



UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA



Introdução aos modelos de fragilidade: Uma maneira de analisar dados correlacionados de sobrevivência

Autor: Matheus Minguillo
Orientador: Professor Dr. Álvaro Vigo

Porto Alegre, 01 de Julho de 2016.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Introdução aos modelos de fragilidade:
Uma maneira de analisar dados
correlacionados de sobrevivência

Autor: Matheus Minguillo

Trabalho de Conclusão de Curso
apresentado para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Álvaro Vigo
Professora Dra. Vanessa Bielefeldt Leotti Torman

Porto Alegre, 01 de Julho de 2016.

Dedico este trabalho à minha família.

“A persistência é o caminho do êxito.”

Charles Chaplin

Agradecimentos

Ao Prof. Álvaro Vigo, pela compreensão, confiança, incentivo, transmissão de pensamentos e sabedoria presente em seus conselhos. Devo também agradecer pelo convívio desde o início da graduação e pela forma como se manteve durante todo o percurso.

Agradeço também aos excelentes profissionais que tive contato no decorrer da graduação. Em especial as professoras Stela Castro e Vanessa Torman por terem, de alguma forma, contribuído com meu crescimento acadêmico e profissional.

Aos amigos que fiz na graduação, com quem compartilhei diversos momentos felizes ao longo desses anos.

Aos colegas do TRF4 e, agora, aos colegas da MStortti.

Aos meus amigos pelo apoio e paciência de entender que muitas vezes foi preciso ficar ausente para me dedicar aos estudos.

À minha namorada, Camila May, que além de todo o apoio e carinho, que me proporciona diariamente, sempre esteve ao meu lado, acreditando e me incentivando muito.

Finalmente, dedico à minha família, cujo suporte tornou possível essa longa caminhada. Ao meu pai, Amancio, à minha mãe, Doracy, à minha sobrinha, Alicia e aos meus irmãos, Daniela, Miguel e Patricia, por terem me incentivado, me apoiando em minhas decisões e por torcerem pelo meu sucesso.

Resumo

O modelo de Cox básico pode ser limitado para um contexto em que os dados de sobrevivência são correlacionados, como ocorre frequentemente em estudos multiníveis. Isto pode ocorrer em estudos multicêntricos ou quando a unidade observacional ou experimental está agregada em um mesmo indivíduo, de modo que existem correlações entre estas unidades. Com base nesse contexto, os modelos de fragilidade são úteis quando queremos lidar com dados de sobrevivência correlacionados. Desta forma, o objetivo principal deste trabalho é revisar a literatura estatística a fim de identificar diferentes abordagens para o uso de modelos de fragilidade e explorar os algoritmos para empregar o método. Rotinas computacionais foram desenvolvidas para mostrar aspectos conceituais e técnicas para o ajuste de modelos de fragilidade utilizando os programas R e SAS. O contexto é de um estudo com acompanhamento de indivíduos ao longo do tempo, registrando-se o tempo até a ocorrência de um evento de saúde adverso. Os modelos ajustados com a inclusão de efeitos aleatórios propostos neste trabalho tornam as estimativas dos efeitos das covariáveis mais consistentes e, além disso, permitem ajustar um modelo mais versátil e que se aproxima dos valores obtidos, utilizando um banco de dados simulados. Os modelos de fragilidade podem ser avaliados pelas estimativas dos coeficientes e permitem estimar a razão de azares e o intervalo de confiança para valores pontuais das covariáveis, a partir de um valor de referência definido pelo usuário.

Sumário

1. INTRODUÇÃO	8
2. OBJETIVOS	10
3. ANÁLISE DE SOBREVIVÊNCIA	11
3.1 CENSURA E TRUNCAMENTO.....	12
3.2 FUNÇÃO DE SOBREVIVÊNCIA E FUNÇÃO DE RISCO	13
3.3 MODELOS PARAMÉTRICOS.....	15
3.3.1 MODELO DE REGRESSÃO EXPONENCIAL.....	15
3.3.2 MODELO DE REGRESSÃO LOGNORMAL	17
3.3.3 MODELO DE REGRESSÃO WEIBULL.....	18
3.4 MÉTODOS NÃO PARAMÉTRICOS.....	20
3.4.1 MODELO DE RISCOS PROPORCIONAIS DE COX	21
4. MODELOS DE FRAGILIDADE.....	24
4.1 FRAGILIDADE GAMA.....	28
4.2 FRAGILIDADE LOGNORMAL.....	29
5. EXEMPLO SIMULADO.....	32
5.1 GERAÇÃO DOS DADOS	33
5.2 ANÁLISE DESCRITIVA.....	35
5.3 MODELO DE COX	37
5.4 MODELOS PARAMÉTRICOS COM FRAGILIDADE.....	42
6. CONSIDERAÇÕES FINAIS	45
7.REFERÊNCIAS BIBLIOGRÁFICAS	47
APÊNDICES.....	49

1. INTRODUÇÃO

Os métodos tradicionais de análise de sobrevivência assumem que existe heterogeneidade não observada e que as observações são independentes. Como resultado, eles não levam em conta o problema da dependência causada pela heterogeneidade não observada (Wienke 2009; Allison 2010). O modelo de Cox básico pode ser limitado para um contexto em que os dados de sobrevivência são correlacionados, como ocorre frequentemente em estudos multiníveis. Além disso, a falta de covariáveis importantes limita a capacidade explicativa do modelo, pois o tempo de sobrevivência pode ter uma variabilidade que é relativamente maior do que a explicada pelo modelo.

Um modelo de fragilidade fornece uma maneira conveniente de lidar com observações correlacionadas ou heterogeneidade não observada. Uma fragilidade é um efeito não observado, suposto aleatório, compartilhado pelas unidades observacionais ou experimentais dentro de um grupo. Na forma mais simples, o efeito aleatório (fragilidade) do grupo possui um efeito multiplicativo na função de risco para todos os indivíduos no mesmo grupo. Assim, indivíduos de grupos com valores grandes para o efeito aleatório (fragilidade) tendem a experimentar o evento antes de indivíduos pertencentes a grupos com valores pequenos no efeito aleatório (Klein e Moeschberger, 1997).

Como o assunto pode se tornar bastante complexo, este trabalho é sugerido como um texto para iniciantes no método e visa introduzir a técnica para ampliar o conhecimento dos leitores. Os objetivos deste trabalho estão apresentados no Capítulo 2. O Capítulo 3 aborda alguns conceitos essenciais da análise de sobrevivência. No Capítulo 4 são apresentados conceitos básicos de fragilidade a inclusão de efeitos aleatórios na modelagem da sobrevivência, bem como a caracterização das distribuições de probabilidade mais utilizadas para a fragilidade.

No Capítulo 5, um exemplo simulado é utilizado para explorar e apresentar procedimentos conceituais e computacionais, para o ajuste de modelos de fragilidade utilizando os programas R e SAS. Além disso, comparo diferentes modelos de Cox e modelos paramétricos a fim de orientar a escolha do melhor

ajuste, bem como estimar os efeitos aleatórios e interpretá-los, inclusive graficamente.

Algumas considerações e recomendações são feitas no Capítulo 6, além de conclusões a respeito do método e sugestões para novos trabalhos. Nos Apêndices estão os códigos necessários para gerar os dados simulados, assim como a modelagem.

2. OBJETIVOS

O objetivo principal do trabalho é apresentar conceitos básicos dos modelos de fragilidade para análise de sobrevivência, bem como procedimentos computacionais para o ajuste destes modelos utilizando os programas R e SAS.

3. Análise de Sobrevivência

A análise de sobrevivência consiste de um conjunto de métodos para a análise de dados em que o desfecho é o tempo até a ocorrência de algum particular evento de interesse, contado a partir de um tempo de origem bem definido. Estes métodos têm sido desenvolvidos e aplicados em diferentes áreas, com diferentes denominações, como por exemplo, na engenharia de qualidade (*reliability analysis*, *failure time analysis*), em economia (*transition analysis*, *duration analysis*), em sociologia (*event history analysis*) ou na pesquisa clínica e epidemiológica (*survival analysis* ou *lifetime data analysis*). Outra denominação usual, em diferentes áreas, é análise do tempo até o evento (*time to event analysis*).

O evento de interesse é definido como a transição de um estado discreto para outro, posicionada no tempo, como uma promoção no trabalho, a falha de um equipamento, o início (ou o diagnóstico) de uma doença ou o óbito. O tempo de origem é fundamental para a definição do tempo até o evento, podendo variar de um estudo para outro. Por exemplo, na pesquisa epidemiológica, indivíduos saudáveis são recrutados para um estudo coorte e acompanhados longitudinalmente para investigar a ocorrência de eventos incidentes, e a data de recrutamento no estudo pode ser usada como tempo de origem. Em um ensaio clínico randomizado para comparação de tratamentos, a data da aleatorização dos indivíduos aos grupos é frequentemente utilizada como tempo de origem, porém, alternativamente, a data início de tratamento poderia ser adotada.

Algumas vezes um evento adverso delimita o término de observação dos indivíduos, sendo chamado de falha, e o tempo observado é denominado tempo de falha. Em particular, se este evento é o óbito, o tempo observado é literalmente o tempo de sobrevivência. Em outras situações, o evento de interesse (alta hospitalar, remissão de doença, etc.) delimita o término do período de observação.

Uma característica importante dos estudos de sobrevivência é a possibilidade de não haver informação completa sobre o verdadeiro tempo de sobrevivência para todos os indivíduos. Vários motivos podem ocasionar a perda de informação temporal, tais como a impossibilidade de aferir a verdadeiro tempo inicial ou a existência de indivíduos ainda em acompanhamento na data do final do estudo,

sem ter sido observado o evento de interesse. Esses tempos incompletos são ditos censurados ou truncados.

Estes conceitos, apresentados de maneira informal, estão extensamente descritos na literatura, destacando-se Carvalho et al (2011), Collett (2004), Hosmer, Lemeshow e May (2008) ou Kleibaum e Klein (2005). A próxima seção sucintamente descreve aspectos sobre censura e truncamento.

3.1 Censura e truncamento

Muitos investigadores consideram que a análise de dados de sobrevivência é apenas a aplicação de dois métodos estatísticos convencionais para um tipo especial de problema: paramétrico, se a distribuição dos tempos de sobrevivência é suposta conhecida e não paramétrico, se a distribuição é considerada desconhecida (Lee e Wang; 2003).

Existem casos em que os tempos de sobrevida não são observados. Isto é, indivíduos de uma coorte podem não ser acompanhados até a ocorrência de um evento de interesse, no período de observação. Esta característica dos dados de sobrevivência é denominada censura e diferencia a análise de sobrevivência dos demais campos da estatística (Hosmer, Lemeshow e May; 2008). Em resumo, trata-se da perda de informação decorrente de não se ter observado a data de ocorrência de uma falha ou desfecho, resultando em observações incompletas.

