$\beta_{prod*et}$	0,0189	0,0139	0,0002	0,0174	0,0073	0,0132	-0,0102	0,0181
$\beta_{prod*rfq}$	-0,0407	0,0188	-0,0467	0,0225	-0,0308	0,0177	-0,0384	0,0237
$\beta_{cam*prof}$	-0,0020	0,0004	-0,0013	0,0005	-0,0017	0,0005	-0,0016	0,0006
$\beta_{et*prof}$	-0,0028	0,0005	-0,0022	0,0005	-0,0025	0,0004	-0,0029	0,0006
$\beta_{rfq*prof}$	-0,0028	0,0006	-0,0020	0,0006	-0,0022	0,0005	-0,0019	0,0007
σ	1,2001	0,0253	1,1592	-	1,1320	0,0258	1,1789	-
σ_b^2	0,4369	0,0582	0,3587	-	0,1292	0,0311	0,2160	-

(diminui), o tempo de funcionamento do poço-coluna até à falha diminui (aumenta). Isso se deve ao fato de quanto maior a produção, maior é a potência com que o poço-coluna trabalha, provocando um maior desgaste nos equipamentos. Também podemos concluir que os poços-colunas que usam o bombeio mecânico como método de elevação apresentam maiores tempos até a falha do que aqueles que usam o bombeio por cavidades progressivas e que os tempos de funcionamento até a falha dos poços-coluna das unidades administrativas CAM, ET e RFQ são maiores que os da ARG.

5.2 Riscos de Falha Preditos

Utilizamos a função de risco (4.12) para prever os riscos de falha considerando os ajustes dos modelos aos dados do banco de dados completo. Na Figura 5.1, são apresentados os *box plots* (diagramas de caixa) dos riscos obtidos na escala logarítmica considerando o estimador de Bayes empírico e o EBLUP, respectivamente. Esses gráficos servem como

uma forma de diagnóstico do modelo, pois os riscos foram preditos para os instantes onde de fato as falhas ocorreram. Nesse sentido, esperamos que um método adequado seja capaz de predizer riscos relativamente "altos" nesses instantes. De acordo com essa figura, constatamos que o EBLUP parece mais adequado em relação ao EBE, pois, em geral, o EBLUP aponta maiores riscos de falha nos instantes referidos.

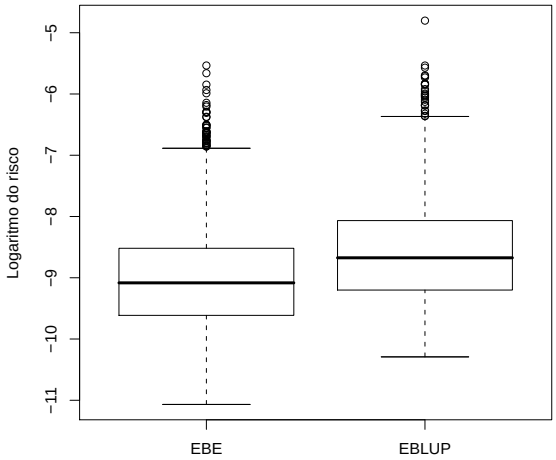

Figura 5.1: *Box plots* dos logaritmos dos riscos considerando o EBE e o EBLUP, respectivamente.

Quando observamos a evolução dessas predições ao longo dos anos, Figuras 5.2 e 5.3, percebemos pela Figura 5.2 que há uma tendência decrescente nos valores medianos dos logaritmos dos riscos preditos. Esse fato levanta a suspeita de que o modelo considerando o EBE está perdendo sua capacidade de detectar as falhas com o passar dos anos. Já em relação a Figura 5.3, essa tendência é menos perceptível, uma vez que nos anos intermediários (2001-2005) a evolução se mostra um pouco mais estável. Todavia, quando comparamos o ano de 2000 com os demais anos, notamos diferenças aparentemente mais significativas em ambas as figuras.

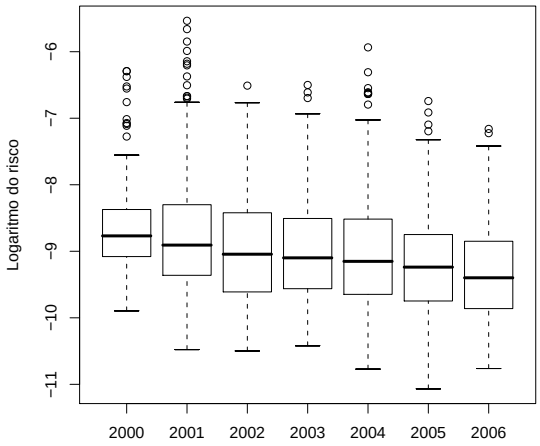

Figura 5.2: *Box plots* dos logaritmos dos riscos por ano - EBE.

