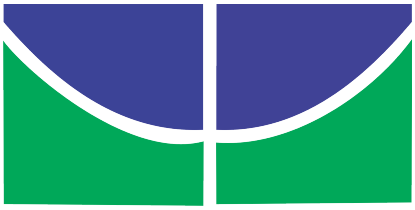

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Elisângela Candeias Biazatti

Modelo de Regressão Log Weibull com fração de cura para dados grupados

Brasília
2017

Elisângela Candeias Biazatti

Modelo de Regressão Log Weibull com fração de cura para dados grupados

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Estatística.

Orientador:
Prof. Dr. Antonio Eduardo Gomes

Coorientadora:
Prof[a]. Dr[a]. Juliana Betini Fachini Gomes

Brasília
2017

Dedicatória

Aos meus pais,
Geraldo *e* ***Sirlene****.*

À minha irmã,
Suzana*.*

Com amor e carinho, dedico.

Agradecimentos

Agradeço primeiramente a DEUS, de onde provém todas as coisas, pela saúde e todas as bençãos em minha vida.

Agradeço aos meus pais, Geraldo e Sirlene e à minha irmã Suzana, pelas orações, apoio incondicional, por todo amor, pelas palavras de força e incentivo. Obrigada por compreenderem minha ausência nos momentos bons e nos momentos difíceis ocorridos durante o período do mestrado.

Aos professores Antonio Eduardo Gomes e Juliana Betini Fachini Gomes, pela orientação, paciência, incentivo, sugestões, contribuição na minha formação e por terem acreditado e confiado em mim para realização deste trabalho.

Aos demais professores do mestrado, pelos ensinamentos e por contribuírem na minha formação.

Aos amigos e aos colegas de estudo do mestrado, pela paciência, companheirismo e pelos conhecimentos compartilhados. Um agradecimento com muito prezar aos amigos Damião e Leandro.

Ao Programa de Pós-Graduação em Estatística - PGEST/UnB - pela oportunidade de realização do meu mestrado.

A todas as pessoas que de alguma forma contribuíram para que esse objetivo fosse alcançado.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Neste trabalho é proposto um modelo de regressão com fração de cura para dados grupados utilizando a distribuição Weibull, que pode ser usada para modelar dados de sobrevivência quando a função de risco tem formas: constante, monotonicamente crescente ou monotonicamente decrescente. O modelo de regressão proposto é indicado para casos em que há no estudo, indivíduos que não apresentam a possibilidade de ocorrência do evento de interesse, indicando a presença de indivíduos curados no estudo. E também em situações em que observa-se um número excessivo de observações empatadas, para corrigir esses empates os dados são grupados em intervalos, ou quando a variável resposta é observada em intervalos de tempo, sendo esses intervalos iguais para todas as unidades amostrais. Dessa forma, os dados grupados são um caso particular de dados de censura intervalar. Um conjunto de dados reais foi utilizado para ilustração do modelo proposto. As estimativas dos parâmetros do modelo foram obtidas pelo método de máxima verossimilhança. Para detectar possíveis observações influentes foi realizada uma análise de sensibilidade no modelo proposto. Toda a análise foi desenvolvida no *software* R.

Palavras-chave: Modelos de regressão; Distribuição log Weibull; Dados de sobrevivência grupados; Fração de cura; Análise de sensibilidade.

Abstract

In this work we propose a regression model with a cure fraction for grouped data using the Weibull distribution, which can be used to model survival data when the risk function is constant, monotonically increasing or monotonically decreasing. The proposed regression model is indicated for cases in which there are individuals who do not present the possibility of occurrence of the event of interest, indicating the presence of individuals cured in the study. Also, in situations where an excessive number of ties observations is observed, the data are grouped in intervals to correct these draws, or when the response variable is observed in time intervals, these intervals being equal for all sample units. In this way, grouped data is a particular case of interval censored data. A set of real data was used to illustrate the proposed model. The estimates of the model parameters were obtained by the maximum likelihood method. In order to detect possible influential observations, a sensitivity analysis was performed in the proposed model. All the analysis was developed in *software* R.

Keywords: Regression models; log Weibull distribution; Grouped survival data; Fraction of cure; Sensitivity analysis.

Capítulo 1

Introdução

A análise de sobrevivência é uma área da estatística com elevada importância em aplicações na biologia, engenharia, medicina, entre outros, pois a variável resposta, nesse tipo de estudo, geralmente é, o tempo decorrido até que ocorra um evento de interesse. Por exemplo, na medicina, esse tempo, chamado de tempo de falha, pode ser o tempo até a morte do paciente ou até a cura de uma doença. Na engenharia, pode ser o tempo até a quebra de um equipamento. Enquanto que na área financeira, o tempo de falha pode ser o tempo que uma pessoa leva para se tornar inadimplente, entre outros. Ressalta-se ainda que, os engenheiros denominam esta área como análise de confiabilidade (Colosimo e Giolo, 2006).

A principal característica dos dados de sobrevivência é a presença de observações incompletas ou parciais, denominadas censuras. A ocorrência dessas observações tem diversos motivos, por exemplo, quando o acompanhamento do indivíduo foi interrompido por causa aleatória ao estudo, o estudo terminou e o indivíduo não apresentou o evento de interesse, ou quando se sabe que o evento de interesse ocorreu em um intervalo de tempo, mas não se sabe o tempo exato. Dessa forma, os mecanismos de censura são, censura à direita, à esquerda e intervalar.

A censura intervalar é o tipo mais geral de censura e se caracteriza quando sabe-se somente que o evento de interesse ocorreu em um determinado intervalo de tempo, mas não se sabe o tempo exato de ocorrência da falha. Isso ocorre em estudos em que os indivíduos são acompanhados em visitas periódicas e é conhecido somente que a falha ocorreu em um certo intervalo de tempo, mas não seu tempo exato.

Um caso particular de dados de censura intervalar são os dados grupados. Essa situação particular acontece quando todas as unidades amostrais são avaliadas nos mesmos instantes (Colosimo e Giolo, 2006). Várias abordagens já foram propostas para esse tipo de dados. Um estudo sobre o modelo de regressão log-Burr XII para dados grupados foi realizado por Hashimoto et al. (2012) e um modelo de regressão log-Beta Burr III para dados grupados foi proposto por Resende (2017).

Além disso, nos estudos de sobrevivência, geralmente, a probabilidade de sobrevivên-

cia diminui conforme o tempo aumenta, isto é, a probabilidade tende a zero quando o tempo tende a ser muito grande. Por outro lado, quando essa característica não ocorre é indicativo da presença de indivíduos curados, ou seja, são indivíduos não suscetíveis ao evento de interesse ou indivíduos que nunca irão falhar. Nesse contexto, os modelos com fração de cura são indicados para esses casos.

Outro ponto que está associado ao tempo de falha é a presença de covariáveis, ou seja, características dos indivíduos em estudo, podem influenciar no tempo de sobrevivência. Sendo assim, o efeito dessas covariáveis deve ser incorporado na análise estatística utilizando um modelo de regressão.

Assim, o objetivo deste estudo é formular um modelo de regressão com fração de cura para dados grupados e então, investigar se os resultados obtidos, a partir do modelo ajustado, são resistentes a pequenas perturbações. Essa etapa é chamada de análise de sensibilidade e pode ser obtida utilizando as metodologias de Influência Global e Local.

Dessa forma, o presente trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada uma revisão bibliográfica de conceitos importantes em análise de sobrevivência, na Seção 2.2.1 é apresentada a distribuição de probabilidade Weibull e na Seção 2.2.2 é apresentada a distribuição de probabilidade log Weibull e o modelo de regressão log Weibull é apresentado na Seção 2.2.3. No Capítulo 3 é proposto o modelo de regressão log Weibull com fração de cura para dados grupados e é apresentada sua função de máxima verossimilhança. No Capítulo 4 é apresentada uma análise do banco de dados de suplementação de Vitamina A (Barreto et al., 1994), na Seção 4.2 são apresentadas as estimativas para o modelo log Weibull com fração de cura para dados grupados para o conjunto de dados de Vitamina A, na Seção 4.3 são apresentadas as estimativas para o modelo considerando a inclusão de covariáveis e na Seção 4.4 é apresentada a análise de sensibilidade. E por fim, no Capítulo 5 são apresentadas as considerações finais sobre o trabalho.

Capítulo 2

Revisão de Literatura

Nesse capítulo são apresentados conceitos básicos de análise de sobrevivência, distribuições de probabilidade, introdução aos modelos de sobrevivência com fração de cura, de forma a compor um embasamento teórico para a compreensão do trabalho.

2.1 Conceitos Básicos em Análise de Sobrevivência

Em análise de sobrevivência, a variável resposta é, usualmente, o tempo até a ocorrência de um evento de interesse (Colosimo e Giolo, 2006). O tempo até a ocorrência de um evento de interesse é denominado tempo de falha, podendo ser o tempo até a morte do indivíduo devido a uma doença, por exemplo. Para definir o tempo de falha é preciso fixar o tempo de início do estudo, a escala de medida a ser utilizada e estabelecer o evento de interesse, que frequentemente é indesejável e conhecido como falha.

A principal característica de dados de sobrevivência é a observação parcial da resposta, denominada censura. Esse tipo de observação pode ocorrer por diversos motivos. Por exemplo, a perda de acompanhamento do paciente no decorrer do estudo por causa diferente ao evento de interesse. Ressalta-se que mesmo incompletas, essas observaçẽs são de extrema importância na análise estatística. É a presença de censura que diferencia a análise de sobrevivência ou análise de confiabilidade de técnicas clássicas como análise de regressão e planejamento de experimentos (Colosimo e Giolo, 2006).

Os mecanismos de censura diferenciam-se da seguinte forma: censura à direita, censura à esquerda e a censura intervalar. Quando o evento de interesse já ocorreu antes do início do estudo, temos a censura à esquerda. A censura intervalar é observada quando o evento de interesse ocorre em um intervalo conhecido de tempo, isto é, $T \in (L, U]$, porém o tempo exato de ocorrência do evento é desconhecido. A censura à direita caracteriza-se pela ocorrência do evento de interesse estar à direita do tempo de estudo, ou seja, o evento de interesse ocorre após finalizado o tempo de estudo. Esse mecanismo de censura é subdividido em três tipos: censura à direita do tipo I, censura à direita do tipo II e censura à direita aleatória.

Por exemplo, em estudos com dados com censura intervalar, em que os pacientes são acompanhados em visitas a cada dois anos e é conhecido somente que os pacientes apresentaram os sintomas em uma dessas visitas, ou seja, não se sabe o tempo exato em que ocorreu o evento de interesse, mas é conhecido que ocorreu no intervalo de tempo de dois anos.

A censura intervalar tem como casos especiais as censuras à direita, à esquerda e os tempos exatos de falha (Lindsey e Ryan, 1998). Quando $L = U$ tem-se os tempos exatos de falha, quando $U = \infty$ tem-se o caso da censura à direita e para $L = 0$ tem-se o caso da censura à esquerda (Colosimo e Giolo, 2006).

Nos dados de sobrevivência grupados, que são um caso particular de censura intervalar, os estudos são conduzidos de modo que todas as unidades amostrais são avaliadas ao mesmo tempo. Esses dados são caracterizados por um número excessivo de empates, ou seja, os tempos de vida aparecem repetidas vezes.

Ao analisar o conjunto de dados, se for detectado um número excessivo de empates, além da presença de censura, os tempos observados são grupados em um certo número de intervalos com a finalidade de eliminar os empates e considerar os intervalos como tempos de vida discreto de modo que as técnicas de análise de sobrevivência discreto possam ser utilizadas (Hashimoto, 2008). O número de intervalos no qual os tempos observados são grupados é impreciso. De acordo com Bolfarine et al. (1991) a metodologia de dados grupados se baseia na tabela de vida, sendo assim, o número de intervalos é arbitrário, considerando que em cada intervalo haja pelo menos uma falha.