Existem três tipos de censura em estudos de sobrevivência (Carvalho et al; 2011):

- Censura à direita: ocorre quando o verdadeiro tempo entre o início e o evento excede o tempo observado, e ignorar este aspecto pode levar a superestimação do risco do evento. As principais causas são o abandono do estudo ou o final do estudo sem que o evento tenha ocorrido.
- Censura à esquerda: ocorre quando o exato momento de ocorrência do evento é desconhecido, entretanto sabe-se que ele ocorreu em algum momento anterior ao da observação.

- Censura intervalar: ocorre quando o exato momento de ocorrência do evento é desconhecido, contudo sabe-se que ocorreu entre duas datas conhecidas.

As censuras também são classificadas como não informativa e informativa. No primeiro caso, as censuras ocorrem ao acaso e não estão relacionados com o desfecho. Por outro lado, censuras informativas ocorrem quando estão relacionadas ao desfecho, como, por exemplo, se os indivíduos abandonam o estudo devido a efeitos adversos a um dos tratamentos (Carvalho et al; 2011).

Outra forma de dados incompletos é denominada truncamento, e ocorre quando indivíduos não são incluídos no estudo por motivos relacionados à ocorrência de um determinado desfecho. Existem dois tipos de truncamento, que estão relacionados com a janela de tempo de observação pré-estabelecida (Carvalho et al; 2011):

- Dado truncado à esquerda: o critério de seleção inclui somente os indivíduos que sofreram o evento após o tempo de início da janela temporal de observação.
- Dado truncado à direita: o critério de seleção dos indivíduos inclui somente aqueles que sofreram o evento de interesse.

Para analisar dados de sobrevivência são necessários métodos que considerem a presença de censura (e também truncamento, quando for o caso). Desconsiderar que o tempo de sobrevivência (observado) é incompleto para alguns indivíduos pode gerar estimativas subestimadas ou superestimadas da probabilidade de sobrevivência ou de associações. A próxima seção apresenta aspectos básicos da função de sobrevivência e função de risco (*hazard function*), bem como abordagens comuns para a estimação destas funções. Estes métodos podem ser usados para análise de dados de sobrevivência quando não há necessidade de ajuste por covariáveis.

3.2 Função de sobrevivência e função de risco

O tempo de sobrevivência (T) pode ser considerado uma variável aleatória contínua não negativa (isto é, $t \geq 0$), com função densidade $f(t)$. Esta

função é usualmente interpretada como a probabilidade de um indivíduo sofrer um evento em um intervalo instantâneo de tempo, tal que

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

A função de distribuição (acumulada) da variável aleatória T é definida como

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

e representa a probabilidade de um indivíduo experimentar o evento em um tempo menor ou igual a t . Assim, a função de sobrevivência $S(t)$ é a probabilidade de um indivíduo sobreviver mais do que t unidades de tempo (isto é, não experimentar o evento antes de t), a partir do tempo de origem, tal que

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{+\infty} f(u) du.$$

A função de risco ou de taxa de falha (*hazard function*) é definida como

$$h(t) = -\frac{d}{dt} [\log S(t)]$$

e representa o risco instantâneo de um indivíduo experimentar o evento, condicional à sobrevivência até o tempo t , descrevendo a forma como a taxa instantânea de falha muda com o tempo.

As funções $f(t)$, $F(t)$ e $h(t)$ estão relacionadas por

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

tal que $S(t)$ e $h(t)$ são inversamente proporcionais. Ainda,

$$S(t) = \exp(-H(t)),$$

em que $H(t)$ é a função de risco acumulado definido como

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

Em um contexto aplicado, a função de sobrevivência $S(t)$ ou a função de risco $h(t)$, ou ambas, precisam ser estimadas. Se a distribuição de probabilidade do tempo, $f(t)$, é suposta conhecida, modelos paramétricos podem ser usados para a estimação. As distribuições utilizadas com mais frequência são a Exponencial, Lognormal e Weibull, e serão descritas na sequência.

3.3 Modelos Paramétricos

Esta seção apresenta modelos paramétricos para análise de sobrevivência, os quais podem ser úteis quando a distribuição até o evento é suposta conhecida (Sabrina Couto, 2005). Quando estes modelos oferecem um bom ajuste aos dados, tendem a produzir estimativas mais precisas sob as quantidades de interesse, pois se baseiam num número menor de parâmetros. Ao mesmo tempo, se o modelo paramétrico for escolhido de forma incorreta, pode conduzir a estimadores consistentes de quantidades erradas (Klein JP & Moeschberger, 2003).

3.3.1 Modelo de regressão exponencial

A distribuição mais simples para estudos de sobrevivência é a exponencial (Lee e Wang, 2003; Andreas Wienke, 2011). No final dos anos 1940, era utilizada principalmente em estudos de confiabilidade de sistemas eletrônicos e, posteriormente em estudos de tempo de vida e em remissão de doenças crônicas e infecciosas (Carvalho et al; 2011). O tempo de sobrevivência T tem uma distribuição exponencial com parâmetro λ (Notação: $T \sim \text{Exponencial}(\lambda)$) se sua função de densidade de probabilidade for

$$f(t) = \lambda \exp(-\lambda t); \lambda > 0,$$

e a função de distribuição acumulada é tal que

$$F(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-\lambda t}, & t \geq 0 \end{cases}$$

Assim, a função de sobrevivência é

$$S(t) = 1 - F(t) = e^{-\lambda t}.$$

Na Figura 1, podemos observar o comportamento da função de sobrevivência para o modelo exponencial, para diferentes valores de λ . Observa-se que quanto maior for o risco λ , mais rápido a função de sobrevivência $S(t)$ decresce.

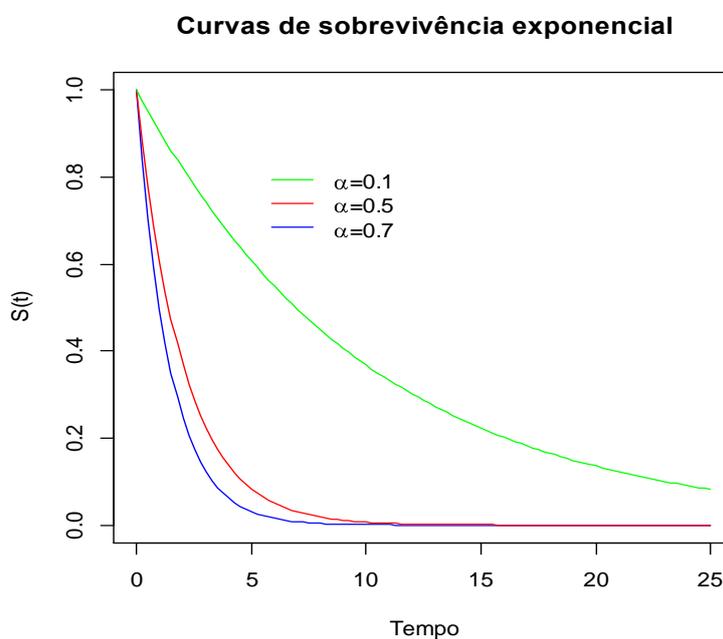


Figura 1 – Função de sobrevivência para o modelo exponencial com diferentes valores de λ .

A função de risco, que descreve a forma com que o risco do evento muda ao decorrer do tempo, é dada por

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda.$$

Sendo assim, a distribuição exponencial é adequada quando o risco for igual ao decorrer do tempo. Para um vetor de preditores \mathbf{x} , o modelo de regressão exponencial é dado por

$$h(t|\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p),$$

em que os coeficientes β_j representam os efeitos das covariáveis, sendo estimados a partir dos dados observados. A função de sobrevivência é definida por

$$S(t|\mathbf{x}) = \exp(-t \exp(\boldsymbol{\beta}'\mathbf{x})).$$

3.3.2 Modelo de regressão Lognormal

O modelo lognormal é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidades log-normal. Um aspecto interessante desta distribuição é que o logaritmo de uma variável com distribuição lognormal, com parâmetros μ e σ , tem distribuição normal, com média μ e desvio-padrão σ .

A função densidade de probabilidade para a variável aleatória tempo de falha T é

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}, \quad t > 0, \mu \in \mathfrak{R}, \sigma > 0,$$

onde μ e σ representam, a média e o desvio-padrão do logaritmo do tempo de falha, respectivamente.

A função de sobrevivência do tempo de falha é dada por

$$S(t) = \phi\left(\frac{-\log(t)+\mu}{\sigma}\right), \quad t > 0, \mu \in \mathfrak{R}, \sigma > 0,$$

e a função taxa de falha

$$h(t) = \frac{f(t)}{S(t)}, \quad t > 0, \mu \in \mathfrak{R}, \sigma > 0,$$

onde $\phi(\cdot)$ é a função de distribuição acumulada de uma distribuição normal padrão.

A característica da função taxa de falha do modelo paramétrico lognormal é ser inicialmente crescente e, quando atingir o ponto máximo passar a decrescer (Letícia Herrmann, 2011).

3.3.3 Modelo de regressão Weibull

Na prática, dificilmente a hipótese de que a função de risco seja constante, ou que os tempos de sobrevivência sejam distribuídos exponencialmente, é válida (Collett, 2004). A distribuição Weibull foi proposta por Waloddi Weibull 1954, em estudos relacionados ao tempo de falha devido à fadiga de metais. Trata-se de uma generalização do modelo exponencial com dois parâmetros positivos. O segundo parâmetro permite uma flexibilidade ao modelo e formas diferentes na função de risco. Atualmente é a distribuição de probabilidade do tempo mais utilizada na análise de sobrevivência, devido à simplicidade da função de risco e sobrevivência.

Se o tempo de sobrevivência T tem uma distribuição Weibull (Notação: $T \sim Weibull(\lambda, \gamma)$), a densidade $f(t)$ e a função de distribuição $F(t)$ são,

$$f(t) = \gamma \lambda^\gamma t^{\gamma-1} e^{-(\lambda t)^\gamma}$$

para $t \geq 0$, $\lambda > 0$ e $\gamma > 0$, em que λ e γ são chamados de parâmetros de escala forma, respectivamente. A função de distribuição é definida como

$$F(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-(\lambda t)^\gamma}, & t \geq 0 \end{cases}$$

de tal forma que a função de sobrevivência e a função de risco são, respectivamente,

$$S(t) = 1 - F(t) = e^{-(\lambda t)^\gamma}$$

e

$$h(t) = \frac{f(t)}{S(t)} = \gamma \lambda^\gamma t^{\gamma-1}.$$

Uma característica interessante é que a taxa de falha é monótona em relação ao tempo, e o parâmetro γ determina a forma como esta taxa se comporta com o tempo:

- Se $\gamma = 1$, a distribuição do tempo é exponencial e a função de risco $h(t)$ é constante no tempo;
- Se $\gamma < 1$, a função de risco $h(t)$ decresce no tempo; e,
- Se $\gamma > 1$, a função de risco $h(t)$ cresce no tempo.