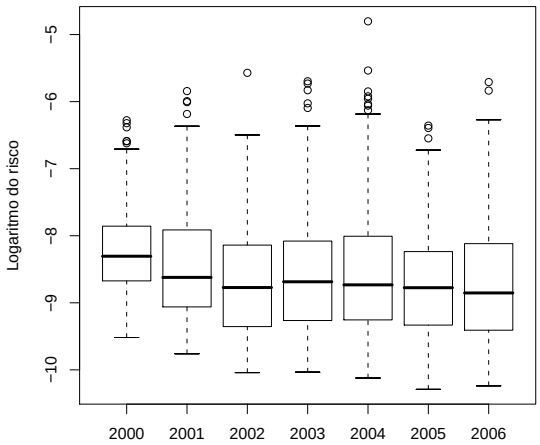

Figura 5.3: *Box plots* dos logaritmos dos riscos por ano - EBLUP.

Os riscos de falha preditos podem ser úteis nas tomadas de decisão por parte dos gestores da PETROBRAS com respeito às manutenções preventivas dos poços-coluna. Com a finalidade de utilizá-los como subsídio para esse fim, sugerimos predizê-los em instantes de interesse considerando as informações mais atuais das covariáveis envolvidas no modelo e, posteriormente, identificar os poços-coluna cujos riscos preditos estejam entre os maiores. Dessa maneira, os poços-coluna identificados deverão ser os eleitos para a manutenção preventiva. Uma opção seria selecionar os poços-coluna com riscos de falha entre os 10%, 5% ou 1% maiores, por exemplo.

5.3 Efeitos Aleatórios Preditos

As análises das Figuras 5.2 e 5.3 sugerem aparentemente que o ajuste sob o EBLUP é mais adequado que o EBE para prever riscos de falha. Com base nos histogramas (Figuras 5.4 e 5.5), construídos através das estimativas da densidade de Kernel (Scott, 1992) implementadas no R, suspeitamos que uma das razões para esse fato está na distribuição assumida para os efeitos aleatórios no EBE. Pois, podemos notar indicativos de assimetria na distribuição dos efeitos aleatórios preditos em ambos os histogramas e, portanto, a suposição de normalidade para os efeitos aleatórios parece não ser adequada. Nesse caso, seria conveniente considerar outras distribuições para os efeitos aleatórios. Mas, vamos deixar essa tarefa para trabalhos futuros.

5.4 Validação do Modelo

Alternativamente ao uso dos riscos de falha preditos na manutenção preventiva, podemos utilizar as probabilidades condicionais de falha preditas. Uma vantagem que essas probabilidades oferecem é que diferentemente dos riscos de falha, elas levam em consideração