Devido à presença de censura, os dados de sobrevivência para o indivíduo $i (i = 1, \ldots, n)$ sob estudo são representados, em geral, pelo par (t_i, δ_i), sendo t_i o tempo de falha ou de censura e δ_i a variável indicadora de falha ou censura, isto é,

$$\delta_i = \begin{cases} 1 & \text{se } t_i \text{ é um tempo de falha} \\ 0 & \text{se } t_i \text{ é um tempo censurado.} \end{cases}$$

Na presença de covariáveis medidas no i-ésimo indivíduo, tais como, $\mathbf{x}_i = (\text{sexo}_i, \text{idade}_i, \text{tratamento}_i)$, dentre outras, os dados ficam representados por $(t_i, \delta_i, \mathbf{x}_i)$. No caso particular de dados de sobrevivência intervalar tem-se, ainda, a representação $(l_i, u_i, \delta_i, \mathbf{x}_i)$ em que l_i e u_i são, respectivamente, os limites inferior e superior do intervalo observado para o i-ésimo indivíduo (Colosimo e Giolo, 2006).

2.1.1 Distribuições do Tempo de Sobrevivência

O tempo de falha ou tempo de sobrevivência, representado pela variável aleatória não negativa T, usualmente contínua, pode ser especificado pela sua função de densidade, função de distribuição, função de sobrevivência e por sua função de taxa de falha (ou risco).

A função de densidade de probabilidade $f(t)$ é definida como o limite da probabilidade de um indivíduo vir a falhar no intervalo de tempo $[t, t+\delta t)$ por unidade de tempo e é expressa por:

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

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

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

Uma vez que se possui a função de distribuição acumulada é possível obter a função de sobrevivência pela seguinte relação entre elas:

$$S(t) = 1 - F(t). \tag{2.1}$$

A função de sobrevivência é definida como sendo a probabilidade de um indivíduo sobreviver, ou seja, não falhar, até um certo tempo t. Essa é uma das principais funções probabilísticas utilizadas em estudos com dados de sobrevivência.

Em termos probabilísticos, a função de sobrevivência, $S(t)$, é representada por:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx,$$

e tem as seguintes propriedades: $S(t) = 1$ quando $t = 0$ e $S(t) = 0$ quando $t \to \infty$.

A função de risco ou função taxa de falha, $h(t)$, é outra função muito importante em análise de dados de sobrevivência. Essa função é muito útil para descrever a distribuição do tempo de vida de pacientes. A função de taxa de falha mostra o risco de um indivíduo falhar no tempo $t + \Delta t$, com $t \to 0$, dado que esse mesmo indivíduo sobreviveu ao tempo t, e é definida por:

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

A função taxa de falha também pode ser expressa em termos das funções de densidade e de sobrevivência pela relação:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.2}$$

Outra função útil em análise de dados de sobrevivência é a função de taxa de falha acumulada ou de risco acumulado. Essa função fornece a taxa de falha acumulada do indivíduo, ou seja, é a soma de todos os riscos em todos os tempos até o tempo t e é definida por:

$$H(t) = \int_0^t h(t)dt = -\log(S(t)).$$

A função de taxa de falha acumulada, $H(t)$, não tem uma interpretação direta, mas pode ser útil na avaliação da função de risco, $h(t)$, e pode apresentar várias formas como ilustra a Figura 2.1:

- Reta Diagonal (A) $\Longrightarrow$ uma função de risco constante é indicada.
- Curva Convexa (B) ou Côncava (C) $\Longrightarrow$ uma função de risco monotonicamente crescente ou decrescente, respectivamente, é indicada.
- Curva Convexa e depois Côncava (D) $\Longrightarrow$ uma função de risco unimodal é indicada.
- Curva Côncava e depois Convexa (E) $\Longrightarrow$ uma função de risco em forma de banheira ou **U** é indicada.

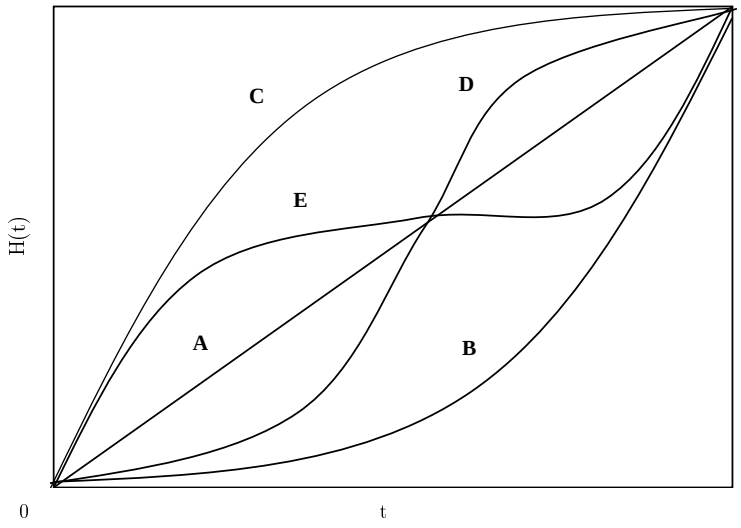

Figura 2.1: Formas que a função de taxa de falha pode assumir.

2.1.2 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier, conhecido também como estimador limite-produto é um estimador não-paramétrico proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência na presença de observações censuradas na amostra. Na construção desse estimador, a quantidade de intervalos de tempo a ser considerado são quantos forem o número total de falhas distintas.

Considere que:

- $t_1 < t_2 < ... < t_k$, os k tempos distintos e ordenados de falha,

- d_j o número de falhas em $t_j, j = 1, ..., k$, e
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é definido como:

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

As principais propriedades do estimador de Kaplan-Meier são basicamente as seguintes (Colosimo e Giolo, 2006):

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

2.1.3 Fração de Cura

Em estudos na área de análise de sobrevivência, considera-se que os indivíduos são suscetíveis ao evento de interesse. Dessa forma, ocorrerá a falha para os indivíduos que experimentarão o evento de interesse delimitado no estudo, ou em algum momento ocorrerá a censura para alguns deles.

Segundo Fachini et al. (2013), há situações que o evento de interesse (falha) não ocorrerá para uma proporção de indivíduos. Esses indivíduos são conhecidos como curados ou não-suscetíveis. Dessa forma, em pesquisas em que o interesse é investigar a recorrência de doenças em indivíduos, muitos indivíduos nunca irão experimentar o evento de interesse, pois existe uma fração de indivíduos curados no estudo.

Na base de dados os indivíduos não suscetíveis, definidos como curados, se apresentam como observações censuradas, visto que não houve falha, ou seja, o evento de interesse não foi observado.

A presença de indivíduos não suscetíveis, em um conjunto de dados, pode ser verificada por meio de um gráfico da função de sobrevivência empírica estimada pelo método de Kaplan-Meier. De acordo com Fachini et al. (2013), verifica-se no gráfico da função de sobrevivência o comportamento da cauda direita e caso ela se apresente em um nível constante acima de zero por um período de tempo, conclui-se que há evidências de indivíduos curados. Dessa forma, a função de sobrevivência é considerada imprópria quando à medida que o tempo aumenta a função não converge para zero.

Segundo Berkson e Gage (1952), quando existe indivíduos não suscetíveis no estudo necessita-se reescrever a função de sobrevivência como uma função de sobrevivência populacional. Essa função é construída na forma de mistura, isto é, dividem-se as observações em duas partes: indivíduos curados com probabilidade $(1-\phi)$ e indivíduos suscetíveis ao evento de interesse com probabilidade ϕ. A função de sobrevivência populacional em forma de mistura é representada por:

$$S_{pop}(t) = (1-\phi) + \phi S(t), \tag{2.3}$$

em que $\phi \in (0,1)$ e $S(t)$ é a função de sobrevivência própria associada à qualquer distribuição de probabilidade.

As propriedades da função são:

$$\lim_{t\to\infty} S_{pop}(t) = (1-\phi)$$

e

$$\lim_{t\to 0} S_{pop}(t) = 1.$$

Assim, a função densidade populacional e a função de risco populacional associadas a (2.3) são, respectivamente, representadas por:

$$f_{pop}(t) = \phi f(t)$$

e

$$h_{pop}(t) = \frac{\phi f(t)}{(1-\phi) + \phi S(t)},$$

em que $f(t)$ é a função densidade de probabilidade própria, ou seja, é a função densidade de probabilidade dos indivíduos suscetíveis.

2.1.4 Inferência

O método de máxima verossimilhança é um método indicado para estudos de tempo de vida. Ele é capaz de incorporar as censuras no seu processo de estimação, ou seja, a função de verossimilhança leva em conta a contribuição dos indivíduos que foram censurados e os indivíduos que falharam.

Então, considere $t_1, \ldots, t_n$ uma amostra de observações aleatórias independentes de tempos de sobrevivência da variável aleatória T contínua. Os dados consistem de n pares observados $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, em que t_i é o tempo de sobrevivência ou censura e δ_i é a variável indicadora de falha ou censura, $i = 1, 2, \ldots, n$:

$$\delta_i = \begin{cases} 1 & \text{se } T_i \quad \text{é tempo de falha;} \\ 0 & \text{se } T_i \quad \text{é tempo de censura.} \end{cases}$$

Na função de verossimilhança, a contribuição de cada observação não-censurada é a sua função de densidade $f(t)$ e a contribuição de cada observação censurada à direita é a sua função de sobrevivência $S(t)$. Dessa forma, a função de verossimilhança pode ser expressa também, em função da função de risco $h(t)$ e da função de sobrevivência $S(t)$ e é definida por:

$$L(\theta) \propto \prod_{i=1}^{n} [f(t_i, \theta)]^{\delta_i} [S(t_i, \theta)]^{1-\delta_i} = \prod_{i=1}^{n} [h(t_i, \theta)]^{\delta_i} [S(t_i, \theta)]. \tag{2.4}$$

No caso de fração de cura a equação (2.4) é expressa por:

$$L(\theta) \propto \prod_{i=1}^{n} [f_{pop}(t_i, \theta)]^{\delta_i} [S_{pop}(t_i, \theta)]^{1-\delta_i},$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos, $S_{pop}(t_i)$ e $f_{pop}(t_i)$ são as funções de sobrevivência e de densidade de probabilidade populacionais, respectivamente, para a variável aleatória T.

Por outro lado, para verificar a significância estatística do parâmetro ϕ pode-se utilizar o teste da razão de verossimilhança (TRV) definido por Maller e Zhou (1996). As hipóteses do teste são definidas por:

$$H_0 : \phi = 1$$
$$H_a : \phi < 1$$

Assim, a estatística do teste da razão de verossimilhança é escrita por:

$$TRV = 2[l(\hat{\boldsymbol{\theta}}_c) - l(\hat{\boldsymbol{\theta}}_s)],$$

em que $l(\hat{\boldsymbol{\theta}}_c)$ é o logaritmo da função de verossimilhança do modelo com fração de cura e $l(\hat{\boldsymbol{\theta}}_s)$ é o logaritmo da função de verossimilhança do modelo sem fração de cura.

Uma vez que se está testando uma hipótese na fronteira do espaço paramétrico do parâmetro, a distribuição da estatística do teste sob H_0 verdadeira não é uma distribuição de qui-quadrado como é geralmente descrita. A distribuição da estatística do teste é uma mistura 50-50 de uma variável aleatória qui-quadrado com um grau de liberdade e uma variável aleatória degenerada no ponto 0. Em outras palavras, a variável aleatória TRV tem distribuição dada por:

$$P(TRV \leq x) = \frac{1}{2} + \frac{1}{2} P(Y \leq x), x \geq 0 \,, \text{em que } Y \sim \chi^2_1. \tag{2.5}$$

Se o valor obtido do cálculo do TRV superar o valor tabelado da distribuição (2.5), rejeita-se a hipótese nula e conclui-se que o valor verdadeiro de ϕ encontra-se dentro do espaço paramétrico (0,1), ou seja, existe a presença de imunes na amostra observada.

Maiores detalhes encontram-se em Maller e Zhou (1996).

2.2 Modelos Paramétricos

Nesta seção serão apresentadas duas distribuições de probabilidade que se destacam por sua comprovada adequação quando se trata de descrever a variável tempo até a falha, na análise estatística de dados de sobrevivência.