A Figura 2 abaixo mostra o comportamento da função de sobrevivência para o modelo Weibull com parâmetro $\lambda=1$ e considerando diferentes valores do parâmetro de forma γ .

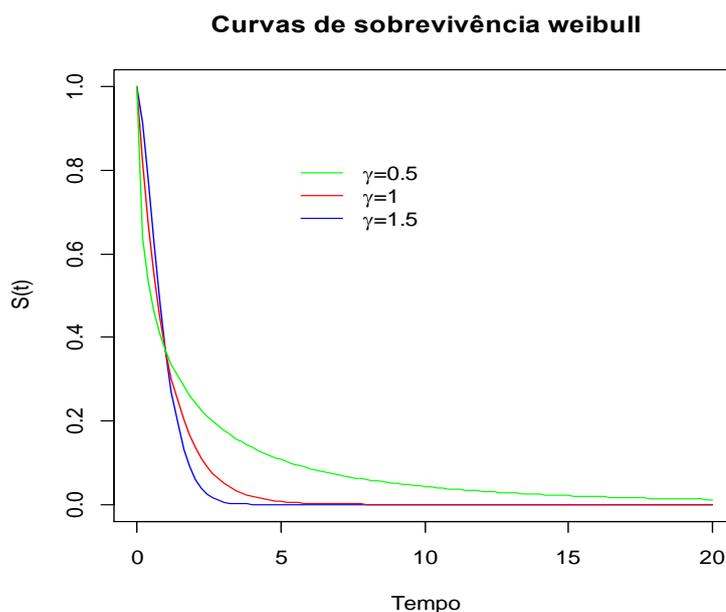


Figura 2 – Função de sobrevivência para o modelo Weibull com parâmetro γ .

Pela flexibilidade, a distribuição de Weibull tem sido bastante usada para modelagem paramétrica em aplicações clínicas e epidemiológicas (Pike, 1966; Whitthamore e Altshuler, 1976; Berry, 2007).

No modelo de regressão Weibull a distribuição do tempo de sobrevivência é Weibull. Este é o único modelo que pode ser considerado tanto um modelo de riscos proporcionais como de tempo de vida acelerado. Consideremos a sua formulação apenas em termos de modelo de riscos proporcionais, que é a mais comum. A função de risco é escrita como

$$h(t|\mathbf{x}) = \gamma t^{\gamma-1} \exp(\boldsymbol{\beta}' \mathbf{x})^\gamma = \gamma t^{\gamma-1} \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)^\gamma$$

e a função de sobrevivência é

$$S(t|\mathbf{x}) = \exp(-t[\exp(\boldsymbol{\beta}' \mathbf{x})]^\gamma).$$

3.4 Métodos não paramétricos

Quando a forma da distribuição do tempo de sobrevivência é desconhecida, ou seja, quando a distribuição de probabilidade do tempo de sobrevida não for especificada (Colton, 1979), estimadores não paramétricos são utilizados. Dessa forma, as funções de sobrevivência e de risco são estimadas sem a necessidade de fazer suposições a respeito da forma da distribuição do tempo de sobrevivência.

As abordagens não paramétricas utilizadas com maior frequência para a estimação da função de sobrevivência são:

- Tabelas de vida (estimador atuarial);
- Estimador de Kaplan-Meier; e,
- Estimador Nelson-Aalen.

Estes métodos estão amplamente descritos na literatura (Carvalho et al. 2011; Andreas Wienke, 2011; Therneau e Grambsch (2000)) e a principal diferença está na definição dos intervalos de tempo. No estimador atuarial estes intervalos podem ser definidos de forma arbitrária, porém intervalos com amplitude grande tendem a gerar estimativas viesadas. Os estimadores Kaplan-Meier e Nelson-Aalen são bastante parecidos, e os intervalos são definidos pelos sucessivos tempos em que ocorrem os eventos, produzindo estimativas assintoticamente não viesadas.

Para amostras de pequeno e médio porte, há evidências da superioridade do método de Kaplan-Meier.

Para análises ajustadas, no entanto, estes métodos têm limitações importantes, especialmente se existirem muitas variáveis envolvidas. O modelo de regressão de Cox, descrito na próxima seção, tem sido utilizado nestas situações.

3.4.1 Modelo de riscos proporcionais de Cox

Os modelos paramétricos de análise de sobrevivência requerem que a distribuição do tempo de sobrevivência seja suposta conhecida. Na prática, a adequação da distribuição de probabilidades postulada deve ser avaliada e, caso não seja identificada uma distribuição adequada, o uso de um modelo paramétrico poderá gerar resultados viesados ou inconsistentes. Nestas situações, o modelo de riscos proporcionais proposto por Cox (1972) é útil para estimar efeitos ou associações dos preditores com o tempo até ocorrência do evento de interesse. O modelo de Cox é robusto e flexível em diversos contextos, como estratificação, variáveis dependentes do tempo, eventos múltiplos, riscos competitivos e fragilidades. Sua força reside na sua capacidade de modelar e testar hipóteses específicas, independentemente da forma da distribuição do tempo de sobrevivência (David D. Hanagal, 2011).

Para definir o modelo, seja T uma variável aleatória contínua que representa o tempo de sobrevivência, $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ um vetor de covariáveis (ou preditores), e $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ o vetor que representa os correspondentes coeficientes de regressão desconhecidos. O modelo de Cox especifica que

$$h(t) = h(t|\mathbf{x}) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}),$$

em que $h_0(t)$ é uma função não negativa do tempo, que representa o risco basal de experimentar o evento quando $\mathbf{x} = \mathbf{0}$. Os coeficientes de regressão $\beta_j; j = 1, 2, \dots, p$ denotam os efeitos (ou associações) dos preditores com o tempo até a ocorrência do evento em estudo.

O modelo postula que as covariáveis têm um efeito multiplicativo na função de risco, de maneira que para dois indivíduos com preditores \mathbf{x}_r e \mathbf{x}_s , respectivamente, a razão das suas funções de risco é

$$\frac{h(t|\mathbf{x}_r)}{h(t|\mathbf{x}_s)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_r)}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_s)} = \exp\{\boldsymbol{\beta}'(\mathbf{x}_r - \mathbf{x}_s)\},$$

não depende do tempo t . Isso significa, por exemplo, se no início do estudo o indivíduo r tem um risco de morte igual a duas vezes o risco do indivíduo s , então esta razão de riscos será a mesma para todo o período de acompanhamento. Devido a esta propriedade, o modelo é chamado de modelo de riscos proporcionais.

Aspectos teóricos do modelo de Cox também estão amplamente descritos na literatura (Hosmer, D.W., 2008, Hanagal, D., 2011). Os parâmetros podem ser estimados sem a necessidade de fazer suposições sobre a forma distribuição do tempo de sobrevivência. O modelo é considerado semi-paramétrico, pois não assume qualquer distribuição para a função de risco basal $h_0(t)$ (Carvalho et al; 2011).

Segundo (Allison, 1984), a contribuição mais importante de Cox foi propor o método de estimação de parâmetros baseado na verossimilhança parcial (*partial likelihood*), que possui similaridades com método de máxima verossimilhança (*ordinary maximum likelihood estimation*).

A verossimilhança parcial $L(\boldsymbol{\beta})$, proposta por Cox (1972), é o produto das verossimilhanças individuais, ao longo dos tempos de falha sem censura, dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \prod_{t \geq 0} \left\{ \frac{Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_j Y_j(t)\exp(\boldsymbol{\beta}'\mathbf{x}_j)} \right\}^{\Delta N_i(t)},$$

em que $Y_j(t)$ é uma variável indicadora que assume o valor 1 se o indivíduo j está em risco no tempo t e 0, caso contrário, e $\Delta N_i(t)$ é a diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente

anterior a t . O numerador da verossimilhança parcial depende apenas das informações dos indivíduos que experimentaram o evento, ao passo que o denominador usa informações a respeito de todos os indivíduos que ainda não experimentaram o evento, inclusive aqueles que serão censurados posteriormente (Carvalho et al, 2011). A importância da utilização da verossimilhança parcial é que esta função depende apenas do vetor de parâmetros β e não envolve a função de risco de base $h_0(t)$.

4. Modelos de Fragilidade

O modelo de Cox básico pode ser limitado para um contexto em que os dados de sobrevivência são correlacionados, como ocorre frequentemente em estudos multiníveis. Isto pode ocorrer em estudos multicêntricos ou quando a unidade observacional ou experimental está agregada em um mesmo indivíduo, de modo que existem correlações entre estas unidades.

Um exemplo é o estudo do tempo até falha de implantes dentários, em que um paciente pode ter mais de um implante (Lin, 1994; Chuang, 2002 & Chuang, 2005). Diferentes abordagens de análise foram propostas, dentre as quais uma abordagem com estimação de equações generalizadas (*GEE- Generalized Estimating Equations*) e modelos de fragilidade (Lin, 1994; Chuang, 2002 & Chuang, 2005).

Em outras situações, variáveis importantes podem não ter sido medidas, resultando em grande heterogeneidade entre indivíduos. Essas diferentes fragilidades não podem ser atribuídas às variáveis observadas (Carvalho et al, 2011).

Um modelo de fragilidade fornece uma maneira conveniente de lidar com observações correlacionadas ou heterogeneidade não observada. Uma fragilidade é um efeito não observado, suposto aleatório, compartilhado pelas unidades observacionais ou experimentais dentro de um grupo (um indivíduo, uma ninhada, um centro de investigação, etc.). Na forma mais simples, o efeito aleatório (fragilidade) do grupo possui um efeito multiplicativo na função de risco para todos os indivíduos no mesmo grupo. Assim, indivíduos de grupos com valores grandes para o efeito aleatório (fragilidade) tendem a experimentar o evento antes de indivíduos pertencentes a grupos com valores pequenos no efeito aleatório (Klein e Moeschberger, 1997).

O termo de fragilidade foi introduzido por Vaupel et al. (1979) em modelos de sobrevivência univariados. Posteriormente, Clayton (1978) promoveu o modelo para uma aplicação multivariada, sem usar a noção de fragilidade, sobre a incidência de doenças crônicas em famílias.