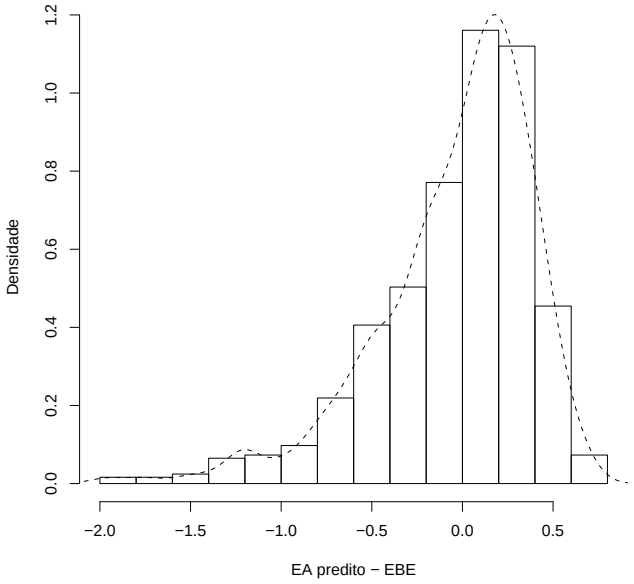

Figura 5.4: Histograma dos efeitos aleatórios (EA) preditos pelo EBE

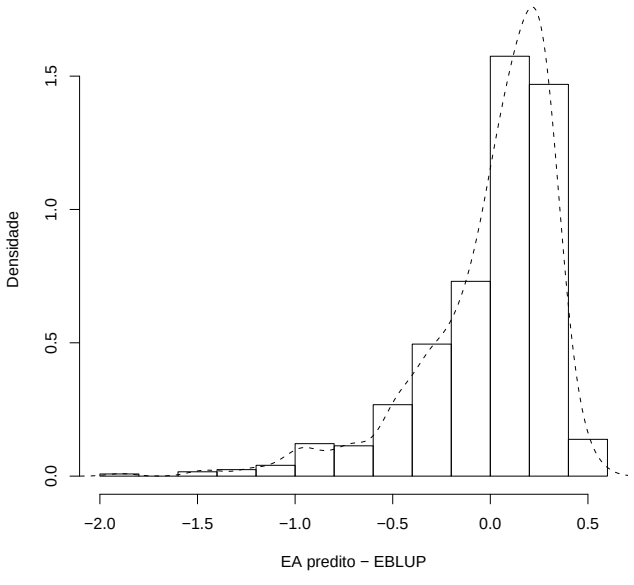

Figura 5.5: Histograma dos efeitos aleatórios (EA) preditos pelo EBLUP

o tempo que o poço-coluna vem funcionando sem falhar desde a última manutenção, além de apresentar valores limitados entre 0 e 1. Assim, poderíamos selecionar os poços-coluna com probabilidades condicionais de falha acima de 50%, 60% , 70%, 80% ou 90%, conforme desejado.

Nesta seção, utilizamos as probabilidades condicionais de falha como uma ferramenta de diagnóstico para avaliar se o modelo de tempo de falha acelerado com efeito aleatório ajustado por ambos os métodos é adequado para fazer previsões de futuras falhas e, consequentemente, identificar os poços-coluna com maiores probabilidades de falha para a manutenção preventiva.

Usamos ideias básicas de validação externa (ver, por exemplo, Harrekll e Frank, 2001) em que removemos de nossa amostra os dados referentes ao ano de 2006 (usados para validação) e ajustamos os modelos estudados ao conjunto de dados do período inicial (2000 a 2005). Assim, com base nas ocorrências de falhas de 2006, desejamos avaliar a capacidade dos modelos em indicar decisões corretas com respeito à manutenção preventiva ou não dos poços em estudo. Para isso, procedemos da seguinte forma:

(i) Com base nos dados de 2000-2005, estimamos as probabilidades condicionais de falha em um período de Δt horas dado que o poço-coluna estava funcionando há t horas por meio da expressão (4.13).

(ii) Fixando um valor p_0 relativamente alto para a probabilidade condicional de falha no intervalo Δt, decidimos sugerir manutenção preventiva (MP) quando o valor predito por (4.13) for maior que p_0 e não sugeri-la (NMP) em caso contrário.

(iii) Com base nos dados de 2006, consideramos como decisões corretas a. indicar MP em poços que de fato falharam em 2006, durante o período correspondente a Δt horas e b. não indicar MP em poços que de fato não falharam em 2006, durante o mesmo período.

(iv) Calculamos as seguintes frequências relativas de "acertos":

a. Número de poços-coluna que falharam dentre os selecionados para MP/número de MP;

b. número de poços-coluna que não falharam dentre os que não foram indicados para MP/número de poços-coluna que não foram indicados para MP e

c. número de decisões corretas/número de poços-coluna avaliados.

(v) Avaliamos os resultados para diferentes valores de Δt e p_0.

As Figuras 5.6 e 5.7 mostram os *box plots* das probabilidades condicionais de falhas preditas dos poços-coluna considerando Δt =500h, 1000h, 2000h e 5000h. Em geral, notamos que essas probabilidades aumentam quando aumentamos o intervalo de predição Δt. Essas figuras, podem nos auxiliar a identificar poços-coluna com probabilidades condicionais de falhas relativamente altas. Esses poços-coluna estarão entre os recomendados para manutenção preventiva, ao menos que suas respectivas probabilidades condicionais de falha não ultrapassem o valor fixado p_0.