2.2.1 Distribuição Weibull

Weibull (1939) propôs originalmente essa distribuição e também discutiu a sua ampla aplicabilidade (Weibull, 1951, 1954) em estudos biomédicos e industriais. A popularidade dessa distribuição em aplicações práticas se deve ao fato dela apresentar uma grande variedade de formas: a sua função de taxa de falha pode ser constante, monotonicamente crescente ou monotonicamente decrescente (Colosimo e Giolo, 2006).

Então, para uma variável aleatória T com distribuição de Weibull, tem-se a função de densidade de probabilidade dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left[-\left(\frac{t}{\alpha}\right)^\gamma\right], t > 0, \tag{2.6}$$

em que $\gamma > 0$ é o parâmetro de forma e $\alpha > 0$ é o parâmetro de escala. O parâmetro α tem a mesma unidade de medida de t e γ não tem unidade de medida.

A função de sobrevivência para a distribuição Weibull é dada por:

$$S(t) = \exp\left[-\left(\frac{t}{\alpha}\right)^\gamma\right], t > 0, \alpha > 0, \gamma > 0. \tag{2.7}$$

Conforme descrito na seção 2.1.1, conhecendo a função distribuição de probabilidade de uma determinada distribuição, pode-se encontrar a função de sobrevivência para essa distribuição, sendo verdadeira a recíproca. Dessa forma, utilizando a relação apresentada na equação (2.1), a função distribuição de probabilidade da distribuição Weibull é definida por:

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^\gamma\right], t > 0.$$

Ao considerar as funções definidas nas equações (2.6) e (2.7) e as inserindo na equação (2.2), define-se a função de risco da distribuição Weibull por:

$$h(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1},$$

para $t \geq 0$, α e $\gamma > 0$. Note que, quando $\gamma = 1$, tem-se a distribuição exponencial e, dessa forma, a distribuição exponencial é um caso particular da distribuição Weibull. Exemplos

do comportamento das funções de densidade, de sobrevivência e de risco da distribuição Weibull para diferentes valores dos parâmetros (γ, α), segundo Colosimo e Giolo (2006), são apresentados na Figura 2.2.

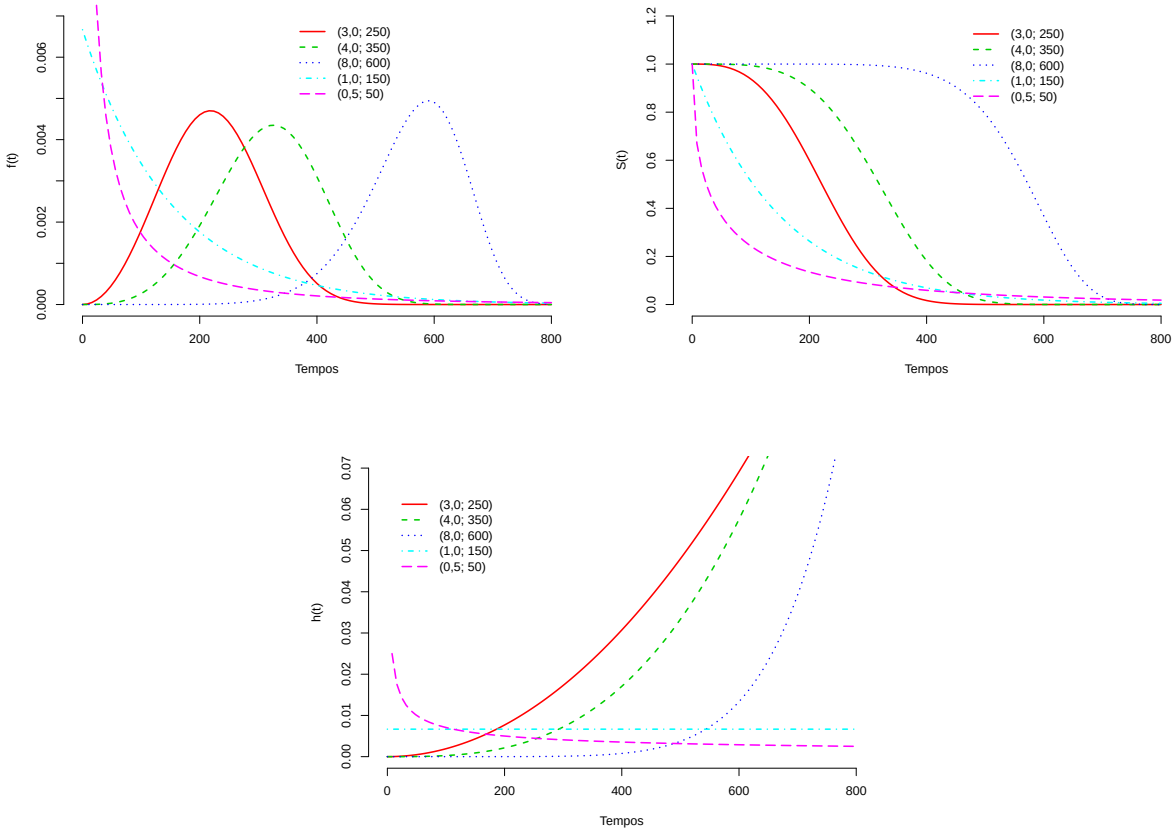

Figura 2.2: Forma típica das funções de densidade de probabilidade, de sobrevivência e de taxa de falha (risco) da distribuição Weibull para alguns valores dos parâmetros (γ, α).

Pode-se observar pela Figura 2.2, que quando $\gamma > 1$ a função de risco é estritamente crescente, quando $\gamma < 1$ a função de risco é estritamente decrescente e quando $\gamma = 1$ a função de risco é constante.

2.2.2 Distribuição Log Weibull

Nesta seção será apresentada a distribuição log Weibull que é o enfoque deste trabalho. Essa distribuição é também conhecida como distribuição do Valor Extremo ou de Gumbel e surge quando se aplica o logaritmo em uma variável aleatória com distribuição Weibull.

Seja T uma variável aleatória com função de densidade Weibull definida na equação (2.6). Considerando-se a transformação $Y = \log(T)$ e as seguintes reparametrizações $\gamma = 1/\sigma$ e $\alpha = \exp\{\mu\}$, então a variável aleatória Y tem distribuição log Weibull com função densidade de probabilidade dada por:

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

em que y e $\mu \in \mathbb{R}$ e $\sigma > 0$. Se $\mu = 0$ e $\sigma = 1$ tem-se a distribuição do Valor Extremo Padrão. Os parâmetros μ e σ são denominados parâmetros de locação e escala, respectivamente.

As funções de sobrevivência e de taxa de falha da variável aleatória Y são dadas, respectivamente, por:

$$S(y) = \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right], -\infty < y < \infty$$

e

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

em que $\sigma > 0$ e $-\infty < \mu < \infty$.

Então, com o objetivo de verificar os possíveis comportamentos da função densidade de probabilidade, foi construído o gráfico da função de densidade para a distribuição log Weibull, considerando diversos valores para seus parâmetros, μ e σ. Os gráficos encontram-se na Figura 2.3.

Nessa figura, verifica-se que a função de densidade de probabilidade da distribuição log Weibull pode acomodar dados simétricos.

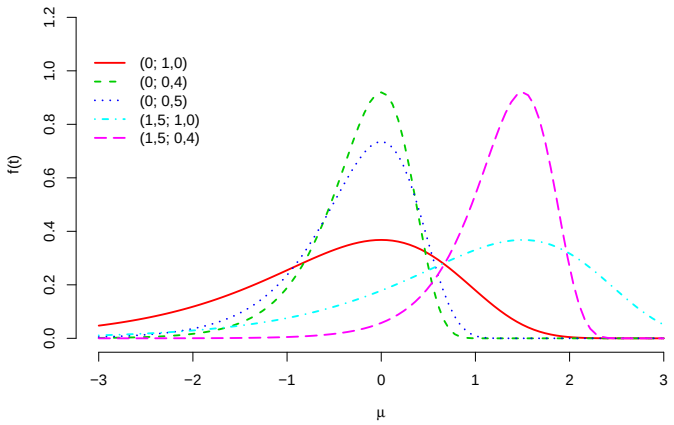

Figura 2.3: Função densidade de probabilidade da distribuição log Weibull para alguns valores dos parâmetros (μ, σ).

Por outro lado, em análise de sobrevivência, muitas vezes é conveniente trabalhar com o logaritmo dos tempos de vida observados, pois é uma forma de incluir o efeito de covariáveis e assim, obter modelos de regressão. Esse fato é explorado no modelo de regressão discutido na seção seguinte.

2.2.3 Modelo de Regressão Log Weibull

Na análise de dados de tempo de vida, muitos estudos envolvem a presença de covariáveis que podem influenciar o tempo de sobrevivência. Por exemplo, o tipo de tratamento, a idade, peso e o sexo de um paciente podem estar relacionados com o tempo de sobrevivência. Dessa forma, essas covariáveis devem estar presentes na análise estatística dos dados. Por isso, nesta seção será considerado a classe de modelos locação e escala, em que o vetor $\mathbf{x}^T = (\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p)$ de variáveis explicativas está relacionado à resposta $Y = \log(T)$ através da estrutura de regressão.

O modelo de regressão de locação e escala pode ser escrito como o modelo log-linear:

$$Y = \mu + \sigma Z, \tag{2.9}$$

em que $Z = (Y - \mu)/\sigma$ é o erro aleatório.

Ao assumir que há uma relação linear entre a variável Y e o vetor de variáveis explicativas $\mathbf{x}^T = (\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p)$, essa relação é introduzida no modelo através do parâmetro de locação ao considerar a relação $\mu = \mathbf{x}^T\boldsymbol{\beta}$. Dessa forma, o modelo de locação e escala é definido por:

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \sigma Z, \tag{2.10}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denota o vetor de coeficientes de regressão associado ao vetor $\mathbf{x}$ de variáveis explicativas, $\sigma > 0$ é o parâmetro de escala e Z é o erro aleatório com função densidade expressa na equação (2.11) que não depende de $\mathbf{x}$.

Assim, ao considerar que Y tem distribuição Valor Extremo, a função densidade e a função de sobrevivência associada ao modelo descrito em (2.9), são definidas, respectivamente, por:

$$f(y|\mathbf{x}) = \frac{1}{\sigma}\exp\left[\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right) - \exp\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right], -\infty < y < \infty$$

e

$$S(y|\mathbf{x}) = \exp\left[-\exp\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right], -\infty < y < \infty.$$

A função densidade associada ao erro do modelo é dada por:

$$f(z) = \exp[z - \exp(z)], -\infty < z < \infty. \tag{2.11}$$

2.3 Análise de Sensibilidade

Um modelo ajustado a um conjunto de dados deve representar de maneira satisfatória os dados observados, afim de evitar-se conclusões indevidas. Dessa forma, é importante

investigar se os resultados obtidos a partir do modelo proposto são afetados quando se realiza algum mecanismo de perturbação nas observações.

Então, a análise de diagnóstico deve consistir de métodos que avaliem o grau de sensibilidade das inferências a pequenas perturbações nos dados ou mesmo no modelo proposto (Fachini et al., 2013).

Nesse contexto, Cook (1977) propôs a metodologia conhecida como Influência Global ou deleção de casos. Essa metodologia avalia o impacto da retirada do i-ésimo indivíduo nos resultados do modelo ajustado.

Uma outra metodologia sugerida por Cook (1986) é a Influência Local, que avalia a influência das observações de forma conjunta, dando peso para as observações, ou seja, sob pequenas perturbações no modelo.

Essas metodologias serão descritas com maiores detalhes a seguir.

2.3.1 Influência Global

Uma importante técnica para efetuar a análise de sensibilidade é por meio da metodologia de deleção de casos. O procedimento de deleção de casos é uma abordagem para avaliar o efeito de retirar a i-ésima observação do conjunto de dados. Dessa forma, é possível observar o quanto a deleção de um caso pode influenciar alguma propriedade do modelo proposto, como as estimativas dos parâmetros e seus erros padrão.