Um modelo de efeito aleatório leva em conta os efeitos da heterogeneidade não observada, causadas por diferentes fontes. Os modelos de fragilidade podem ser úteis para análise de dados de sobrevivência quando existe heterogeneidade não observada entre indivíduos ou entre grupos. Carvalho et al (2011) exemplificaram um problema de heterogeneidade não observada em um estudo sobre a sobrevivência de pacientes em hemodiálise de diferentes centros de tratamento no estado do RJ (Carvalho et al, 2011). A ausência de indicadores socioeconômicos, estado nutricional do paciente, heterogeneidade dos pacientes (superdispersão), o tamanho dos centros de diálise, qualidade no atendimento, atenção recebida, qualidade dos equipamentos, formação dos profissionais, etc., resulta em uma variabilidade das observações maior do que o esperado, afetando assim a estimação dos parâmetros e limitando a capacidade explicativa do modelo. Sendo assim, a inclusão de um efeito aleatório permite avaliar a heterogeneidade entre os centros de diálise ou dos pacientes.

No estudo sobre implantes dentários (Chuang et al, 2005), o objetivo foi determinar se o risco de falha do implante varia entre os indivíduos e identificar fatores associados com o risco de insucesso do implante. A hipótese de que o risco de falha do implante varia entre os indivíduos foi testada, aplicando o modelo de riscos proporcionais de Cox com fragilidade, definida por um efeito aleatório compartilhado pelos implantes de um mesmo paciente. Alguns indivíduos foram mais propensos a ter falha nos implantes do que outros.

Estes exemplos mostram que a estrutura multinível das unidades observacionais ou experimentais pode surgir de diferentes maneiras. No exemplo da hemodiálise, os pacientes estão agrupados dentro dos centros de diálise, sugerindo que podem ser mais parecidos dentro dos centros e heterogêneos entre os centros. Esta heterogeneidade não observada pode ser considerada na análise por meio de um efeito aleatório para os centros.

O efeito aleatório, denominado de fragilidade e aqui denotado por Z é o termo que descreve o risco comum ou a heterogeneidade do indivíduo, agindo como um fator na função de risco. Podem-se distinguir duas categorias principais de modelos de fragilidade, univariada e multivariada. Em um modelo univariado, a fragilidade refere-se a um indivíduo com uma observação e é responsável pela heterogeneidade não observada no nível do indivíduo. A segunda é a classe de

modelos de fragilidade multivariados que levam em conta o tempo de sobrevivência multivariada e são utilizados para modelar as respostas a partir de observações agrupadas. Os aglomerados podem ser em múltiplos indivíduos por grupo, que ocorrem naturalmente, ou a partir de múltiplas medições dentro de um indivíduo. Como por exemplo, riscos concorrentes, recorrência de eventos em um mesmo indivíduo, ocorrência de uma doença em parentes, etc. Estes modelos são muitas vezes referidos como modelos de fragilidade compartilhados (Amrhein, 2014).

Como extensão do modelo de Cox clássico, a inclusão de efeitos aleatórios no modelo de fragilidade é realizada a partir de uma variável aleatória não observável Z que independe do tempo, com $E(Z)=1$ e $Var(Z)=\xi$. Posteriormente, iremos abordar as distribuições de fragilidade gama e lognormal. A fragilidade tem um efeito multiplicativo sobre a função de risco de base $h_0(t)$ e nos preditores. Valores pequenos de ξ indicam pouca heterogeneidade, isto é, não se torna necessário incluir o termo de fragilidade no modelo em questão. De modo análogo, valores altos de ξ indicam grande heterogeneidade entre indivíduos ou grupos. O modelo de riscos proporcionais com fragilidade $Z = z$ pode ser escrito como

$$h(t) = zh_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p),$$

em que $h_0(t)$ é a função de risco basal, x_1, x_2, \dots, x_p são preditores e $\beta_1, \beta_2, \dots, \beta_p$ são os respectivos coeficientes de regressão. O modelo pode ser paramétrico ou semi-paramétrico, podendo ser visto como uma generalização do modelo de riscos proporcionais clássico.

Um indivíduo ou grupo i com fragilidade $Z_i > 1$ tende a experimentar algum determinado evento com taxa mais rápida que no modelo de Cox básico. Do mesmo modo, aqueles que apresentarem fragilidades $Z_i < 1$ tendem a ter tempos maiores até a ocorrência do evento. Outra consideração a ser ressaltada é que, quando a variância ξ se aproxima de zero, o modelo se reduz ao modelo de Cox básico (Carvalho et al, 2011).

O modelo apresentado anteriormente também pode ser representado pela sua função de sobrevivência condicional

$$S(t|z, \mathbf{x}) = \exp\left(-z \int_0^t h(s|\mathbf{x}) ds\right) = \exp(-zH(t|\mathbf{x})),$$

em que $H(t) = \int_0^t h(s) ds$ representa a função de risco acumulado. A função $S(t|z, \mathbf{x})$ pode ser interpretada como a fração de indivíduos que sobrevivem até o tempo t , condicional ao vetor de preditores \mathbf{x} e fragilidade z .

É importante ressaltar que, até este momento, o modelo de fragilidade foi definido para um indivíduo, porém este modelo não é observável. Assim, é necessário definir o modelo populacional, tal que a função de sobrevivência populacional é a média das funções de sobrevivência individuais. Por meio de integração da transformada de Laplace, a função de sobrevivência marginal pode ser calculada por

$$S(t|\mathbf{X}) = \int S(t|Z, \mathbf{X}) g(z) dz = E[S(t|Z, \mathbf{X})]$$

em que $g(z)$ é a densidade da variável aleatória Z , isto é, a fragilidade, que deve ser suposta conhecida.

A escolha da distribuição de probabilidade para a fragilidade é um aspecto importante, a fim de estimar a variância ξ do efeito aleatório. As distribuições de fragilidade mais frequentemente utilizadas são as distribuições Gama (Clayton 1978; Vaupel et al. 1979), a distribuição de três parâmetros *power variance function* (Hougaard 1986a), a distribuição de Poisson composta (Aalen 1988, 1992) e a distribuição lognormal (McGilchrist e Aisbett, 1991).

A suposição da distribuição da fragilidade precisa ter plausibilidade biológica, e tanto a distribuição gama quanto a lognormal permitem essa aproximação, devido a sua flexibilidade (Carvalho et al, 2011). Dependendo dos valores de seus parâmetros, as distribuições variam desde uma forma exponencial até uma curva em forma de sino que lembra uma distribuição normal.

4.1 Fragilidade Gama

A distribuição gama tem sido amplamente aplicada como uma mistura de distribuições (Greenwood and Yule, 1920; Vaupel et al., 1979; Congdon, 1995; dos Santos et al., 1995; Hougaard, 2000). Do ponto de vista computacional e analítico, ele se encaixa muito bem aos dados de falha, pois é fácil para derivar as expressões com forma fechada para as funções de sobrevivência, para a densidade e para a função de risco. Isto só é possível devido à simplicidade da transformação de Laplace. Além disso, esta é também a razão pela qual esta distribuição tem sido utilizada na maioria das aplicações publicadas até o presente momento. A distribuição é flexível e apresenta diversidade de formas.

A distribuição gama é caracterizada por dois parâmetros, e uma notação usual é $Z \sim \Gamma(\gamma, \alpha)$. Os parâmetros γ e α são chamados de parâmetros de forma e escala, respectivamente. A média e a variância da variável aleatória Z são $E(Z) = \gamma\alpha$ e $Var(Z) = \gamma\alpha^2$. Quando $\gamma = 1$ a distribuição gama é idêntica à distribuição exponencial e, para valores de γ grandes, ela assume uma curva em forma de sino que lembra a de uma distribuição normal.

Apesar destas vantagens, é necessário mencionar que nenhuma razão biológica torna a distribuição gama preferível em relação às demais distribuições de fragilidade. Quase todos os argumentos a favor da distribuição gama são fundamentados em aspectos matemáticos e computacionais. Abbring e Van den Berg (2005) procuram explicar o uso de distribuições gama de fragilidades na análise de dados de sobrevivência. Os autores mostram que, em uma grande classe de modelos de fragilidade, a distribuição da fragilidade entre os sobreviventes converge para uma distribuição gama de acordo com os pressupostos.

No modelo de fragilidade é necessário que $E(Z) = 1$ e $Var(Z) = \xi$, o que pode ser obtido fazendo $\gamma = \frac{1}{\xi}$ e $\alpha = \xi$. A Figura 3 mostra o comportamento de uma variável aleatória Z com distribuição gama e $E(Z) = 1$, variando o valor da variância.

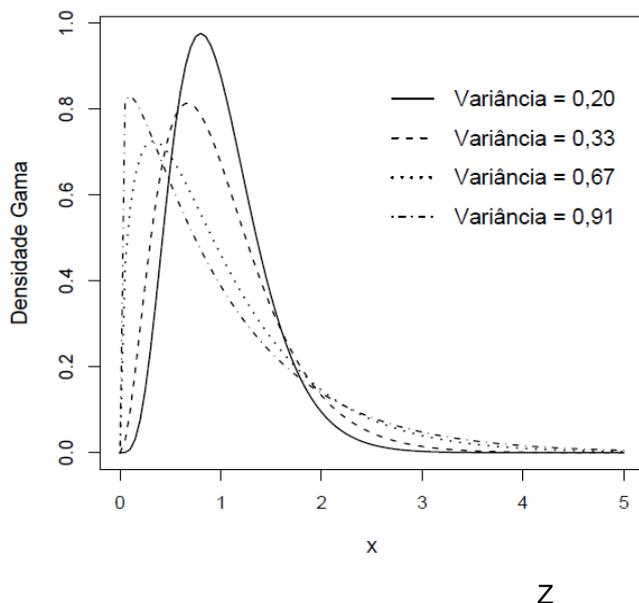


Figura 3: Funções de densidade de probabilidade de diferentes distribuições gama com média 1 e variâncias 0,2, 0,33, 0,67, 0,91 (Carvalho et al, 2011).

4.2 Fragilidade Lognormal

Assim como a distribuição gama, a distribuição lognormal é uma das mais importantes e frequentemente utilizadas em modelos de fragilidade. São especialmente úteis em estruturas de dependência de modelagem em modelos multivariados de fragilidade, como por exemplo, em McGilchrist e Aisbett (1991), McGilchrist (1993), Lillard (1993), Lillard et al. (1995), Xue e Brookmeyer (1996), Sastry (1997), Gustafson (1997), Ripatti e Palmgren (2000), Ripatti et al. (2002), Huang e Wolfe (2002). Contudo, a distribuição lognormal também tem sido aplicada em casos univariados, por exemplo, por Flinn e Heckman (1982).