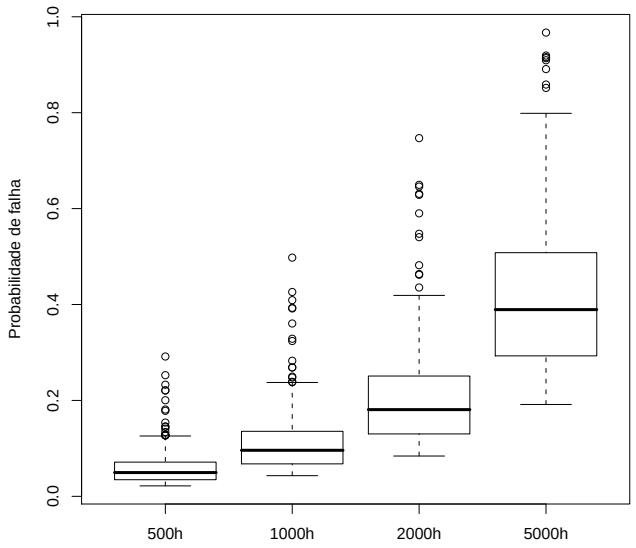

Figura 5.6: *Box plots* das probabilidades condicionais de falha - EBE.

A partir das probabilidades condicionais de falhas preditas e fixando p_0 =0,5, 0,6, 0,7, 0,8 e 0,9, identificamos os poços-coluna que apresentaram valores dessas probabilidades

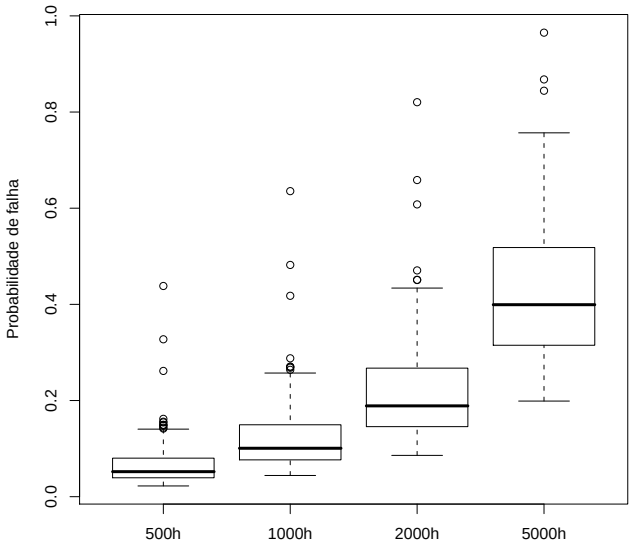

Figura 5.7: *Box plots* das probabilidades condicionais de falha - EBLUP.

acima de p_0 em cada caso. Na prática, os poços-coluna identificados seriam selecionados para manutenção preventiva.

Consideramos um "acerto" quando um poço-coluna identificado de fato falhou em 2006 ou quando um poço-coluna não identificado de fato não falhou em 2006. Um número de acertos relativamente alto implica que o modelo ajustado é adequado. A seguir, apresentamos os resultados dos percentuais de acertos para as frequências discriminadas no item (iv). Os resultados que serviram de base para os cálculos desses percentuais encontram-se no Apêndice B.

A Tabela 5.2 apresenta os percentuais de acertos em relação aos poços indicados para manutenção preventiva por Δt, p_0 e método de predição. Embora esses percentuais não sejam altos, notamos que o EBE apresenta melhores resultados que o $EBLUP$. Como exemplo de interpretação dessa tabela, o valor 60,0 significa que 60% dos poços-coluna identificados com probabilidade condicional de falha acima de 0,9 de fato falhou em 5000

Tabela 5.2: Percentual de acertos referente aos poços indicados para manutenção preventiva

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		-	-	-	-	-	-	-	-	-	-
1000		-	-	-	-	-	0	0	-	-	-
2000		37,5	40,0	0	-	-	0	0	0	0	-
5000		19,0	17,8	28,6	50,0	60,0	23,7	15,7	0	0	0

*horas.

horas. Neste caso, os 40% restantes soaram como "falso alarme". As células vazias significam que nenhum poço-coluna foi indicado, pois suas probabilidades condicionais de falha preditas em Δt não ultrapassam o p_0 correspondente. Veja, por exemplo, os *box plots* das probabilidades condicionais de falha em 500 horas. Nestes, notamos que essas probabilidades não ultrapassam o valor 0,5.

Em relação aos percentuais de acertos referente aos poços-coluna não indicados para manutenção preventiva, dispostos na Tabela 5.3, os resultados são muito satisfatórios. Em todos os casos, os percentuais ultrapassam 75% e a metade deles, ultrapassa 95%. Na prática, isso implicaria uma economia significativa dos recursos da empresa. Notamos também que os resultados para os diferentes métodos de predição são, em geral, análogos.