Então, seja $\hat{\boldsymbol{\theta}}_{(i)}$ o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ obtido a partir de $l_{(i)}(\boldsymbol{\theta})$, em que $l_{(i)}(\boldsymbol{\theta})$ é o logaritmo da função de verossimilhança com a i-ésima observação deletada. Para avaliar a influência da i-ésima observação nos estimadores compara-se a diferença entre $\hat{\boldsymbol{\theta}}_{(i)}$ e $\hat{\boldsymbol{\theta}}$. Se a diferença $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ for relativamente grande, a observação é considerada influente. Dessa forma, deve ser dada maior atenção às observações que se enquadram nessa situação.

Para avaliar o efeito da i-ésima observação, duas medidas são consideradas: a distância de Cook Generalizada e o Afastamento da Verossimilhança.

A Distância de Cook Generalizada é uma medida de Influência Global, definida como a norma padronizada de $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ e é dada por:

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \boldsymbol{M} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{2.12}$$

em que $\boldsymbol{M} = -\ddot{L}(\hat{\boldsymbol{\theta}})$ ou $\boldsymbol{M} = [-\ddot{L}(\hat{\boldsymbol{\theta}})]^{-1}$ e $\ddot{L}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$.

O Afastamento da Verossimilhança é outra medida para avaliar a sensibilidade causada pela i-ésima observação e é dada por:

$$LD_i(\theta) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})\}.$$

Assim, os gráficos dessas medidas *versus* o índice das observações indicará se alguma

observação está influenciando o ajuste do modelo.

2.3.2 Influência Local

A Influência Local é outra abordagem introduzida por Cook (1986), essa metodologia propõe dar pesos para as observações em vez de removê-las, ou seja, perturbar os componentes do modelo proposto utilizando a função Afastamento da Verossimilhança.

Considerando um vetor de perturbação $\boldsymbol{w}$, de dimensão $m \times 1$ pertencente a um subconjunto aberto $\Omega \subseteq \mathbf{R}^m$ e $l(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança do modelo postulado, então, o logaritmo da função de verossimilhança do modelo perturbado é $l(\boldsymbol{\theta}|\boldsymbol{w})$. Segundo Fachini et al. (2008), a dimensão m do vetor de perturbação depende do esquema de perturbação, em geral, está relacionada com a dimensão do vetor $\boldsymbol{\theta}$ ou com o tamanho da amostra.

Alem disso, é assumido que $l(\boldsymbol{\theta}|\boldsymbol{w})$ é duas vezes continuamente diferenciável em $(\boldsymbol{\theta}, \boldsymbol{w})^T$ e que o modelo postulado está encaixado no modelo perturbado, ou seja, supõe-se que existe $\boldsymbol{w_0} \in \boldsymbol{\Omega}$ tal que $l(\boldsymbol{\theta}|\boldsymbol{w_0}) = l(\boldsymbol{\theta})$ para todo $\boldsymbol{\theta} \in \mathbf{R}^{(p+3)}$.

Assim, seja $\hat{\boldsymbol{\theta}}$ o estimador de máxima verossimilhança de $\boldsymbol{\theta}$, obtido pela maximização de $l(\boldsymbol{\theta})$ e seja $\hat{\boldsymbol{\theta}}_{(w)}$ o estimador de $\boldsymbol{\theta}$ sob $l(\boldsymbol{\theta}|\boldsymbol{w})$. Conforme Fachini et al. (2008), ao comparar $\hat{\boldsymbol{\theta}}$ com $\hat{\boldsymbol{\theta}}_{(w)}$, quando $\boldsymbol{w}$ varia em Ω, se a distância for pequena é um indicativo de que o modelo ajustado é estável no que se refere ao particular esquema de perturbação que foi utilizado.

Por outro lado, comparar $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(w)}$ diretamente, pode não ser simples em consequência de vários motivos, tais como, diferença de escala, definição de um vetor arbitrário de perturbação, unidade de medida, entre outros. Dessa forma, Cook (1986) propõe analisar o Afastamento da Verossimilhança:

$$LD(\boldsymbol{w}) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(w)})\}$$

Dessa forma, informações relevantes em relação as características importantes do modelo sob investigação podem ser obtidas através de uma análise sobre o comportamento geométrico da função $LD(\boldsymbol{w})$ quando $\boldsymbol{w}$ varia em Ω. Como uma medida de diagnóstico, Cook (1986) propõe utilizar a curvatura normal $C_{\boldsymbol{d}}$ ao redor de $\boldsymbol{w}_0$, que é o ponto onde duas verossimilhanças são iguais, em uma direção unitária $\mathbf{d}$, $\| \mathbf{d} \| = 1$ do espaço Ω. A superfície geométrica é expressa por:

$$\boldsymbol{\alpha}(\boldsymbol{w}) = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ LD_{(\boldsymbol{w})} \end{pmatrix},$$

com $\boldsymbol{w}$ variando em Ω.

Conforme Cook (1986) a curvatura normal $C_{\mathbf{d}}$ na direção $\mathbf{d}$ é dada por:

$$C_{\mathbf{d}} = 2 \parallel \mathbf{d}^T \boldsymbol{\Delta}^T \ddot{L}^{-1} \boldsymbol{\Delta} \mathbf{d} \parallel, \tag{2.13}$$

em que

$$\boldsymbol{\Delta} = \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \boldsymbol{\theta} \partial \boldsymbol{w}^T}$$

e

$$\ddot{L}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

ambas matrizes avaliadas em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$. As possíveis influências que pequenas perturbações exercem sobre os componentes do modelo, como estimativas dos parâmetros entre outras, são avaliadas através da equação (2.13).

Os elementos que exercem maior influência sobre $LD(\boldsymbol{w})$ sob pequenas perturbações podem ser revelados através de um gráfico de $\mathbf{d}_{max}$. É sugerido considerar a direção $\mathbf{d}_{max}$ correspondente à maior curvatura $C_{\mathbf{d}_{max}}$.

Então, as observações mais influentes para o esquema de perturbação considerado são identificadas pelo autovetor normalizado $\mathbf{d}_{max}$, que corresponde ao maior autovalor $C_{\mathbf{d}_{max}}$ da matriz $F = \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}(\boldsymbol{\theta})^{-1} \boldsymbol{\Delta}$. É importante saber a direção que produz a maior influência local na estimativa dos parâmetros que é dada pelo $\mathbf{d}_{max}$. O gráfico do autovetor $\mathbf{d}_{max}$ contra a ordem das observações é utilizado para identificar as observações influentes.

Por outro lado, Lesaffre e Verbeke (1998) sugerem considerar a direção do i-ésimo indivíduo que corresponderia a $\mathbf{d}_i = (0, \ldots, 1, \ldots, 0)^T$, tal que o i-ésimo elemento recebe o valor um. Dessa forma, a curvatura normal chamada influência local total do i-ésimo indivíduo, é dada por:

$$C_i = 2|\Delta_i^T \ddot{\boldsymbol{L}}(\boldsymbol{\theta})^{-1} \Delta_i|,$$

em que Δ_i é a i-ésima linha de Δ. O gráfico de C_i contra a ordem das observações pode ser usado para identificar observações influentes. Também pode ser considerado observações influentes tal que $C_i \geq 2\bar{C}$ em que $\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$.

Os três esquemas de perturbações na análise de inflência local são: ponderação de casos, perturbação de uma variável explicativa e perturbação na variável resposta.

2.3.3 Impacto das Observações Influentes

O impacto das observações influentes detectadas deve ser analisado de forma à avaliar as estimativas e sensibilidade do modelo. Para fazer esta análise, deve-se obter novas estimativas para os parâmetros do modelo considerando a retirada dessas observações influentes, individual e conjuntamente.

Assim, seja $\hat{\boldsymbol{\theta}}$ as estimativas de máxima verossimilhança do modelo obtido a partir do conjunto de dados com todas as observações e $\hat{\boldsymbol{\theta}}_{(I)}$ as estimativas de máxima verossimilhança do modelo obtido ao considerar as subamostras referentes à exclusão individualmente e em grupo dos valores possivelmente influentes.

Então, a influência de cada observação detectada pode ser quantificada pela mudança relativa do modelo, que é definida por:

$$RC = \frac{\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(I)}}{\hat{\boldsymbol{\theta}}}, \tag{2.14}$$

em que (I) se refere ao índice das observações retiradas da amostra.

Logo, o impacto das observações influentes é verificado comparando-se os resultados obtidos pela expressão (2.14). Quanto maior o valor da mudança relativa, maior é a influência exercida pela observação ou grupo de observações no modelo. É analisado também o p-valor dos coeficientes, bem como se há alteração no sinal das estimativas dos coeficientes.

Capítulo 3

Material e Métodos

3.1 Material

A base de dados fornecida pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia será utilizada para ilustrar o modelo de regressão proposto. Esses dados foram obtidos de um estudo conduzido por Barreto et al. (1994), cujo objetivo foi avaliar o efeito da suplementação de Vitamina A em episódios de diárreia recorrente em crianças pequenas.

O conjunto de dados é composto de 1207 crianças com idade entre 5 e 48 meses no início do estudo, que receberam placebo ou Vitamina A em uma pequena cidade no nordeste do Brasil entre dezembro de 1990 e dezembro de 1991. O tempo de sobrevivência para estudo foi definido como sendo o tempo (em dias) da primeira dose da Vitamina A até a ocorrência do primeiro episódio de diarréia. Um episódio de diarréia foi considerado como sendo uma sequência de dias com diarréia e um dia com diarréia foi definido quando 3 ou mais movimentos liquídos ou semiliquídos foram relatados em um período de 24 horas.

A informação sobre a ocorrência de diarréia recolhida em cada visita corresponde a um período de recordação de 48-72 horas. O número de movimentos líquidos e semilíquidos por 24 horas foi registrado. Observou-se que das 1207 crianças observadas no estudos, 925 apresentaram um episódio de diarréia e 282 apresentaram tempos de censura.

As variáveis explicativas consideradas no modelo, associadas a cada criança são: x_{i1}: idade no ínicio do estudo (em meses); x_{i2}: tratamento (0=placebo, 1=vitamina A); e x_{i3}: sexo (0=feminino, 1=masculino). Os tempos de vida foram grupados em doze intervalos $\{[4, 14), [14, 18), \ldots, [126, 185)\}$. Não há um critério que estabeleça o número de intervalos, a única suposição que se tem é a presença de pelo menos uma falha em cada intervalo, dessa forma por conveniência adotou-se $k = 12$ intervalos, com base nos intervalos definidos por Hashimoto (2008).

3.2 Métodos

Ao considerar o modelo de regressão log Weibull definido na Seção 2.2.3 e a metodologia de fração de cura definida na Seção 2.1.3, nesta Seção será proposto o modelo de regressão com fração de cura para dados grupados.

3.2.1 Especificação do Modelo de Regressão com Fração de Cura para Dados Grupados

Se o conjunto de dados observado, além da presença censura, apresentar um excessivo número de empates, os tempos observados são grupados em um certo número de intervalos com a finalidade de eliminar os empates e considerar os intervalos como tempos de vida discreto de modo que as técnicas de análise de sobrevivência discreta possam ser utilizadas (Hashimoto, 2008). Segundo Bolfarine et al. (1991) a metodologia de dados grupados se baseia na tabela de vida, sendo assim, o número de intervalos é arbitrário, de acordo com o interesse do pesquisador, sabendo que em cada intervalo haja pelo menos uma falha.

Conforme Hashimoto (2008), os intervalos são construídos de tal modo que os eixos dos tempos são divididos em k intervalos definidos por pontos de corte $a_1, \ldots, a_k, a_{k+1}$. Dessa forma, o j-ésimo intervalo é denotado por $I_j = [a_j, a_{j+1})$ para $j = 1, \ldots, k$ e os tempos de vida t_i são agrupados em k intervalos.

A estrutura de regressão é especificada em termos da probabilidade de um indivíduo sobreviver a um certo tempo condicional a sua sobrevivência ao tempo de visita anterior.