Se W é uma variável aleatória com distribuição normal com média μ e variância σ^2 , então uma fragilidade Z com distribuição lognormal pode ser definida como $Z = \exp(W)$ (Hutchinson e Lai, 1991). Assim, a esperança e a variância da fragilidade Z estão relacionadas com parâmetros μ e σ^2 , tal que

$$E(Z) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \text{ e } \text{Var}(Z) = \exp(\sigma^2) - 1.$$

Desse modo, para que $E(Z)=1$ e $Var(Z)=\xi$, é suficiente utilizar a parametrização $\mu = -\frac{1}{2}\sigma^2$ e $\sigma^2 = \ln(1+\xi)$. A Figura 4 mostra o comportamento da densidade da fragilidade lognormal, para diferentes valores da variância.

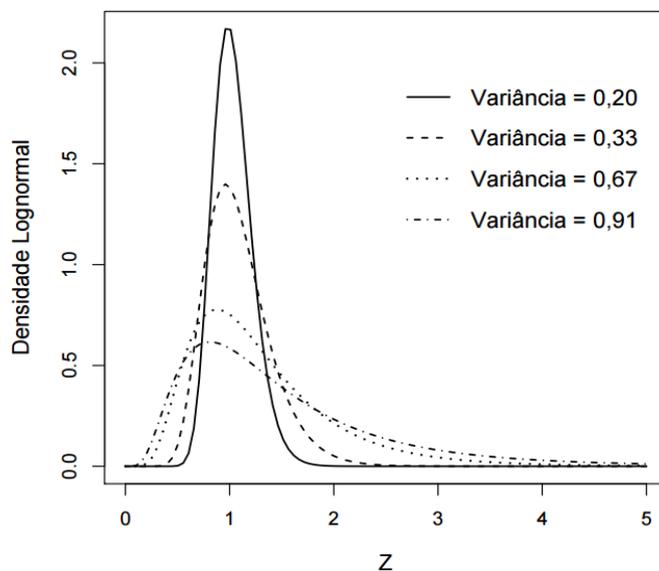


Figura 4: Funções de densidade de probabilidade de diferentes distribuições lognormal com média 1 e variâncias 0,2, 0,33, 0,67, 0,91 (Carvalho et al, 2011).

Um modelo de fragilidade compartilhada impõe que todos os indivíduos pertencentes a um grupo compartilham a mesma fragilidade Z , por isso o nome do modelo. Foi introduzido por Clayton (1978) e posteriormente estudado em Hougaard (2000), Therneau e Grambsch (2000), Duchateau et al. (2002 e 2003) e Duchateau e Janssen (2004). Assim, a variável de fragilidade está associada a grupos de indivíduos, em vez de indivíduos como tais. O modelo de risco para cada indivíduo pode ser escrito como

$$h(t, Z) = Zh_o(t),$$

e assume que, condicional à fragilidade, os tempos de falha são independentes. O valor da fragilidade Z é constante ao decorrer do tempo e comum aos indivíduos do

grupo e, por isso, é a causa para a dependência entre o tempo de vida dentro dos pares.

O próximo capítulo utiliza um conjunto de dados simulados para explorar aspectos computacionais para o ajuste de modelos de fragilidade utilizando os programas R e SAS.

5. Exemplo Simulado

Este exemplo é utilizado para mostrar aspectos conceituais e computacionais dos modelos de fragilidade. O contexto é de um estudo com acompanhamento de indivíduos ao longo do tempo, registrando-se o tempo até a ocorrência de um evento de saúde adverso. O tempo de origem pode ser considerado como sendo a data de diagnóstico de uma doença e o tempo final pode ser definido como a data de óbito ou censura.

Foi considerada uma amostra de $n = 10000$ indivíduos provenientes de 20 centros de investigação com características diferentes. Assim, espera-se que indivíduos de um mesmo centro possam mais similares entre si, e diferentes de indivíduos de outros centros. Essa heterogeneidade entre centros foi incorporada aos dados utilizando uma variável aleatória Z (a fragilidade) com distribuição lognormal com

$$E(Z) = e^{\frac{1}{2}} \approx 1,65$$

e

$$\text{Var } Z = e^2 - \left(e^{\frac{1}{2}} \right)^2 \approx 4,67 .$$

A fragilidade pode representar, por exemplo, a heterogeneidade não observada, devido às diferenças de qualidade do tratamento pelos centros de investigação.

Foram considerados os preditores Grupo (dicotômico), X_2 e X_3 , os dois últimos com distribuição normal padrão. A variável Grupo pode representar, por exemplo, dois tipos de tratamentos ou exposições, enquanto X_2 e X_3 podem ser considerados fatores de prognóstico ou confundidores contínuos padronizados.

É importante salientar que os dados foram simulados separadamente para cada centro de investigação, para especificar estruturas de correlação entre os preditores. A Seção 5.1 descreve os métodos computacionais utilizados para geração da amostra, utilizando o programa SAS, versão 9.4.

5.1 Geração dos dados

De forma sucinta, o algoritmo para a geração da amostra pode ser descrito pelas seguintes etapas:

- 1) Para cada centro de investigação k ($k=1,2,\dots,20$) foram simuladas n_k observações independentes do vetor aleatório $q=(Q_1, Q_2, Q_3)'$ com distribuição normal multivariada com $E(Q_j)=0$, $Var(Q_j)=\sigma_j^2=1$; $j=1,2,3$, e matriz de correlações

$$\rho_k = \begin{pmatrix} 1 & r_{j12} & r_{j13} \\ & 1 & r_{j23} \\ & & 1 \end{pmatrix}$$

utilizando a função “randnormal ()”. Estas observações foram armazenadas na matriz Q com dimensão $(n_k \times 3)$;

- 2) Foi definida a matriz U_k ($n_k \times 3$) por meio da transformação $U_j(Q_j)=\Phi(Q_j)$, $j=1,2,3$, em que $\Phi(\cdot)$ é a função de distribuição da normal padrão. Desse modo, as colunas de U são variáveis aleatórias com distribuição Uniforme em $[0,1]$, porém não são independentes;
- 3) Foi criada a matriz X_k ($n_k \times 3$), cujas colunas foram definidas por meio da transformação $X_j = F^{-1}(U_j)$, em que $F^{-1}(\cdot)$ é a inversa da função de distribuição especificada: para X_1 foi usada a distribuição Uniforme em $[0,1]$; para X_2 e X_3 foi usada a distribuição normal padrão;
- 4) Para cada centro de investigação k ($k=1,2,\dots,20$) foi simulado um valor da distribuição log-normal, para representar as correspondentes fragilidades. Este valor foi armazenado na variável (coluna) FRAILTY, no arquivo de **FRAILTYk**;
- 5) Para cada centro de investigação k ($k=1,2,\dots,20$) foi criado um banco de dados chamado **CENTROk**, com n_k linhas e as seguintes colunas:
 - i) As colunas X_1, X_2 e X_3 da matriz X_k e a coluna FRAILTY contendo o valor do arquivo **FRAILTYk**, de tal forma que todos os indivíduos do k -ésimo centro compartilham a mesma fragilidade;
 - ii) Uma variável para identificação do centro: $CENTRO = k$; e,

iii) Uma variável de identificação dos indivíduos em cada centro:
 $ID = CENTRO * 100 + _n_$, em que $_n_$ é uma variável de sistema do programa SAS que assume, sequencialmente, o número da linha de dados no processamento do *datastep*;

iv) Uma variável que representa o grupo de exposição (tratamento) do indivíduo, definida como

$$GRUPO = \begin{cases} 1, & \text{se } x_1 \leq u \\ 0, & \text{se } x_1 > u \end{cases},$$

em que u é o valor de uma variável aleatória com distribuição Uniforme em $U[0,1]$, independente de x_1 ;

6) Foi criado um banco de dados contendo todas as observações, empilhando os bancos Centro1, Centro2, ..., Centro20;

7) Foi criado o banco de dados do exemplo, adicionando as variáveis TEMPO e CENSURA, como descrito abaixo:

i) O tempo de sobrevivência (TEMPO) foi gerado utilizando uma distribuição exponencial, e está relacionado com os preditores por meio da relação

$$\text{linpred} = \exp(1,2 \times Grupo - 0,2 \times x_2 + 0,8 \times x_3 - Frailty),$$

utilizada para definir o parâmetro β da distribuição exponencial que gera o tempo até o evento, como

$$f(t) = \frac{1}{\beta} e^{-\frac{1}{\beta}t} I_{(0,+\infty)}(t)$$

em que $\beta = 5 \times \text{linpred}$.

ii) Para definir as censuras foram gerados valores de uma variável C com distribuição exponencial com parâmetro $\beta^* = 10$, tal que

$$f(c) = \frac{1}{\beta^*} e^{-\frac{1}{\beta^*}c} I_{(0,+\infty)}(c).$$

A variável que identifica a presença de censura foi definida como

$$CENSURA = \begin{cases} 1, & \text{se } c < t \text{ ou } t > 20 \\ 0, & \text{caso contrário} \end{cases},$$

em que o valor 20 representa o final do estudo.

5.2 Análise descritiva

A Tabela 1 apresenta uma descrição da amostra, por centro, com respeito ao número de indivíduos global e por grupo, frequência de eventos, o tempo mediano (IC 95%) até o evento ou censura, e média e desvio padrão dos preditores X2 e X3. A última coluna mostra o valor simulado para a fragilidade em cada centro.

Tabela 1 – Descrição da amostra, por centro.