Para os percentuais de acertos gerais (decisões corretas), isto é, considerando os acertos referentes às manutenções preventivas (MP) e às não manutenções preventivas (MNP) conjuntamente, os resultados também são muito satisfatórios (Tabela 5.4). Esses resultados sugerem que os modelos ajustados por ambos os métodos proporcionam um bom uso dos recursos da empresa para a manutenção preventiva dos poços-coluna.

Tabela 5.3: Percentual de acertos referente aos poços-coluna não indicados para manutenção preventiva

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		98,8	98,8	98,8	98,8	98,8	98,8	98,8	98,8	98,8	98,8
1000		96,9	96,9	96,9	96,9	96,9	96,9	96,9	96,9	96,9	96,9
2000		86,5	86,3	85,8	85,9	85,9	85,8	85,8	85,8	85,8	85,9
5000		76,0	76,5	77,7	78,0	77,9	77,7	76,0	76,7	77,1	77,2

*horas.

Tabela 5.4: Percentual de decisões corretas considerando MP e MNP

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		98,8	98,8	98,8	98,8	98,8	98,8	98,8	98,8	98,8	98,8
1000		96,9	96,9	96,9	96,9	96,9	96,6	96,6	96,9	96,9	96,9
2000		85,3	85,6	85,6	85,9	85,9	85,0	85,0	85,6	85,6	85,9
5000		61,3	68,4	74,5	77,3	77,6	62,3	66,6	74,5	76,4	77,0

*horas.

Capítulo 6

Considerações Finais

6.1 Conclusão

Ao longo deste trabalho tratamos de procedimentos de predição de efeitos aleatórios em modelos lineares mistos com enfoque na abordagem clássica e bayesiana. Verificamos que para utilizar esses procedimentos em modelos de tempo de falha acelerado com efeito aleatório para tratar dados de sobrevivência correlacionados, devemos tomar alguns cuidados com respeito à presença de censuras.

Do ponto de vista da aplicação, a ideia de usar as probabilidades condicionais de falha para prever falhas futuras mostrou-se adequada, proporcionando bons resultados em relação ao emprego e economia de recursos para manutenção preventiva dos poços-coluna por parte da PETROBRAS. Com base nos percentuais de decisões corretas, concluímos que os ajustes do modelo de tempo de falha acelerado com efeito aleatório através do estimador de Bayes empírico e do EBLUP proporcionaram resultados análogos.

6.2 Trabalhos Futuros

Como motivações de continuidade deste trabalho citamos:

(i) Ajuste do modelo através do método bayesiano empírico assumindo outras combinações de distribuições para os efeitos aleatórios e variável resposta;

(ii) Estudo mais aprofundado de métodos de imputação de dados para tratar da ocorrência de censuras em uma abordagem semi-paramétrica (EBLUP);

(iii) Análise de resíduos em modelos de tempo de falha acelerado com efeito aleatório;

(iii) Comparar resultados do EBE e do EBLUP por simulação;

(iv) Estudar o método de validação “cross-validation”.

Referências Bibliográficas

Ageel M.I. (2002). A novel means of estimating quantiles for 2-parameter Weibull distribution under the right random censoring model. Journal of Computational and Applied Mathematics, **149**, 373-380.

Anderson, J. E. & Louis, T. A. (1995). Survival analysis using a scale change random effects model. *Journal of the American Statistical Association*, **90**, 669-679.

Carlin, B. P. & Louis, T. A. (1998). *Bayes and empirical Bayes methods for data analysis*. Chapman e Hall, New York.

Collett, D. (1994). *Modelling survival data in medical research*. Chapman e Hall, New York.

Colosimo, E. A. & Giolo, S. R. (2006). Análise de sobrevivência aplicada. Edgard Blüncher, São Paulo.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal Royal Statistical Society*, B, **34**, 187-220.

Dantas, M. A. (2008). Modelagem de dados de falhas de equipamentos de subsuperfície em poços de petróleo da Bacia potiguar. Dissertação de Mestrado, Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Norte, Natal, RN.

Das, K., Jiang, J. & Rao, J. N. K. (2004). Mean Squared Error of empirical predictor. *The Annals of Statistics*, **32**, 818-840.

Davison, W. C. (1991). Variable metric method for minimization. *SIAM Journal on optimization*, **1**, 1-17.

Demidenko, E. (2004). Mixed Models: Theory and Applications. Wiley, New York.