Considere que os tempos de vida são grupados em k intervalos mutuamente exclusivos, $I_j = [a_j, a_{j+1}), j = 1, \ldots k,$. A probabilidade de falha do i-ésimo indivíduo, condicionada a covariável $\mathbf{x}_i$ e ao i-ésimo indivíduo estar vivo em a_j é dada por

$$p_j(\mathbf{x}_i) = P(T_i < a_{j+1} | T_i \geq a_j, \mathbf{x}_i), \tag{3.1}$$

em termos da função de sobrevivência $S(.)$, a expressão (3.1) é escrita como:

$$p_j(\mathbf{x}_i) = P(a_j \leq T_i < a_{j+1} | T_i \geq a_j, \mathbf{x}_i) = 1 - \frac{S(a_{j+1})}{S(a_j)}.$$

Por outro lado, a probabilidade de um indivíduo não falhar no j-ésimo intervalo é definido por:

$$P(T_i \geq a_{j+1} | T_i \geq a_j) = 1 - P(T_i < a_{j+1} | T_i \geq a_j, \mathbf{x}_i) = 1 - p_j(\mathbf{x}_i),$$

em termos da função de sobrevivência, tem-se que:

$$1 - p_j(\mathbf{x}_i) = \frac{S(a_{j+1})}{S(a_j)}.$$

Para relacionar o efeito das covariáveis à variável resposta $Y = \log(T)$ utiliza-se uma estrutura de regressão baseada na distribuição log Weibull (2.8), de modo que o modelo $Y|\mathbf{x}$ pode ser representado pela expressão (2.10) e a função de sobrevivência de $Y|\mathbf{x}$ é dada por:

$$S(y|\mathbf{x}) = \exp\left[-\exp\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right], -\infty < y < \infty.$$

Ao considerar o logaritmo dos tempos de vida e assumir que eles são modelados pela distribuição log Weibull, a função de sobrevivência de $\log(a_j)|\mathbf{x}$ e $\log(a_{j+1})|\mathbf{x}$ são dadas por:

$$S[\log(a_j)|\mathbf{x}] = \exp\left[-\exp\left(\frac{\log(a_j) - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right]$$

e

$$S[\log(a_{j+1})|\mathbf{x}] = \exp\left[-\exp\left(\frac{\log(a_{j+1}) - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right], \tag{3.2}$$

respectivamente.

Como este trabalho propõe um modelo de regressão com fração de cura para dados grupados e assume distribuição log Weibull para modelar a variável resposta, considerando as funções definidas na equação (3.2) e as inserindo na função sobrevivência populacional definida na equação (2.3), a função de sobrevivência de $\log(a_j)|\mathbf{x}$ e $\log(a_{j+1})|\mathbf{x}$ são representadas por:

$$S_{pop}[\log(a_j)|\mathbf{x}] = (1-\phi) + \phi\left\{\exp\left[-\exp\left(\frac{\log(a_j) - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right]\right\}$$

e

$$S_{pop}[\log(a_{j+1})|\mathbf{x}] = (1-\phi) + \phi\left\{\exp\left[-\exp\left(\frac{\log(a_{j+1}) - \mathbf{x}^T\boldsymbol{\beta}}{\sigma}\right)\right]\right\}, \tag{3.3}$$

respectivamente. Este modelo será referido como o modelo de regressão log Weibull com fração de cura para dados de sobrevivência grupados.

3.2.2 Estimador de Máxima Verossimilhança

Seja $(y_1; \mathbf{x}_1), ..., (y_n; \mathbf{x}_n)$ uma amostra observada de n observações independentes, em que y_i representa o logaritmo do tempo de falha ou o logaritmo do tempo de censura e $\mathbf{x}_i^T = (1, \mathbf{x}_{i1}, ..., \mathbf{x}_{ip})$ é o vetor de variáveis explicativas associado ao i-ésimo indivíduo. Considerando que o logaritmo dos tempos, y_i, são grupados em k intervalos denotados por $I_j = [\log(a_j), \log(a_{j+1})]$, para $j = 1, ..., k$. A função de verossimilhança pode ser obtida considerando as variáveis explicativas $\mathbf{x}_i$ de modo que a contribuição do i-ésimo indivíduo

no j-ésimo intervalo é dado por:

- Se o i-ésimo indivíduo falhar no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $1 - S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]$.
- Se o i-ésimo indivíduo sobreviver (ou seja, estiver sob risco) no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $S[\log(a_{j+1})|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]$.
- Se o i-ésimo indivíduo for censurado no tempo c_i no j-ésimo intervalo, sua contribuição para a função de verossimilhança é dada por $S[\log(c_i)|\mathbf{x}]/S[\log(a_j)|\mathbf{x}]$, em que $\log(c_i) \in I_j$.

Assim, a função de verossimilhança para dados grupados é dado por:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{k} \left\{ \prod_{i \in F_j} [1 - S[\log(a_{j+1})|\mathbf{x}_i]/S[\log(a_j)|\mathbf{x}_i]] \right. \\ &\quad \times \prod_{i \in R_j} [S[\log(a_{j+1})|\mathbf{x}_i]/S[\log(a_j)|\mathbf{x}_i]] \\ &\quad \left. \times \prod_{i \in C_j} [S_i[\log(c_i)|\mathbf{x}]/S_i[\log(a_j)|\mathbf{x}]] \right\}, \end{aligned} \tag{3.4}$$

em que F_j denota o conjunto de indivíduos que falharam no j-ésimo intervalo, R_j denota o conjunto de indivíduos sob risco no j-ésimo intervalo, e C_j denota o conjunto dos indivíduos censurados no j-ésimo intervalo. Visto que na prática, a equação (3.4) é complicada de ser utilizada, pois os tempos de $\log(c_i)$ são desconhecidos, se os dados foram grupados em intervalos, considera-se como contribuição do i-ésimo indivíduo censurado no j-ésimo intervalo:

$$\frac{S[\log(c_i)]}{S[\log(a_j)]} \approx \left\{ \frac{S[\log(a_{j+1})]}{S[\log(a_j)]} \right\}^{1/2}. \tag{3.5}$$

Então, inserindo a equação (3.5) na equação (3.4) e considerando as funções de sobrevivência (3.2), o logaritmo da função de verossimilhança para o modelo de regressão log Weibull para dados grupados tem a seguinte forma:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{k} \left\{ \sum_{i \in F_j} \log \left\{ 1 - \left[\frac{\exp[-\exp(z_{ij+1})]}{\exp[-\exp(z_{ij})]} \right] \right\} \right. \\ &\quad + \sum_{i \in R_j} \log \left[\frac{\exp[-\exp(z_{ij+1})]}{\exp[-\exp(z_{ij})]} \right] \\ &\quad \left. + \frac{1}{2} \sum_{i \in C_j} \log \left[\frac{\exp[-\exp(z_{ij+1})]}{\exp[-\exp(z_{ij})]} \right] \right\}, \end{aligned}$$

em que $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta}^T)^T$, $z_{ij+1} = [\log(a_{j+1}) - \mathbf{x}_i^T\boldsymbol{\beta}]/\sigma$ e $z_{ij} = [\log(a_j) - \mathbf{x}_i^T\boldsymbol{\beta}]/\sigma$.

Seguindo a ideia apresentada para dados grupados e considerando as funções de sobrevivência (3.3), tem-se que a função de verossimilhança para o modelo de regressão com fração de cura para dados grupados é dada por:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{k} \left\{ \prod_{i \in F_j} [1 - S_{pop}[\log(a_{j+1})|\mathbf{x}_i]/S_{pop}[\log(a_j)|\mathbf{x}_i]] \right. \\ &\quad \times \prod_{i \in R_j} [S_{pop}[\log(a_{j+1})|\mathbf{x}_i]/S_{pop}[\log(a_j)|\mathbf{x}_i]] \\ &\quad \left. \times \prod_{i \in C_j} [S_{pop_i}[\log(c_i)|\mathbf{x}]/S_{pop_i}[\log(a_j)|\mathbf{x}]] \right\}. \end{aligned}$$

Assim, o logaritmo da função de verossimilhança para o modelo de regressão log Weibull com fração de cura para dados grupados é representada por:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{j=1}^{k} \left\{ \sum_{i \in F_j} \log \left\{ 1 - \left[\frac{(1-\phi) + \phi[\exp[-\exp(z_{ij+1})]]}{(1-\phi) + \phi[\exp[-\exp(z_{ij})]]} \right] \right\} \right. \\ &\quad + \sum_{i \in R_j} \log \left[\frac{(1-\phi) + \phi[\exp[-\exp(z_{ij+1})]]}{(1-\phi) + \phi[\exp[-\exp(z_{ij})]]} \right] \\ &\quad \left. + \frac{1}{2} \sum_{i \in C_j} \log \left[\frac{(1-\phi) + \phi[\exp[-\exp(z_{ij+1})]]}{(1-\phi) + \phi[\exp[-\exp(z_{ij})]]} \right] \right\}, \end{aligned} \tag{3.6}$$

em que $z_{ij+1} = [\log(a_{j+1}) - \mathbf{x}_i^T\boldsymbol{\beta}]/\sigma$ e $z_{ij} = [\log(a_j) - \mathbf{x}_i^T\boldsymbol{\beta}]/\sigma$.

O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ para o vetor de parâmetros $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^T)^T$, do modelo de regressão log Weibull com fração de cura para dados grupados é encontrado resolvendo-se o seguinte sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \theta} = 0$$

e pode ser obtido numericamente maximizando o logaritmo da função de verossimilhança definida em (3.6). A implementação do modelo, estimação dos parâmetros e inferência será realizada utilizando o *software* R (R Develepment Core Team, 2017).

Então, sob certas condições de regularidade, o vetor de parâmetros $\hat{\boldsymbol{\theta}}$ tem distribuição assintótica normal multivariada, com média $\boldsymbol{\theta}$ e matriz de variância e covariância $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$, ou seja,

$$\hat{\boldsymbol{\theta}} \sim N_{(p+3)}(\boldsymbol{\theta}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}),$$

em que $\boldsymbol{I}(\boldsymbol{\theta}) = E[\ddot{\boldsymbol{L}}(\boldsymbol{\theta})]$, tal que

$$\ddot{L}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

Nesse contexto, a matriz de informação de Fisher $\boldsymbol{I}(\boldsymbol{\theta})$, não pode ser obtida devido a presença de observações censuradas, alternativamente pode-se utilizar a matriz de informação observada $[\ddot{\boldsymbol{L}}(\boldsymbol{\theta})]$ avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Os elementos da matriz $\ddot{\boldsymbol{L}}(\boldsymbol{\theta})$ são:

$$\ddot{\boldsymbol{L}}(\boldsymbol{\theta}) = \begin{pmatrix} \ddot{\boldsymbol{L}}_{\phi\phi} & \ddot{\boldsymbol{L}}_{\phi\sigma} & \ddot{\boldsymbol{L}}_{\phi\beta} \\ . & \ddot{\boldsymbol{L}}_{\sigma\sigma} & \ddot{\boldsymbol{L}}_{\sigma\beta} \\ . & . & \ddot{\boldsymbol{L}}_{\beta\beta} \end{pmatrix},$$

sendo cada submatriz obtida numericamente para o modelo de regressão com fração de cura para dados grupados.