Centro	n	Grupo				Eventos		Censuras		Tempo Mediano (IC 95%) [†]			X2		X3		Fragilidade
		0	1	n	%	n	%	Mediana	Inferior	Superior	Média	Desvio-padrão	Média	Desvio-padrão			
1	650	330	320	647	99,54	3	0,46	0,0907	0,0790	0,1031	0,0217	0,9475	-	0,9300	4,1347		
2	450	218	232	431	95,78	19	4,22	0,2689	0,2230	0,2985	0,0367	1,0365	0,0207	0,9510	3,1991		
3	450	230	220	345	76,67	105	23,3	1,8445	1,6790	2,2227	0,0701	0,9778	0,0075	1,0232	1,0934		
4	500	276	224	444	88,8	56	11,2	0,7327	0,5960	0,8063	-	0,9968	0,0048	0,9886	2,0947		
5	610	315	295	515	84,43	95	15,6	0,9404	0,8190	1,0808	0,0160	1,0197	0,0038	0,9969	1,7701		
6	600	279	321	368	61,33	232	38,7	3,9858	3,5360	4,5224	0,0297	1,0337	0,0562	0,9917	0,4148		
7	550	278	272	267	48,55	283	51,5	5,8647	5,0080	7,0005	0,0135	0,9713	0,0392	0,9598	0,1289		
8	350	173	177	315	90	35	10	0,6670	0,5350	0,7710	0,0282	0,9939	0,0102	1,0113	2,1500		
9	600	294	306	438	73	162	27	2,2173	1,9080	2,5746	0,0532	1,0371	0,0060	0,9957	1,0843		
10	400	194	206	231	57,75	169	42,3	3,9971	3,3800	4,5408	0,0072	1,0448	0,0574	1,0018	0,3773		
11	500	249	251	337	67,4	163	32,6	2,7700	2,4440	3,3172	0,0170	1,0051	0,0271	0,9735	0,7284		
12	370	186	184	339	91,62	31	8,38	0,3333	0,2670	0,4020	0,0956	1,0509	0,1054	1,0016	2,8811		
13	500	250	250	403	80,6	97	19,4	1,1317	1,0020	1,2912	0,0097	0,9732	0,0134	0,9860	1,5484		
14	500	242	258	300	60	200	40	4,3618	3,9900	5,1703	0,0067	1,0095	0,0561	0,9841	0,1483		
15	430	230	200	402	93,49	28	6,51	0,3198	0,2580	0,3837	0,0355	0,9850	0,0008	0,9431	2,9519		
16	700	337	363	699	99,86	1	0,14	0,0046	0,0040	0,0053	0,0205	0,9966	0,0129	0,9732	7,1280		
17	600	293	307	356	59,33	244	40,7	4,4890	3,7460	5,3986	0,0053	0,9992	0,0140	1,0006	0,3193		
18	340	157	183	230	67,65	110	32,4	2,9615	2,2330	3,6131	0,0649	1,0017	0,0338	0,9941	0,6511		
19	400	193	207	295	73,75	105	26,3	1,6299	1,3450	1,9443	0,0331	0,9999	0,0794	1,0444	1,2090		
20	500	245	255	278	55,6	222	44,4	4,7484	4,3170	5,6157	0,1200	1,0298	0,1120	1,0058	0,1688		

[†] Estimado pelo método de Kaplan-Meier

A Figura 5.1 mostra o gráfico de dispersão do tempo de acordo com o valor da fragilidade, evidenciado que sujeitos que compartilham valores grandes da fragilidade tendem a apresentar tempo até o evento menor. A Figura 5.2 mostra o impacto da fragilidade compartilhada pelos indivíduos nos centros, sugerindo

presença de grande heterogeneidade na sobrevivência por centro. As funções de sobrevivência estimadas pelo método de Kaplan-Meier apresentadas na Figura 5.1 mostraram que a sobrevivência é maior para centros com fragilidades baixas.

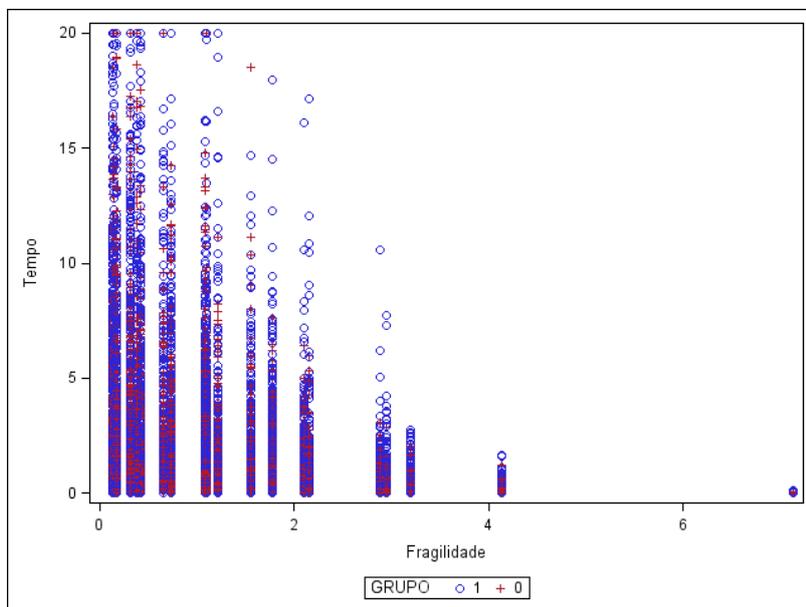


Figura 5.1 – Dispersão do tempo de acordo com os valores de fragilidade, por grupo.

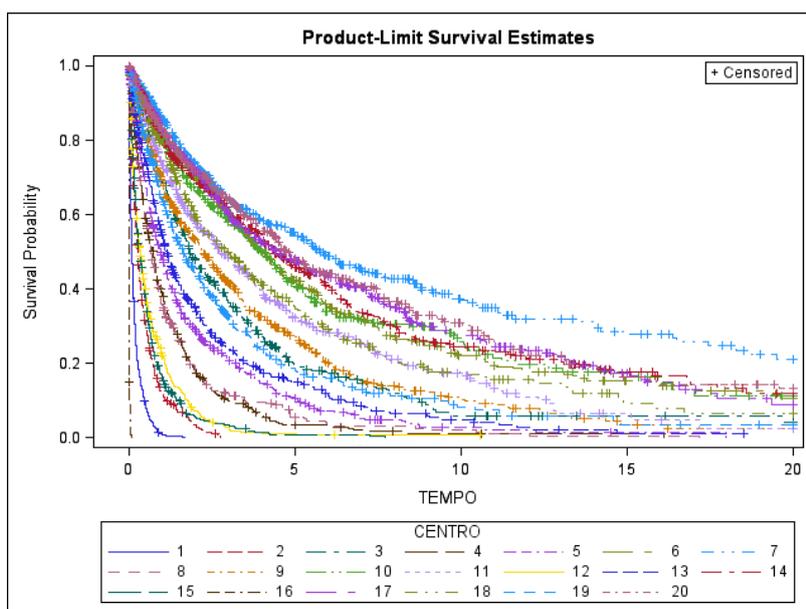


Figura 5.2 – Estimativas da função de sobrevivência global, por centro, pelo método de Kaplan-Meier.

5.3 Modelo de Cox

Na Tabela 2 encontram-se os resultados do modelo de Cox sem fragilidade. Os preditores mostraram associações (efeitos) significativas, porém seria esperado que os coeficientes de regressão fossem próximos dos valores utilizados na simulação ($\beta_{Grupo} = -1,2$, $\beta_{x_2} = 0,2$ e $\beta_{x_3} = -0,8$). As rotinas SAS e R utilizadas para ajustar o modelo de Cox sem fragilidade são mostradas abaixo.

Sintaxe SAS

```
procphregdata=Exemplo1;
title"Modelo de Cox (sem fragilidade)";
model TEMPO*CENSURA(1) = Grupo X2 X3 / rl;
run;
```

Sintaxe R

```
modelo1<- coxph(Surv(TEMPO, 1- CENSURA) ~ GRUPO + X2 +
X3, data=simulacao)

summary(modelo1)
```

Tabela 2 – Estimativas[†] dos coeficientes de regressão (EP) e da razão de azares obtidas pelo modelo de Cox sem fragilidade.

Preditor	Coeficientes de regressão		Razão de azares	
	Beta	EP*	Estimativa pontual	IC 95%
Grupo	-0,75229	0,02597	0,471	(0,448-0,496)
X2	0,13517	0,01501	1,145	(1,112-1,179)
X3	-0,47600	0,01565	0,621	(0,603-0,641)

* EP = Erro padrão

[†] Resultados produzidos pelo SAS, praticamente idênticos aos do programa R.

A heterogeneidade entre centros não é considerada no modelo da Tabela 2, resultando em estimativas de coeficientes de regressão muito discrepantes dos valores esperados. A fragilidade compartilhada pelos indivíduos de um mesmo centro pode ser considerada utilizando um modelo de Cox com fragilidade, o qual considera os indivíduos aninhados aos centros. Isto pode ser realizado

especificando um efeito aleatório para centro. Esse efeito aleatório permite medir o perfil de risco de cada centro.

A rotina SAS utilizada para ajustar o modelo com fragilidade lognormal é mostrada abaixo. A instrução **random** especifica um efeito aleatório para os centros, com distribuição lognormal (**dist=lognormal**) e a opção **solution** requisita as estimativas de parâmetros da fragilidade. Desta forma, os indivíduos são agregados de forma hierárquica aos centros, e a variável CENTRO deve ser definida como categórica (**class CENTRO**). Este modelo especifica efeitos fixos para as variáveis Grupo, X2 e X3, e um efeito aleatório para CENTRO.

Sintaxe SAS

```
procphreg data=Exemplo1;
title "Modelo de Cox (com fragilidade lognormal)";
class CENTRO;
model TEMPO*CENSURA(1) = Grupo X2 X3 / rl;
random CENTRO / dist=lognormal solution;
run;
```

Para ajustar o modelo utilizando o programa R, foi usada a sintaxe abaixo. O comando *coxph* ajusta os parâmetros do modelo de regressão de Cox. O termo **frailty** especifica um efeito aleatório para os centros, com distribuição lognormal. A distribuição *default* é a gama. Para indicar a distribuição lognormal use-se o argumento **dist="gauss"**. O objeto, chamado aqui de modelo2, contém o resultado do ajuste do modelo de Cox. A função *summary* lista o resultado completo do modelo.

Sintaxe R

```
modelo2<- coxph(Surv(TEMPO, 1-CENSURA) ~ GRUPO + X2 + X3
+ frailty(CENTRO, dist = "gauss"), data = simulacao)

summary(modelo2)
```

A Tabela 3 apresenta os resultados do modelo de Cox com fragilidade lognormal, isto é, especificando um efeito aleatório com distribuição lognormal para os centros. Nitidamente as estimativas dos coeficientes de regressão são bastante similares aos valores esperados, portanto, as estimativas de razão de azares são bastante diferentes das da Tabela 2. A variância dos efeitos aleatórios dos centros é 2,7980 e foi significativamente diferente de zero ($p < 0,0001$), evidenciando que as observações são correlacionadas dentro dos centros e existe heterogeneidade entre centros.

Tabela 3 – Estimativas[†] dos coeficientes de regressão (EP) e da razão de azares obtidas pelo modelo de Cox com fragilidade compartilhada lognormal.

Preditor	Coeficientes de regressão		Razão de azares	
	Beta	EP*	Estimativa pontual	IC 95%
Grupo	-1,22173	0,02722	0,295	(0,279-0,311)
X2	0,20658	0,01567	1,229	(1,192-1,268)
X3	-0,78391	0,01695	0,457	(0,442-0,472)

* EP = Erro padrão

[†] Resultados produzidos pelo SAS, praticamente idênticos aos do programa R.