Harrekll, Jr. & Frank E. (2001) Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. Springer: New York.

Henderson. C. R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics, **31**, 387-396.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387-396.

Hourgaard, P. (2000). Analysis of multivariate survival data. New York: Springer-Verlag.

Hughes, J.P. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics*, **55**, 625-629.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The statistical analysis of failure time data*. Jonh Wiley & Sons, New York.

Karunamuni, R. J. (2002). An empirical Bayes derivation of best linear unbiased predictors. *International Journal of Mathematics and Mathematical Sciences*, **31**, 703-714.

Keiding, N., Andersen, P. K. & Klein, J. P. (1997). The role of frailty models and accelerated failure time molels in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, **16**, 215-224.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795-806.

Klein, J. P., Pelz, C. & Zhang, M. (1999). Modelling random effects for censored data by a multivariate normal regression model. *Biometrics*, **55**, 497-506.

Jiang, J. (1997). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statistica Sinica*, **8**, 861-885.

Lambert, P., Collett, D., Kimber, A. & Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine*, **23**, 3177-3192.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Jonh Wiley and Sons, 2ª edição, New York.

Liu, Q. & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624-629.

Morris, C. & Christiansen, C. (1995). Fitting Weibull duration moldels with random effects. *Lifetime Data Analysis*, **1**, 347-359.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association,* **84***, 487-493.*

Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, **7**, 55-64.

Pettitt, N. A. (1986). Censored observations, repeated measures and mixed effects models: An approach using EM algorithm and normal errors. *Biometrika*, **73**, 635-643.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.R-project.org>.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, **65**, 161-172.

Rao, C. R. (1971a). Estimation of variance and covariance components - MINQUE theory. *Journal of the Multivariate Analysis*, **1**, 257-275.

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of the Multivariate Analysis*, **1**, 445-456.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. Stat. Sci., **6**, 15-51.

Santos, P. B. (2009). Estimação em modelos de tempo de falha acelerado para dados de sobrevivência correlacionados. Dissertação de Mestrado, Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN.

SAS Institute (2009). SAS/STAT user's Guide, version 9.1. Cary: SAS Institute.

Scott, D. W. (1992). Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992). Variance components. John Wiley & Sons, 2ª edição, New York.

Taga, M. F. L. (2008). Regressão linear com medidas censuradas. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP.

Valença, D. M. (2003). Teste de homogeneidade e estimação para dados de sobrevivência agrupados e com erros de medida. Tese de Doutorado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP.

Vaupel, J. W., Manton, K. G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

Apêndice A

Outros Métodos de Predição

A seguir apresentamos as expressões do melhor preditor (BP) e do melhor preditor linear (BLP) em modelos lineares mistos. O leitor é referido a Searle et al. (1992) para ver as demonstrações dos resultados.

Melhor Predição

Pode ser mostrado que minimizando (3.6) em relação à $\tilde{\mathbf{b}}$, encontramos que o melhor preditor (BP) de $\mathbf{b}$ é a média condicional de $\mathbf{b}$ dado $\mathbf{Y}$, isto é,

$$\tilde{\mathbf{b}} = E(\mathbf{b}|\mathbf{Y}). \tag{A.1}$$

Duas propriedades importantes desse resultado são que o mesmo é válido para todas as funções densidade de probabilidade $f(\mathbf{b}, \mathbf{y})$ e que o melhor preditor é não viciado, isto é, $E(\tilde{\mathbf{b}}) = E(\mathbf{b})$, que é a definição de preditor não viciado no âmbito da predição.

Dependendo da expressão de $f(\mathbf{b}, \mathbf{y})$, temos uma maior ou menor grau de dificuldade para determinar o melhor preditor de b. A normalidade, apresentada a seguir, é um caso em que determinamos explicitamente a expressão do melhor preditor.

Normalidade:

Como já mencionado, o problema de estimação do melhor preditor $\tilde{\mathbf{b}}$ demanda o conhecimento da função densidade conjunta $f(\mathbf{b}, \mathbf{y})$. Suponha que essa seja normal com parâmetros conhecidos, isto é,

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{Y} \end{bmatrix} \sim N_{n+k} \left(\begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{V} \end{bmatrix} \right).$$

Pode ser mostrado que a função densidade condicional de $\mathbf{b}$ dado $\mathbf{Y}$ é normal com média

$$\tilde{\mathbf{b}} = E(\mathbf{b}|\mathbf{Y}) = \mathbf{C}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y}),$$

que é a expressão do melhor preditor. É importante ressaltar que caso os parâmetros $\boldsymbol{\mu}_\mathbf{Y}$, $\mathbf{C}$ e $\mathbf{V}$ não sejam conhecidos o melhor preditor não existe.