3.2.3 Influência Global

A análise de influência global definida na Seção 2.3.1 pode ser obtida para o modelo de regressão log Weibull com fração de cura para dados grupados por meio do cálculo da Distância de Cook Generalizada expressa em (2.12), em que $\hat{\boldsymbol{\theta}}_{(i)}$ é o vetor de estimativas sem a i-ésima observação e $\hat{\boldsymbol{\theta}} = (\hat{\phi}, \hat{\sigma}, \hat{\boldsymbol{\beta}}^T)^T$ é o vetor de estimativas com todas as observações. Em que $\hat{\boldsymbol{\theta}}$ é obtido ao maximizar $l(\hat{\boldsymbol{\theta}})$ definido em (3.6) e $\hat{\boldsymbol{\theta}}_{(i)}$ é obtido ao maximizar $l(\hat{\boldsymbol{\theta}}_{(i)})$ definida por:

$$\begin{aligned} l(\boldsymbol{\theta}_{(i)}) \;=\; & \sum_{j=1}^{k} \left\{ \sum_{i \in F_j} \log \left\{ 1 - \left[\frac{(1-\phi) + \phi[\exp[-\exp(z^*_{ij+1})]]}{(1-\phi) + \phi[\exp[-\exp(z^*_{ij})]]} \right] \right\} \right. \\ & + \sum_{i \in R_j} \log \left[\frac{(1-\phi) + \phi[\exp[-\exp(z^*_{ij+1})]]}{(1-\phi) + \phi[\exp[-\exp(z^*_{ij})]]} \right] \\ & \left. + \frac{1}{2} \sum_{i \in C_j} \log \left[\frac{(1-\phi) + \phi[\exp[-\exp(z^*_{ij+1})]]}{(1-\phi) + \phi[\exp[-\exp(z^*_{ij})]]} \right] \right\} \end{aligned}$$

em que $z^*_{(i)j+1} = [\log(a_{j+1}) - \mathbf{x}^T_{(i)}\boldsymbol{\beta}]/\sigma$ e $z^*_{(i)j} = [\log(a_j) - \mathbf{x}^T_{(i)}\boldsymbol{\beta}]/\sigma$.

Capítulo 4

Resultados e Discussões

4.1 Análise Descritiva

Uma análise descritiva dos dados referente ao estudo realizado para avaliar o efeito da suplementação de Vitamina A em episódios de diarréia será apresentada nesta seção. Na Figura 4.1 é apresentado um histograma construído para se ter uma ideia sobre o comportamento dos dados. Na construção do histograma considerou-se os tempos de falha e de censura.

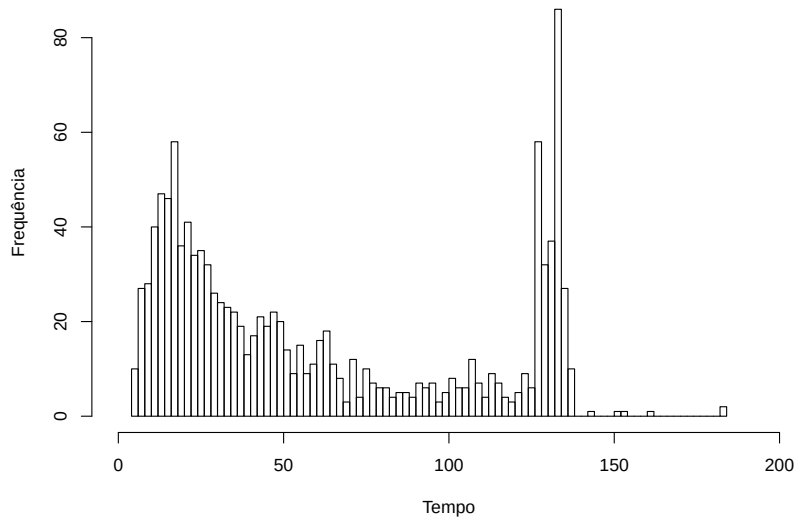

Figura 4.1: Histograma dos dados de Vitamina A

A Figura 4.1 mostra uma alta frequência de observações nos tempos iniciais do estudo e próximo ao final do estudo. Também pode ser visto a partir da Figura 4.1 que os valores para a variável resposta estão entre 4 e 184. Esse fato pode ser indicativo de que existe um número significativo de empates, fato observado também nos resultados da Tabela 4.1. Por esta razão um modelo de regressão para dados grupados é uma alternativa

interessante para modelar os dados de Vitamina A. Os tempos de vida foram grupados em doze intervalos $\{[4,14),[14,18),\ldots,[126,185)\}$. A Tabela 4.1 apresenta os intervalos de tempo de vida e os números de falhas, censura e indivíduos em risco em cada um dos doze intervalos.

Tabela 4.1: Descrição dos tempos de vida para os dados de Vitamina A desconsiderando-se as covariáveis.

Intervalo I_j	Número de falhas	Número de censuras	Número sob risco
$[4,14)$	124	0	1207
$[14,18)$	106	0	1083
$[18,23)$	103	0	977
$[23,29)$	100	1	874
$[29,37)$	92	3	773
$[37,48)$	92	6	678
$[48,61)$	90	1	580
$[61,73)$	67	1	489
$[73,90)$	46	3	421
$[90,108)$	49	6	372
$[108,126)$	46	11	317
$[126,185)$	10	250	260

A função taxa de falha acumulada é útil para auxiliar na identificação do modelo mais apropriado, conforme visto na Seção 2.1.1. A função taxa de falha acumulada para os dados de Vitamina A, que foi construída considerando os dados sem serem grupados é apresentada na Figura 4.2 e indica uma função de risco decrescente.

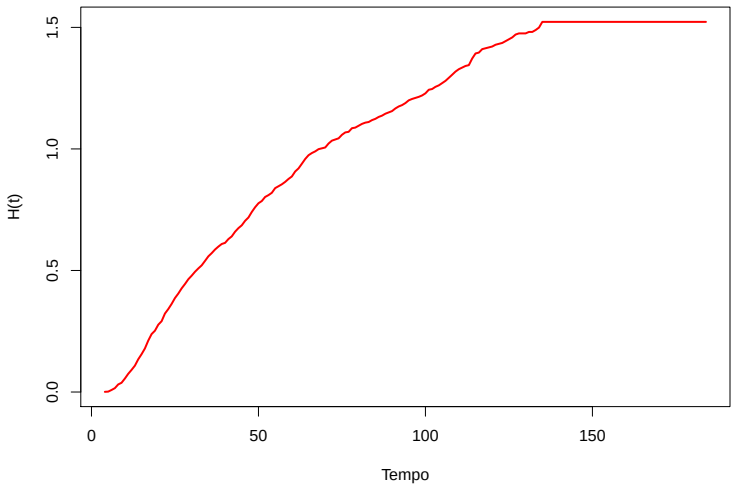

Figura 4.2: Função taxa de falha acumulada para os dados de Vitamina A

Na Figura 4.3 é apresentada a função de sobrevivência para os dados de Vitamina A, obtida pelo estimador não-paramétrico de Kaplan-Meier. Observa-se que a função de sobrevivência decresce ao longo do tempo e verifica-se uma presença maior de censuras

nos tempos finais do estudo, o que pode ser uma explicação para a alta frequência de observações nesses tempos. Observa-se também que a cauda à direita apresenta-se em um nível constante acima de zero por um período considerado suficientemente grande o que caracteriza a existência de uma proporção de indivíduos curados.

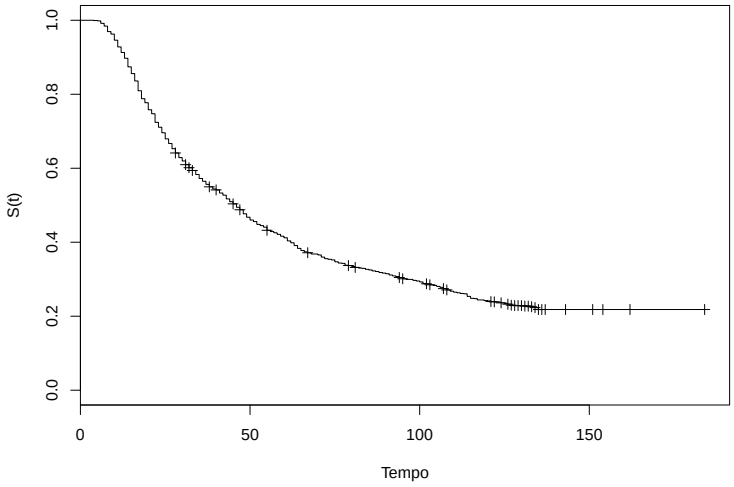

Figura 4.3: Função de sobrevivência estimada por Kaplan-Meier

Na Figura 4.4 é apresentada a função de sobrevivência estimada por Kaplan-Meier para a variável tratamento. Ao analisar a Figura 4.4, visualmente não é possível concluir se existe diferença significativa entre os tipos de tratamentos recebidos pelos indivíduos estudados.

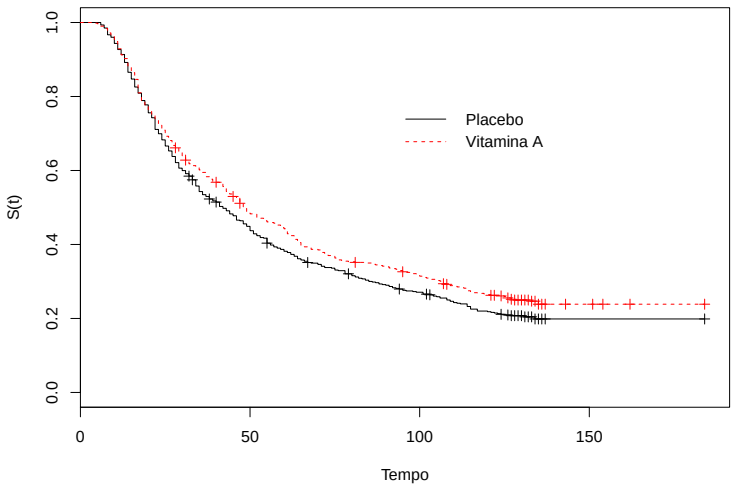

Figura 4.4: Função de sobrevivência estimada por Kaplan-Meier para covariável x_{i2}:tratamento

As curvas de sobrevivência estimadas pelo estimador de Kaplan-Meier para a covariável

sexo dos indivíduos em estudo estão representadas na Figura 4.5. Através dessa figura é possível observar que não existe diferença significativa entre as curvas de sobrevivência do sexo masculino e feminino.

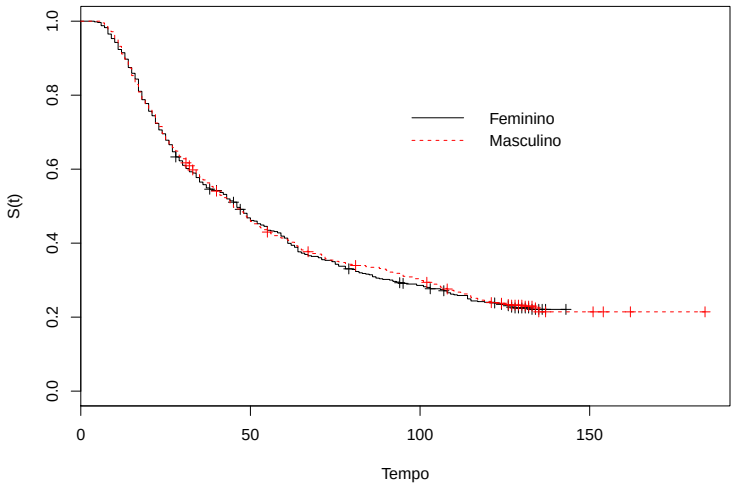

Figura 4.5: Função de sobrevivência estimada por Kaplan-Meier para covariável x_{i3}:sexo

Ainda como parte da análise descritiva dos dados, será utilizado o teste não-paramétrico de Wilcoxon para comparação das curvas de sobrevivência. As hipóteses desse teste são definidas por:

$$H_0 : S_1 = S_2$$
$$H_a : S_1 \neq S_2$$

em que S_1 é a função de sobrevivência vinculada a categoria 1 da covariável e S_2 é a função de sobrevivência vinculada a categoria 2 da covariável.

Os resultados do teste para as covariáveis são apresentados na Tabela 4.2. A variável x_{i3}: sexo (0=feminino, 1=masculino) não é significativa pelo teste de Wilcoxon aos níveis de significância de 5% e 10%. Isto é, não há diferença entre as curvas de sobrevivência do sexo feminino e do sexo masculino. A variável x_{i2}: tratamento (0=placebo, 1=vitamina A) é significativa ao nível de significância de 10% pelo teste de Wilcoxon. Isto é, existe diferença significativa entre o tratamento placebo e Vitamina A. A covariável x_{i1}: idade não é categórica, portanto não foi avaliada nesta etapa da análise.