A Tabela 4 apresenta as estimativas dos efeitos aleatórios para os centros e das correspondentes fragilidades (IC 95%) com distribuição normal. A coluna da fragilidade pode ser obtida pela exponencial da estimativa pontual (exemplo, para o Centro 1, $e^{2,473} = 11,858$), representando a fragilidade lognormal. A mesma operação pode ser usada para obter intervalo com 95% de confiança para a fragilidade lognormal, exponenciando os limites inferior e superior de confiança para o efeito aleatório. Indivíduos pertencentes aos centros com estimativas de fragilidade lognormal > 1 tendem a experimentar o evento mais cedo do que sob o modelo de independência.

Tabela 4 – Estimativas[†] dos efeitos aleatórios (IC 95%) e das fragilidades, por centro.

Centro	Efeito aleatório			Fragilidade Lognormal (IC 95%)
	Estimativa Pontual	EP*	IC 95%	
1	2,473	0,386	(1,7157;3,2303)	11,858(5,561;25,286)
2	1,508	0,387	(0,7495;2,2662)	4,517(2,116;9,643)
3	-0,590	0,387	(-1,3489;0,1695)	0,554(0,260;1,185)
4	0,351	0,387	(-0,4067;1,1088)	1,421(0,666;3,031)
5	0,114	0,386	(-0,6426;0,8714)	1,121(0,526;2,390)
6	-1,312	0,387	(-2,0705;-0,5526)	0,269(0,126;0,575)
7	-1,712	0,389	(-2,4738;-0,9510)	0,180(0,084;0,386)
8	0,501	0,388	(-0,2588;1,2610)	1,651(0,772;3,529)
9	-0,643	0,387	(-1,4008;0,1148)	0,526(0,246;1,122)
10	-1,302	0,389	(-2,0644;-0,5392)	0,272(0,127;0,583)
11	-0,954	0,388	(-1,7137;-0,1948)	0,385(0,180;0,823)
12	1,066	0,388	(0,3060;1,8250)	2,902(1,358;6,203)
13	-0,202	0,387	(-0,9601;0,5564)	0,817(0,383;1,744)
14	-1,423	0,388	(-2,1834;-0,6625)	0,241(0,113;0,516)
15	1,184	0,387	(0,4256;1,9427)	3,268(1,531;6,978)
16	5,380	0,390	(4,6154;6,1438)	216,939(101,033;465,812)
17	-1,400	0,387	(-2,1592;-0,6407)	0,247(0,115;0,527)
18	-1,010	0,389	(-1,7727;-0,2475)	0,364(0,170;0,781)
19	-0,503	0,388	(-1,2628;0,2579)	0,605(0,283;1,294)
20	-1,527	0,388	(-2,2878;-0,7657)	0,217(0,101;0,465)

* EP = Erro padrão

[†] Resultados produzidos pelo SAS, similares aos do programa R.

No programa R, estas estimativas ficam armazenadas no objeto *frail* gerado em cada modelo e estão dispostas abaixo, em ordem crescente de centro.

	Estimativas dos efeitos aleatórios, por centro.					
[1]	2,4728	1,5077	-0,5897	0,3511	0,1144	-1,3115
[7]	-1,7122	0,5011	-0,6429	-1,3016	-0,9542	1,0654
[13]	-0,2018	-1,4228	1,1841	5,3790	-1,3998	-1,0100
[19]	-0,5024	-1,5266				

Estimativas das fragilidades lognormal, por centro.						
[1]	11,8554	4,5163	0,5545	1,4206	1,1212	0,2694
[7]	0,1805	1,6505	0,5257	0,2721	0,3851	2,9020
[13]	0,8173	0,2410	3,2676	216,8120	0,2466	0,3642
[19]	0,6051	0,2173				

A Figura 5.3 mostra o histograma das estimativas pontuais dos efeitos aleatórios para os centros, porém é mais fácil interpretar a heterogeneidade entre centros na Figura 5.4, visto que também considera as estimativas de variabilidade.

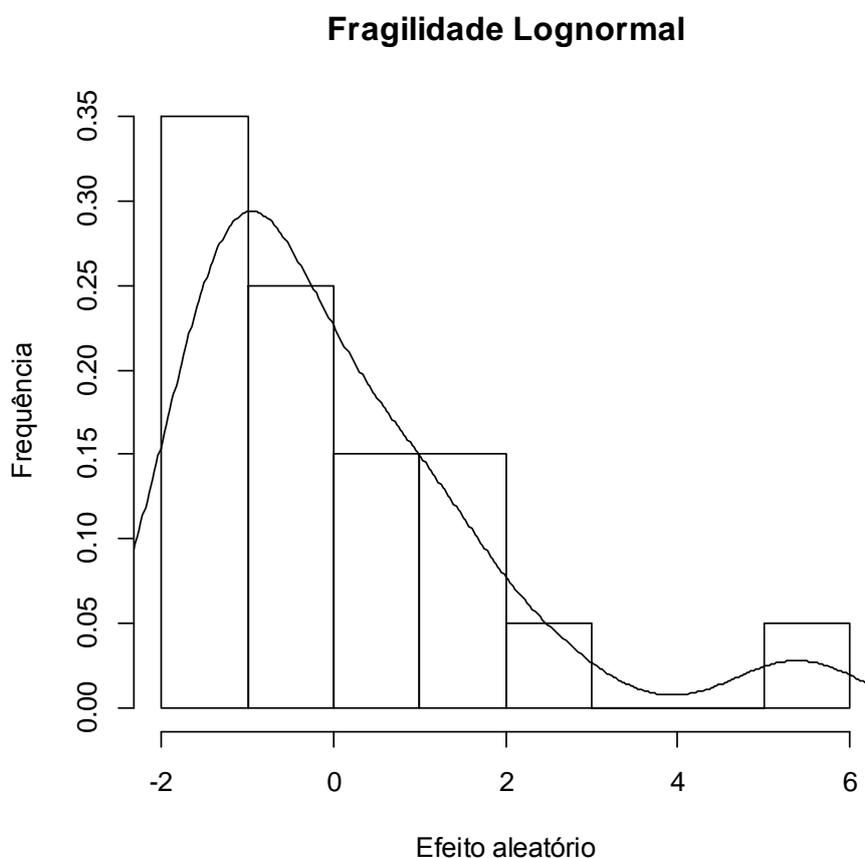


Figura 5.3 – Histograma do efeito aleatório do modelo com fragilidade.

Para a elaboração do gráfico da Figura 5.4 foi necessário carregar as funções que estão no arquivo Rfun.r disponibilizada na página <http://sobrevida.fiocruz.br/rlatex.html> (acessada em 28/04/2016). Os centros foram ordenados pela estimativa pontual da fragilidade. Este gráfico mostra que os centros com menor risco associado ao evento são os centros 7 e 20, identificadas na

fragilidade lognormal. Do mesmo modo, os centros identificados como tendo maior risco significativo foram os centros 1 e 16.

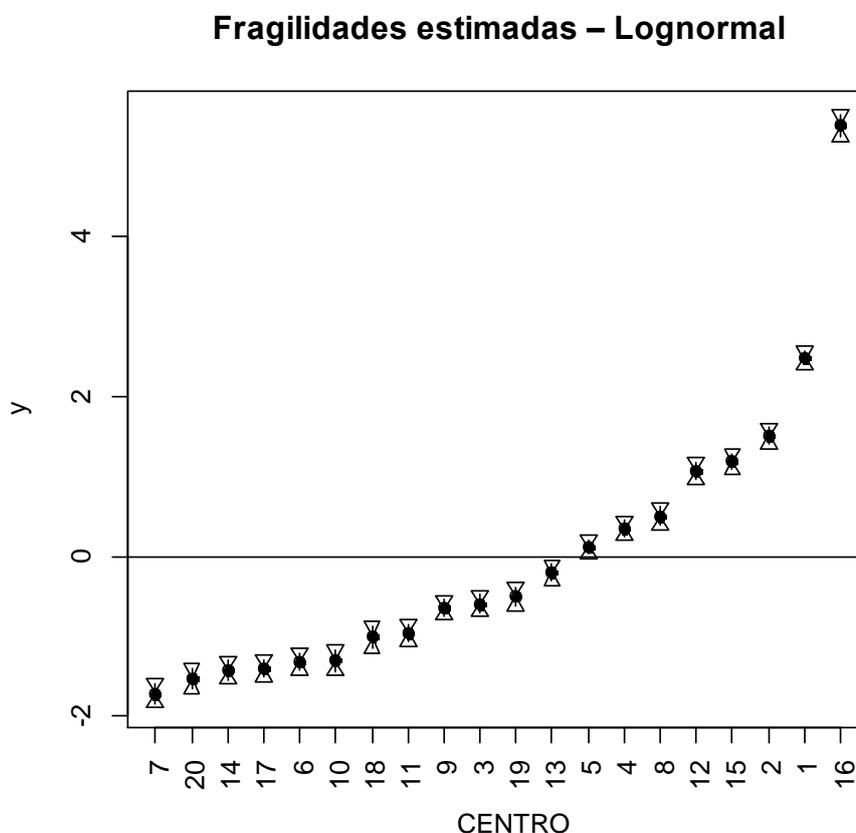


Figura 5.4 – Estimativas das fragilidades (IC 95%), por centro, estimadas pelo modelo de Cox com erro aleatório com distribuição lognormal.

5.4 Modelos paramétricos com fragilidade

Quando a distribuição do tempo é conhecida é possível utilizar um modelo paramétrico com o termo de fragilidade. No exemplo em discussão, foi usada a distribuição exponencial para simular o tempo até o evento. A Tabela 5 mostra as estimativas dos parâmetros do modelo paramétrico, especificando a distribuição Weibull para o tempo e erro aleatório com distribuição normal. A sintaxe abaixo foi usada no procedimento PROC NL MIXED do programa SAS para ajustar o modelo. Os termos “*gamma*” e “*alpha*” representam os parâmetros de forma e de escala da

distribuição Weibull; “*linp*” representa o preditor linear do modelo, em que b_0 é um intercepto e b_1 , b_2 e b_3 são os coeficientes de regressão associados aos preditores e “*ll*” representa o logaritmo da função de verossimilhança. O erro aleatório é representado por “*z*”, com distribuição normal com média zero e variância $e^{2 \times \text{logsig}}$, isto é, $z \sim N(0, e^{2 \times \text{logsig}})$, e a opção “**out=EB**” é usada para obter as estimativas dos efeitos pelo estimador de Bayes empírico (*empirical Bayes*). A instrução “**estimate**” foi usada para estimar a razão de azares (IC 95%), em que a constante 2,093024054 é o valor da distribuição *t* de Student com 19 graus de liberdade que delimita o nível de confiança de 95%, e as demais constantes são os correspondentes erros padrão associados aos coeficientes de regressão.