Melhor Predição Linear

O melhor preditor dado em (A.1) não é necessariamente linear em $\mathbf{Y}$. Agora, vamos voltar nossa atenção para os preditores de $\mathbf{b}$ que são lineares em $\mathbf{Y}$ com a seguinte forma:

$$\tilde{\mathbf{b}} = \mathbf{a} + \mathbf{B}\mathbf{Y}, \tag{A.2}$$

para algum vetor $\mathbf{a}$ e matriz $\mathbf{B}$. Pode ser mostrado que minimizando (3.6) para $\tilde{\mathbf{b}}$ de (A.2) a fim de obter o melhor preditor linear de $\mathbf{b}$ encontramos a seguinte expressão:

$$\tilde{\mathbf{b}} = \mathbf{C}\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{Y}), \tag{A.3}$$

em que $\boldsymbol{\mu}_\mathbf{Y}$, $\mathbf{C}$ e $\mathbf{V}$ são definidas como anteriormente.

Nesse caso a determinação de $\tilde{\mathbf{b}}$ não demanda o conhecimento da forma de $f(\mathbf{b}, \mathbf{y})$, que não é necessariamente a distribuição normal, mas apenas de $\boldsymbol{\mu}_{\mathbf{Y}}$, $\mathbf{C}$ e $\mathbf{V}$. Todavia, a expressão do melhor preditor linear de $\mathbf{b}$ encontrada aqui é a mesma do melhor preditor de $\mathbf{b}$ sob a suposição de normalidade. É fácil observar também que $\tilde{\mathbf{b}}$ de (A.3) é não viciado.

Apêndice B

Tabelas

Como no conjunto de dados original houve poços-coluna que censuraram durante o período de 2000 a 2005 e outros que surgiram ou que só vieram a apresentar falhas em 2006, não é possível verificar acertos com relação a esses poços-coluna. Portanto, os seguintes resultados referem-se a um total de 326 poços-coluna avaliados.

Tabela B.1: Número de poços indicados para manutenção preventiva e que de fato falharam

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		0	0	0	0	0	0	0	0	0	0
1000		0	0	0	0	0	0	0	0	0	0
2000		3	2	0	0	0	0	0	0	0	0
5000		16	8	6	4	3	22	8	0	0	0

*horas.

Tabela B.2: Número de poços indicados para manutenção preventiva

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		0	0	0	0	0	0	0	0	0	0
1000		0	0	0	0	0	1	1	0	0	0
2000		8	5	1	0	0	3	3	1	1	0
5000		84	45	21	8	5	93	51	9	3	1

*horas.

Tabela B.3: Número de poços não indicados para manutenção preventiva e que de fato não falharam

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		322	322	322	322	322	322	322	322	322	322
1000		316	316	316	316	316	315	315	316	316	316
2000		275	277	279	280	280	277	277	279	279	280
5000		184	215	237	248	250	181	209	243	249	251

*horas.

Tabela B.4: Número de poços não indicados para manutenção preventiva

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		326	326	326	326	326	326	326	326	326	326
1000		326	326	326	326	326	325	325	326	326	326
2000		318	321	325	326	326	323	323	325	325	326
5000		242	281	305	318	321	233	275	317	323	325

*horas.

Tabela B.5: Número de decisões corretas

		EBE					EBLUP				
Δt^*	p_0 :	0,5	0,6	0,7	0,8	0,9	0,5	0,6	0,7	0,8	0,9
500		322	322	322	322	322	322	322	322	322	322
1000		316	316	316	316	316	315	315	316	316	316
2000		278	279	279	280	280	277	277	279	279	280
5000		200	223	243	252	253	203	217	243	249	251

*horas.