Tabela 4.2: Resultados do teste de Wilcoxon para a comparação das categorias das covariáveis

Covariável	Estatística de teste	Valor p
Tratamento	3,0	0,0819
Sexo	0	0,835

4.2 Modelo Log Weibull com Fração de Cura para Dados Grupados

Com o objetivo de verificar o efeito da cura no modelo, nesta seção são apresentadas as estimativas dos parâmetros, seus respectivos erros-padrão, bem como o ajuste para os modelos log Weibull com e sem fração de cura para dados grupados.

As estimativas e os erros-padrão do vetor de parâmetros $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^T)^T$ estão presentes na Tabela 4.3.

Tabela 4.3: Estimativas de máxima verossimilhança para os parâmetros do modelo log Weibull para dados grupados e do modelo log Weibull com fração de cura para dados grupados

Parâmetros	Sem Fração de Cura		Com Fração de Cura	
	Estimativas	Erro Padrão	Estimativas	Erro Padrão
μ	4,235130	0,05774009	3,8178265	0,03435705
σ	1,447764	0,07102306	0,8480850	0,03447603
ϕ	-	-	0,8005633	0,01276915

O gráfico da função de sobrevivência empírica, da função de sobrevivência estimada pelo ajuste do modelo log Weibull para dados grupados e da função de sobrevivência estimada pelo ajuste do modelo log Weibull com fração de cura para dados grupados foi feito com o objetivo de verificar a qualidade do ajuste do modelo.

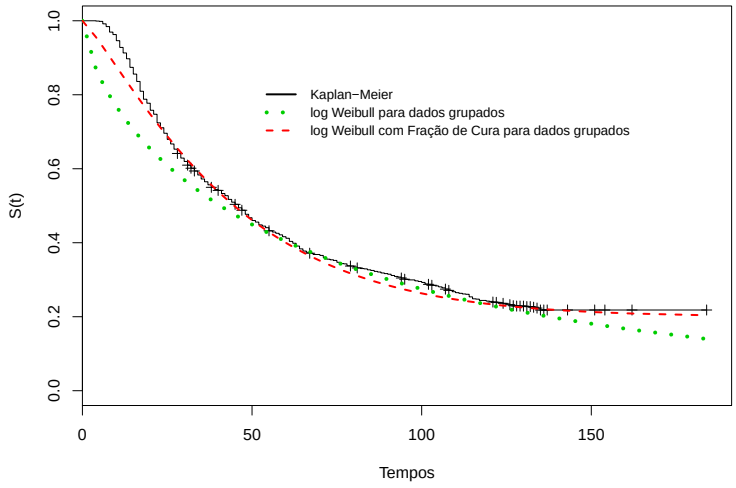

Figura 4.6: Estimativa da função de sobrevivência para o modelo log Weibull para dados grupados, log Weibull com fração de cura para dados grupados e Kaplan-Meier, para os dados de Vitamina A.

Pode-se dizer pelo gráfico apresentado na Figura 4.6 que o modelo log Weibull com fração de cura para dados grupados apresenta um bom ajuste.

Para avaliar a significância estatística do parâmetro ϕ foi aplicado o Teste de Razão de Verossiminhaça (TRV) como um teste paramétrico para a presença de imunes, como apresentado na Subseção 2.1.4.

Sob a hipótese $H_0 : \phi = 1$, e considerando a expressão (2.5), tem-se:

$$\frac{1}{2} + \frac{1}{2}P(Y \leq c_{0,95}) = 0,95\,, \text{em que}\, Y \sim \chi^2_1.$$

Assim $c_{0,95}$ satisfaz $P(Y \leq c_{0,95}) = 0,90$. Da tabela de χ^2_1 obtemos $c_{0,95} = 2,71$.

O valor obtido do cálculo do TRV é $109,9893$. Como $109,9893 > 2,71$ rejeita-se H_0 ao nível de 5% de significância e considera-se que há forte evidência de $\phi < 1$ para estes dados, ou seja, o valor verdadeiro de ϕ encontra-se dentro do espaço paramétrico (0,1) e sugere fortemente a presença de imunes. Com base nos resultados encontrados, na próxima seção é apresentado o ajuste do modelo de regressão log Weibull com fração de cura para os dados de Vitamina A.

4.3 Modelo de Regressão Log Weibull com Fração de Cura para Dados Grupados

Afim de avaliar o efeito das covariáveis conjuntamente ao efeito da cura, nesta seção são analisados esses fatores considerando os modelos de regressão log Weibull com e sem fração de cura para dados grupados.

Na Tabela 4.4 são apresentadas as estimativas dos parâmetros do modelo de regressão log Weibull para dados grupados e do modelo de regressão log Weibull com fração de cura para dados grupados.

Tabela 4.4: Estimativas de máxima verossimilhança para os parâmetros do modelo de regressão log Weibull para dados grupados e do modelo de regressão log Weibull com fração de cura para dados grupados

Parâmetros	Sem Fração de Cura			Com Fração de Cura		
	Estimativas	Erro Padrão	p-valor	Estimativas	Erro Padrão	p-valor
Intercepto	2,96797722	0,123830464	<0,0001	2,9306728	0,109532725	<0,0001
Idade	0,04428948	0,003746431	<0,0001	0,0373567	0,004395113	<0,0001
Tratamento	0,20219116	0,085208445	0,0176	0,2135306	0,078510887	0,0065
σ	1,29034778	0,058221323	-	0,9965450	0,056544623	-
ϕ	-	-	-	0,8939624	0,021959165	-

O modelo foi ajustado com as três variáveis explicativas, porém a variável x_{i3}:sexo (0=feminino, 1=masculino) não foi significativa no modelo, assim como visto na análise descritiva, por essa razão ela não foi incluída na análise com covariáveis.

O teste paramétrico para a presença de imunes também foi realizado para o modelo de regressão log Weibull com fração de cura para dados grupados e apresentou $TRV = 34,084$. Sob a hipótese $H_0 : \phi = 1$ e considerando a expressão (2.5), tem-se:

$$\frac{1}{2} + \frac{1}{2}P(Y \leq c_{0,95}) = 0,95\,, \text{em que}\, Y \sim \chi^2_1.$$

Da tabela de χ^2_1 obtêm-se $c_{0,95} = 2,71$. Como $34,084 > 2,71$ rejeita-se H_0 ao nível de 5% de significância e considera-se que há forte evidência de $\phi < 1$ para estes dados.

Então, ao considerar o modelo de regressão log Weibull com fração de cura para dados grupados, verifica-se que a variável idade (x_1) e tratamento (x_2) são significativas ao nível de 5% e 10% de significância. Dessa forma, conclui-se que ao aumentar em uma unidade a variável idade, o tempo de sobrevivência aumenta. Em relação ao tratamento, crianças que receberam suplementação de Vitamina A apresentarão maior probabilidade de sobrevivência estimada do que as crianças que receberam placebo, pois a estimativa do coeficiente para essa covariável é positiva. Esse resultado confirma o resultado encontrado na análise descritiva na Figura 4.4.

4.4 Análise de Sensibilidade

Conforme descrito na Secão 2.3 é importante verificar se existem observações influenciando o ajuste do modelo. A medida de Influência Global foi calculada como definida na Seção 2.3.1, para investigar esse fato.

4.4.1 Análise de Influência Global

Ao considerar a Distância de Cook Generalizada expressa em (2.12), a observação #980 é a que mais se destaca, como mostra a Figura 4.7, indicando que essa pode ser considerada como uma possível observação influente.

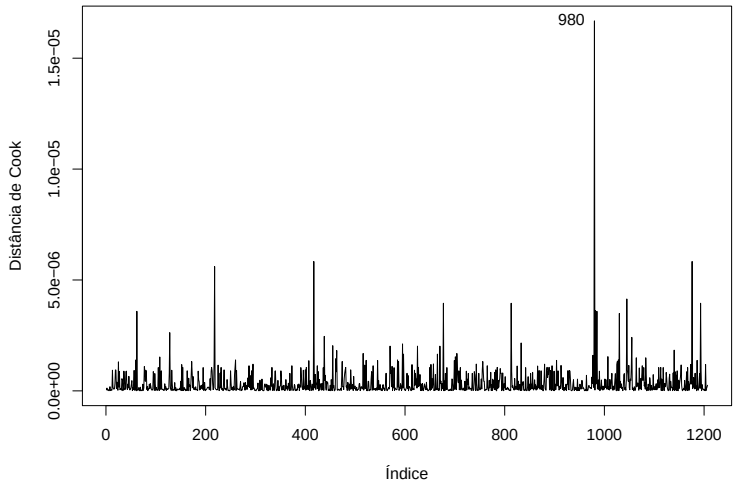

Figura 4.7: Distância de Cook

4.4.2 Análise das Observações Influentes

Após realizar a análise de sensibilidade através da metodologia de Influência Global, verificou-se que a observação #980 pode ser influente. As informações do indivíduo 980 encontram-se na Tabela 4.5.

Tabela 4.5: Resposta das variáveis explicativas do indivíduo 980.

Indivíduo (i)	T_i	δ_i	X_{i1}	X_{i2}	X_{i3}
980	83	1	5	Placebo	Feminino

Ao analisar a Tabela 4.5 pode-se dizer que o tempo de falha do indivíduo (980) é relativamente alto considerando que recebeu o tratamento placebo, visto que crianças que receberam suplementação de Vitamina A apresentarão maior probabilidade de sobrevivência estimada do que as crianças que receberam placebo. Observa-se ainda, que iniciou o estudo com 5 meses, o que corresponde aos indivíduos mais novos no início do estudo e como visto na análise descritiva o sexo da criança não exerce influência.

Para melhor investigar se esse indivíduo está influenciando o modelo, foram estimados os parâmetros a partir de uma subamostra da amostra original, que foi selecionada retirando-se a observação considerada influente. A mudança relativa do modelo (RC) definida na Seção 2.3.3, as estimativas dos parâmetros e a significância dos mesmos foram utilizadas para verificar o impacto de retirar a observação 980 da análise.

Na Tabela 4.6 encontram-se a mudança relativa de cada parâmentro, as estimativas dos parâmentros e seus respectivos p-valor:

Tabela 4.6: Mudança relativa [RC], estimativas dos parâmetros e correspondentes (p-valor)

Subamostra	ϕ	σ	β_0	β_1	β_2
	[-]	[-]	[-]	[-]	[-]
I - {completo}	0,89396	0,99655	2,93067	0,03736	0,21353
	(-)	(-)	(<0,0001)	(<0,0001)	(0,0065)
	[0,00617]	[-0,00947]	[0,00919]	[-0,00309]	[-0,09224]
I - {980}	0,888457	1,00599	2,90375	0,03747	0,23325
	(-)	(-)	(<0,0001)	(<0,0001)	(0,0034)

Ao analisar a Tabela 4.6, verifica-se que as estimativas dos parâmetros não sofreram grandes mudanças e que a significância dos coeficientes se manteve e observa-se ainda valores pequenos da mudança relativa, o que indica que o modelo de regressão log Weibull com fração de cura para dados de sobrevivência grupados proposto é robusto nesta aplicação.

Capítulo 5

Considerações Finais

Neste trabalho foi proposto o modelo de regressão log Weibull com fração de cura para dados de sobrevivência grupados. Foi utilizado para aplicação desse modelo o conjunto de dados de Vitamina A, ao analisar as covariáveis, concluiu-se que a covariável x_{i3}:sexo não é significativa. Dessa forma, as covariáveis significativas no modelo ajustado foram x_{i1}:idade e x_{i2}:tratamento.

O *software* R foi utilizado para estimação dos parâmetros e demais análises, por meio de funções implementadas e outras já existentes.

Uma análise de sensibilidade foi realizada utilizando a metodologia de Influência Global, com o intuito de verificar a presença de possíveis observações influentes. Através dos resultados da análise pôde-se concluir que o modelo se mostrou robusto na aplicação dos dados de Vitamina A.

Quantidades distintas de diferentes intervalos foram construídos observando a disposição do número de falhas, censuras e indivíduos em risco nos intervalos, porém verificou-se que não houve mudanças significativas nas estimativas dos parâmetros e seus respectivos erros padrão, bem como no ajuste do modelo.