Apêndice C

Programas

```
#=========================================================================#
# Programa para ajustar o modelo de tempo de falha acelerado com efeito  #
# aleatório usando o SAS (versão 9.1, 2009) - EBE                         #
#=========================================================================#
data Work.BDcompleto;  #Nome do banco de dados
set Work.BD;
y=log(tempo);
run;
proc nlmixed  cov   #Pacote do SAS
    data=Work.BDcompleto;
#.........................................................................
# Estimativas iniciais dos parâmetros
#.........................................................................
parms b0 = 7.46884  b1 = -0.05021 b3 = 0.51965 b4 = 0.05872 b5 = 1.34214
b6 = 0.96126 b7 = 1.82455 b8 = 0.00216 b9 = 0.03204 b10 = 0.01859
```

```
b11 = -0.07772 b13 = -0.00209 b14 = -0.00268 b15 = -0.00271 sigma=1.3365
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;
#..............
#Preditor linear
#..............
    eta = b + b0 + 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;
    s = (y-eta)/sigma; #Erro aleatório
```

```
#.....................................................
#Função log-verossimilhança em (4.3) e sob o modelo Weibull
#.....................................................
    logp = delta*(s-log(sigma))- exp(s);
    model y ~ general(logp);
    random b ~ normal(0,teta) subject=poco OUT=RE;  #Efeitos aleatórios
    run;

#=========================================================================#
# Programa para ajustar o modelo de tempo de falha acelerado com efeito  #
# aleatório usando o R (versão 2.8.1, 2008) - EBLUP                       #
#=========================================================================#
dados<-read.table("BDcompleto.csv",sep=";",dec=",",h=T)
attach(dados)
require(survival)
#Matriz X
X=survreg(Surv(tempo,delta)~prod+met+idade+unid+profb+unid*prod+unid*profb,
data=dados,dist= "weibull",x = T)$x
#Matriz Z
n=table(poco)
M=NULL
for (i in 1:615) M=c(M,rep(1,n[i]),rep(0,2374))
M=c(M,rep(1,n[616]))
Z=matrix(M,2374,616)
#Matriz diagonal de ordem 2374
In=diag(rep(1,2374))
```

```
#Matriz de variâncias e covariâncias de Y (V)
theta=0.3587  #Componentes de variância estimados através do I-MINQUE
sigma=1.1592
V=theta*Z%*%t(Z)+sigma^2*In   #t(Z) é a transposta de Z
# EBLUE dos betas
y=log(tempo)
EBLUE=(solve(t(X)%*%solve(V)%*%X))%*%t(X)%*%solve(V)%*%y
#t(X) é a transposta de X
#solve(V) é a inversa de V
var=solve(t(X)%*%solve(V)%*%X) #Matriz de variâncias e covariâncias do EBLUE
EP=sqrt(diag(var))             # Erro padrão do EBLUE (EP)
# EBLUP de b (efeitos aleatórios)
C=theta*t(Z) #Covariância de b e Y
EBLUP=C%*%solve(V)%*%(y-X%*%EBLUE)

#=========================================================================#
# Programa para encontrar o I-MINQUE dos componentes de variância         #
#=========================================================================#
#Estimativas iniciais (‘‘sementes’’)
theta0=0
sigma0=1.3365
sigma^2_e=(pi^2/6)sigma0^2
#Matriz V estimada (Vo)
Vo=theta0*Z%*%t(Z)+(pi^2/6)sigma^2_e*In  #(pi^2/6): variância da VEP
# Matriz Po
Po=solve(Vo)-solve(Vo)%*%X%*%(solve(t(X)%*%solve(Vo)%*%X))%*%t(X)%*%solve(Vo)
```

```
# Soma de quadrados de todos os elementos das respectivas matrizes
s_oo=sum(Po^2)
s_1o=sum((Po%*%Z)^2)
s_11=sum((t(Z)%*%Po%*%Z)^2)
u_o=sum((Po%*%y)^2)
u_1=sum((t(Z)%*%Po%*%y)^2)
#MINQUE
theta=(s_oo*u_1-s_1o*u_o)/(s_oo*s_11-s_1o^2)
sigma2_e=(u_o-s_1o*theta)/s_oo

#Observação: para determinar o I-MINQUE utilizamos as estimativas do
MINQUE obtida em cada iteração como sementes. O processo iterativo
termina quando a diferença entre duas estimativas consecutivas de
cada um dos componentes de variância não ultrapassam o valor de 0,0001.

#=============================================================================#
# Programas para calcular os riscos e as probabilidades condicionais de falha #
#=============================================================================#
#Função de risco da Weibull
h=function(t){gamma*exp(-gamma*(b+X%*%EBLUE))*t^(gamma-1)}
#Função de sobrevivência da Weibull
S=function(t){exp(-exp(-gamma*(b+X%*%EBLUE))*t^gamma)}
#Probabilidades
Probs=function(\Delta t){(S(t)-S(t+\Delta t))/S(t)}
```

Livros Grátis

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

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

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