Uma vantagem do modelo de regressão log Weibull com fração de cura para dados grupados em relação ao modelo de regressão log Burr XII para dados grupados proposto por Hashimoto et al. (2012) é a significância do coeficiente da covariável tratamento, dado que essa covariável é importante, considerando que o objetivo do estudo era avaliar o efeito da suplementação da Vitamina A. Em relação ao modelo de regressão log-Beta Burr III para dados grupados proposto por Resende (2017), a vantagem foi incorporar a fração de cura dos dados, que se mostrou significativa.

Os resultados obtidos na aplicação dos dados de Vitamina A mostram que o modelo de regressão log Weibull com fração de cura para dados grupados é adequado.

5.1 Perspectivas para Trabalhos Futuros

1. Implementar a Técnica de Influência Local para o modelo de regressão log Weibull com fração de cura para dados grupados.

2. Realizar uma Análise de Resíduos para o modelo proposto.

3. Estudar outros modelos probabilísticos para dados grupados.

Referências Bibliográficas

Barreto, M. L., L. M. P. Santos, A. M. O. Assis, M. P. N. Araújo, G. G. Farenzena, P. A. B. Santos e R. L. Fiaccone. Effect of vitamin a supplementation on diarrhoea and acute lower-respiratory-tract infections in young children in brazil. Lancet 344, 228-231. 1994.

Berkson, J. e R. P. Gage. Survival curve for cancer patients follwing treatment. Journal of the American Statistical Association, 47, 501-515. 1952.

Bolfarine, H., J. Rodrigues, e J. A. Achcar. Análise de sobrevivência. In II Escola de Modelos de Regressão. IM-UFRJ, Rio de Janeiro. 1991.

Colosimo, E. A. e S. R. Giolo. Análise de Sobrevivência Aplicada. Edgard Blucher. 2006.

Cook, R.D. Assesment of local influence (with dicussion). Journal of the Royal Statistical Society, London, v. 48, 133-169. 1986.

Cook, R.D. Detection of influential observations in linear regression. Technometrics, Alexandria, v. 19, 15-18. 1977.

Fachini, J. B., Ortega, E. M. M., Louzada Neto, F. Influence diagnostics for polyhazard models in the presence of covariates. Statistical Methods and Applications, 17, 413-433. 2008.

Fachini, J. B., Ortega, E. M. M., Cordeiro, G. M. A bivariate regression model with cure fraction. Journal of Statistical Computation and Simulationv, 1, 1-16. 2013.

Hashimoto, E. M. Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados. Dissertação (Mestrado), Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo. 2008.

Hashimoto, E. M., Ortega, E. M. M.,Cordeiro, G. M., Barreto, M. L. The log-burr XII regression model for grouped survival data. Journal of Biopharmaceutical Statistics, 22, 141-159. 2012.

Hosmer, D. W., Lemeshow, S. Applied Survival Analysis, John Wiley and Sons, New York. 1999.

Kaplan, E. L. e P. Meier. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481. 1958.

Lesaffre, E.; Verbeke, G. Local influence in linear mixed models. Biometrics, Washington, v. 54, p.570-582. 1998.

Lindsey, J. C. e L. M. Ryan. Tutorial in biostatistics: methods for intervalcensored data. Statistics in Medicine 17, 219-238. 1998.

Maller, R. A. e X. Zhou. Survival Analysis with Long Term Survivors. Wiley series in probability and statistics. 1996.

Resende, V. S. Modelo de regressão log-beta burr III para dados grupados. Dissertação de mestrado, Universidade de Brasília. 2017.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2017. Disponível em: http://www.R-project.org/

Weibull, W. A statistical theory of the strength of materials. Ingeniors Vetenskaps Akademien Handlingar n. 151: The Phenomenon of Rupture in Solid, 293-297. 1939.

Weibull, W. A statistical distribution of wide applicability. Journal of Applied Mechanics, 18, 293-297. 1951.

Weibull, W. A statistical representation of fatigue failure in solids. Royal Institute of Technology. 1954.

Apêndice A

Programa Estimação

```
dados2<-read.csv(file.choose(), header = TRUE,sep=";")
tempo<-dados2$tempo
censura<-dados2$cens

n=length(tempo)

xx0<-dados2$x0
xx1<-dados2$x1
xx2<-dados2$x2
xx3<-dados2$x3

require(survival)

y<-log(tempo)
mve1<-survreg(Surv(y,censura)~xx1 + factor(xx2), dist="extreme")

intervaloJ <-  c(4,14,14,18,18,23,23,29,29,37,37,48,48,61,
61,73,73,90,90,108,108,126,126,185)

linf <- intervaloJ[(1:12)*2-1]
lsup <- intervaloJ[-c((1:12)*2-1)]
matrizJ <- cbind(linf,lsup)

k<-12
d1<-matrix(1,k,n)
d2<-matrix(0,k,n)

for (i in 1:n)
{
for(j in 1:k)
{
if(tempo[i]>=linf[j] & tempo[i]<lsup[j]){

d1[j,i]=0
```

```
if(censura[i]==1){
d2[j,i]=1

}
else {
d2[j,i]=0
}
}
if(tempo[i]<linf[j])
{
d1[j,i]=10000
d2[j,i]=10000
}
}
}

k<-12
n<-length(tempo)
vero<-matrix(0,k,n)
unsk<-matrix(1,1,k)
unsn<-matrix(1,n,1)
zij<-mat.or.vec(n,12)
zij2<-mat.or.vec(n,12)

logv<-function(para){
phi<-para[1]
sigmaw<-para[2]
beta0<-para[3]
beta1<-para[4]
beta2<-para[5]

for(j in 1:k)
{
for (i in 1:n)
{

xbeta<-xx0*beta0+xx1*beta1+xx2*beta2

zij[i,j]<-((log(matrizJ[j,2])-xbeta[i])/sigmaw)
zij2[i,j]<-((log(matrizJ[j,1])-xbeta[i])/sigmaw)

sobrelinffc<-(1-phi)+phi*(exp(-exp(zij2[i,j])))
sobrelsupfc<-(1-phi)+phi*(exp(-exp(zij[i,j])))

if(d1[j,i]==0 && d2[j,i]==1)#Falha
{
vero[j,i]=sum(log(1-(sobrelsupfc/sobrelinffc)))
}
```

```
if(d1[j,i]==0 && d2[j,i]==0)#Censura
{
vero[j,i]=sum((1/2)*log(sobrelsupfc/sobrelinffc))
}
if(d1[j,i]==1 && d2[j,i]==0)#Risco
{
vero[j,i]=sum(log(sobrelsupfc/sobrelinffc))
}
if(d1[j,i]==10000 && d2[j,i]==10000) #Individuos que passaram
#para o proximo intervalo
{
vero[j,i]=0
}
}
}

if ((sigmaw>0) && (phi > 0) && (phi < 1))

return ((-1)*(unsk%*%vero%*%unsn))

else return (-Inf)
}

vllfct <- optim(c(0.5,mve1$scale,mve1$coefficients),logv,NULL,
hessian=T)

logVerofct<-(-1)*vllfct$value
inversaafct<-solve(vllfct$hessian)
varparfct<-diag(inversaafct)
erropadfct<-sqrt(varparfct)

Z0=vllfct$par[3]/erropadfct[3]
2*(1-pnorm(abs(Z0)))
Z1=vllfct$par[4]/erropadfct[4]
2*(1-pnorm(abs(Z1)))
Z2=vllfct$par[5]/erropadfct[5]
2*(1-pnorm(abs(Z2)))
```

Apêndice B

Programa Influência Global

```
dados2<-read.csv(file.choose(), header = TRUE,sep=";")
tempo<-dados2$tempo
censura<-dados2$cens

n=length(tempo)

xx0<-dados2$x0
xx1<-dados2$x1
xx2<-dados2$x2
xx3<-dados2$x3

library(survival)

y<-log(tempo)
mve1<-survreg(Surv(y,censura)~xx1 + factor(xx2), dist="extreme")

phicoef<- matrix(0,n,1)
sigcoef<- matrix(0,n,1)
b0coef<- matrix(0,n,1)
b1coef<- matrix(0,n,1)
b2coef<- matrix(0,n,1)

for(l in 1:n){
dad<-dados2[-l, ]
tempo<-dad[ ,1]
censura<-dad[ ,2]
xx0<-dad[ ,3]
xx1<-dad[ ,4]
xx2<-dad[ ,5]
xx3<-dad[ ,6]

n<-length(tempo)

intervaloJ <-  c(4,14,14,18,18,23,23,29,29,37,37,48,48,61,61,73,
```

```
73,90,90,108,108,126,126,185)
linf <- intervaloJ[(1:12)*2-1]

lsup <- intervaloJ[-c((1:12)*2-1)]

matrizJ <- cbind(linf,lsup)
matrizJ

k<-12
d1<-matrix(1,k,n)
d2<-matrix(0,k,n)

for (i in 1:n)
{
for(j in 1:k)
{
if(tempo[i]>=linf[j] & tempo[i]<lsup[j]){

d1[j,i]=0
if(censura[i]==1){
d2[j,i]=1
}
else {
d2[j,i]=0
}
}
if(tempo[i]<linf[j])
{
d1[j,i]=10000
d2[j,i]=10000
}
}
}

k<-12
vero<-matrix(0,k,n)
unsk<-matrix(1,1,k)
unsn<-matrix(1,n,1)
zij<-mat.or.vec(n,12)
zij2<-mat.or.vec(n,12)

logv<-function(para){
phi<-para[1]
sigmaw<-para[2]
beta0<-para[3]
beta1<-para[4]
beta2<-para[5]
```

```
for(j in 1:k)
{
for (i in 1:n)
{

xbeta<-xx0*beta0+xx1*beta1+xx2*beta2

zij[i,j]<-((log(matrizJ[j,2])-xbeta[i])/sigmaw)
zij2[i,j]<-((log(matrizJ[j,1])-xbeta[i])/sigmaw)

sobrelinffc<-(1-phi)+phi*(exp(-exp(zij2[i,j])))
sobrelsupfc<-(1-phi)+phi*(exp(-exp(zij[i,j])))

if(d1[j,i]==0 && d2[j,i]==1)#Falha
{
vero[j,i]=sum(log(1-(sobrelsupfc/sobrelinffc)))
}
if(d1[j,i]==0 && d2[j,i]==0)#Censura
{
vero[j,i]=sum((1/2)*log(sobrelsupfc/sobrelinffc))
}
if(d1[j,i]==1 && d2[j,i]==0)#Risco
{
vero[j,i]=sum(log(sobrelsupfc/sobrelinffc))
}
if(d1[j,i]==10000 && d2[j,i]==10000) #Individuos que passaram
#para o proximo intervalo
{
vero[j,i]=0
}
}
}

if ((sigmaw>0) && (phi > 0) && (phi < 1))

return ((-1)*(unsk%*%vero%*%unsn))

else return (-Inf)

}

vllfc <- optim(c(0.5,mve1$scale,mve1$coefficients),logv,NULL,
hessian=T)

phicoef[l]<- vllfc$par[1]
sigcoef[l]<- vllfc$par[2]
b0coef[l]<- vllfc$par[3]
b1coef[l]<- vllfc$par[4]
```

```
b2coef[l]<- vllfc$par[5]

}

ck<-cbind(phicoef,sigcoef,b0coef,b1coef,b2coef)

inversaafct<-solve(vllfct$hessian)

hes<-vllfct$hessian

DC<-matrix(0,n,1)
DC1<-matrix(0,n,1)

for (u in 1:n){

DC[u]<-t(ck[u,]-vllfct$par)%*%inversaafct%*%(ck[u,]-vllfct$par)

DC1[u]<-t(ck[u,]-vllfct$par)%*%hes%*%(ck[u,]-vllfct$par)

}
DC
DC1

#graficos

#DC1
indice<-c(1:n)
plot(indice,DC1,type = "l",xlab="Índice", ylab="Distância de
Cook")
identify(indice,DC1)

#DC
indice<-c(1:n)
plot(indice,DC,type = "l",xlab="Índice", ylab="Distância de
Cook")
identify(indice,DC)
```