Sintaxe SAS

```

procnlmixeddata=Exemplo1;
  bounds gamma >0;
  linp = b0 + b1*Grupo + b2*X2 + b3*X3 + z;
  alpha = exp(-linp);
  G_t = exp(-(alpha*Tempo)**gamma);
  g = gamma*alpha*((alpha*Tempo)**(gamma-1))*G_t;
  ll = (Censura=0)*log(g) + (Censura=1)*log(G_t);
  model Tempo~ general(ll);
  random z ~ normal(0,exp(2*logsig)) subject=CENTRO out=EB;
  predictG_tout=Gt;
  id CENTRO;
  estimate'Grupo: HR (1 x 0)'exp(-b1);
  estimate'Grupo: IC95 HR_Inf'exp(-b1-2.093024054*0.02572);
  estimate'Grupo: IC95 HR_Inf'exp(-b1+2.093024054*0.02572);
  estimate'HR: X2 (Aumento de 1 unidade)'exp(-b2);
  estimate'X2: IC95 HR_Inf'exp(-b2-2.093024054*0.01564);
  estimate'X2: IC95 HR_Inf'exp(-b2+2.093024054*0.01564);
  estimate'HR: X3 (Aumento de 1 unidade)'exp(-b3);
  estimate'X3: IC95 HR_Inf'exp(-b3-2.093024054*0.01605);
  estimate'X3: IC95 HR_Inf'exp(-b3+2.093024054*0.01605);
run;

```

A Tabela 5 mostra as estimativas dos parâmetros do modelo. Como esperado, a estimativa do parâmetro de forma (**gamma**) foi aproximadamente igual a 1, haja vista que o tempo até o evento foi simulado utilizando uma distribuição exponencial. As estimativas dos coeficientes e da razão de azares são similares àquelas obtidas pelo modelo de Cox com fragilidade. A variância dos efeitos aleatórios dos centros é $e^{2 \times 0.5204} = 2,8315$ e foi significativamente diferente de zero ($p=0,0009$), evidenciando que existe heterogeneidade não observada entre centros ou, equivalentemente, que as observações são correlacionadas dentro dos centros.

As estimativas dos efeitos aleatórios (fragilidade normal) e respectivos erros padrão também foram muito similares às aquelas mostradas na Tabela 4.

Tabela 5 – Estimativas[†] dos coeficientes (EP) e razão de azares obtidas pelo modelo de paramétrico com fragilidade compartilhada lognormal.

Preditor	Coeficientes de regressão		Razão de azares	
	Beta	EP*	Estimativa pontual	IC 95%
Gamma	0,9975	0,008890	-	-
Intercepto	-0,1211	0,3766	-	-
Grupo	1,2268	0,02572	0,2932	0,2779 – 0,3095
X2	-0,2076	0,01564	1,2307	1,1910 – 1,2716
X3	0,7873	0,01605	0,4551	0,4400 – 0,4706
logsig	0,5204	0,1582	-	0,1892 – 0,8516

* EP = Erro padrão

[†] Resultados produzidos pelo SAS, praticamente idênticos aos do programa R.

A sintaxe R para ajuste do modelo é mostrada abaixo. O comando **survreg** estima modelos paramétricos de sobrevivência. O termo **frailty** especifica um efeito aleatório para os centros. Para indicar a distribuição Weibull usa-se o argumento **dist="weibull"**. O objeto, chamado **model.weibull**, contém o resultado do ajuste do modelo. A função **summary** lista o resultado completo do modelo. Os resultados e interpretações são similares aos do PROC NLMIXED e não serão mostrados.

Sintaxe R

```
model.weibull <- survreg(Surv(TEMPO, 1-CENSURA) ~ GRUPO + X2 + X3 +
  frailty(CENTRO), data=simulacao, dist="weibull")

summary(model.weibull)
```

6. CONSIDERAÇÕES FINAIS

O modelo de riscos proporcionais de Cox é comumente utilizado para avaliar os efeitos de covariáveis observáveis. Contudo, a falta de covariáveis importantes limita a capacidade explicativa do modelo, pois o tempo de sobrevivência pode ter uma variabilidade que é relativamente maior do que a explicada pelo modelo.

Quando as observações são correlacionadas e/ou existe heterogeneidade não observada, a inclusão de efeitos aleatórios pode resultar em estimativas dos efeitos das covariáveis mais consistentes. Nesse contexto, a aplicação de um modelo de fragilidade compartilhada se dá quando se supõe não somente a existência de heterogeneidade não observada entre grupos, mas também a dependência dos tempos de sobrevivência dentro de um grupo ou entre tempos dentro de um mesmo indivíduo. Desse modo, este trabalho apresenta a inclusão da fragilidade no modelo de riscos proporcionais de Cox e no modelo de regressão paramétrico.

Com o intuito de mostrar aspectos conceituais e computacionais dos modelos de fragilidade, foi simulado um exemplo em que o contexto é de um estudo com acompanhamento de indivíduos ao longo do tempo, registrando-se o tempo até a ocorrência de um evento de saúde adverso. O desempenho das estimativas dos parâmetros das covariáveis observadas nos modelos foram avaliadas e comparadas.

A simulação mostrou que, no modelo de Cox básico, as estimativas de coeficientes de regressão são muito discrepantes dos verdadeiros valores utilizados na simulação. Contudo, no modelo de Cox com fragilidade lognormal, ou seja, especificando um efeito aleatório com distribuição lognormal para os centros, as estimativas dos coeficientes de regressão são bastante similares aos valores esperados, evidenciando a heterogeneidade entre os centros e o impacto de utilizar o modelo correto.

Além disso, as estimativas dos coeficientes e da razão de azares dos modelos paramétrico com fragilidade são similares às aquelas obtidas pelo modelo de Cox com fragilidade. Sendo assim, os modelos com fragilidade propostos neste

trabalho parecem ser ideais, pois ao se incluir o efeito aleatório no modelo, a estimativa pontual dos coeficientes de regressão é similar ao verdadeiro valor. Ademais, o efeito aleatório para o centro foi significativo, evidenciando a grande heterogeneidade entre os centros.

Contudo, como apontado por (Carvalho et al, 2011), os modelos de efeitos aleatórios apresentam um vasto leque de possibilidades na modelagem. Aqui foram apresentadas as técnicas mais usuais e, como continuidade a este trabalho, sugere-se uma exploração mais profunda de métodos, simulação com replicação e com menores tamanhos de amostra. Assim como uma investigação mais aprofundada em métodos de escolha de modelos, avaliação de ajuste, etc. Em especial, dedicação aos modelos paramétricos e métodos bayesianos para estimar fragilidade no modelo de Cox, os quais não foram abordados.

7.REFERÊNCIAS BIBLIOGRÁFICAS

Amrhein, J., (2014) Introduction to frailty models.

Bena, J. (2008) Survival Methods for Correlated Time-to-Event Data.

Carvalho, M. S. et al. Análise de Sobrevivência: Teoria e Aplicações em Saúde. 2ª edição, Rio de Janeiro: Ed. Fiocruz, 2011.

Chuang, S.K., Cai, T., Douglass, C.W., Wei, L.J., Dodson, T.B. (2002) Predicting dental implant survival by use of the marginal approach of the semi-parametric survival methods for clustered observations. Journal of Dental Research 84, 54–58.

Chuang, S.K., Cai, T., Douglass, C.W., Wei, L.J., Dodson, T.B. (2005) Frailty approach for the analysis of clustered failure time observations in dental research. Journal of Dental Research 84, 54–58.

Collett, D. (2004) Modelling Survival Data in Medical Research. Chapman & Hall/CRC, London.

Duchateau, L., Janssen, P. (2008) The Frailty Model. Springer, New York.

Hosmer, D.W., Lemeshow, S. & May, S. (2008) Applied Survival Analysis-Regression Modeling of Time to Event Data.

Hanagal, D. (2011) Modeling survival data using frailty models. Chapman & Hall/CRC, London.

Klein, J.P., Moeschberger, M.L. (2003) Survival Analysis - Techniques for Censored and Truncated Data. Springer, New York.

Lee, E.T., & Wang, J.W., (2003) - Statistical Methods for Survival Data Analysis

Li, H., Zhong, X. (2002) Multivariate survival models induced by genetic frailties, with application to linkage analysis. Biostatistics 3, 57–75.

Lin, D.Y., (1994), Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach, Statistics in Medicine, 13, 2233-2247.

Silva, Sabrina Letícia Couto da (2005) Abordagem bayesiana para modelos paramétricos na análise de sobrevivência

Therneau, T.M., Grambsch, P.M. (2000) Modeling Survival Data. Springer, New York.

Vaupel, J., Manton, K., Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16, 439–54.

Wienke, A., (2003) Frailty Models URL: <http://www.demogr.mpg.de/papers/working/wp-2003-032.pdf>, acesso em 29/04/2016.

URL: <https://openaccess.leidenuniv.nl/bitstream/handle/1887/11456/01.pdf?sequence=6> (acessada em 29/04/2016).

APÊNDICES

Apêndice A – Sintaxe para gerar banco de dados

Apêndice B – Ajuste dos modelos usando o programa SAS

Apêndice C – Ajuste dos modelos usando o programa R

Apêndice A – Sintaxe para gerar banco de dados

```

optionsps=58 ls=80 nocenternodatenonumberformchar='|----|+|---+=|-\<>';
%macro centros(centro=, ncentro=, seed1=, seed2=, corr1=, corr2=, corr3=);
* Gera distribuicao dos preditores correlacionados;
    prociml;
        callrandseed(&seed1);
* Especifica a matriz de correlacoes da distribuicao normal multivariada com medias
0 e variancias 1;
        sigma = {1.0 &corr1. &corr2.,
                &corr1. 1.0 &corr3.,
                &corr2. &corr3. 1.0};
* Gera 10000 observacoes da normal multivariada;
    Z = randnormal(&ncentro, {0,0,0,0}, sigma);
        * Se F a f.d. da v.a. X, entao U=F(X)~U(0,1)
        funcao 'cdf' aplica a f.d. da v.a.;
    U= cdf("normal", Z); * As colunas de U saov.a. U(0,1), mas naosao ind.;

* Se U~U(0,1), entao X=invF(U) ~ F;
    X1 = quantile("Uniform", U[,1],0,1); /* X1 ~ Uniform */
    X2 = quantile("Normal", U[,2],0,1); /* X2 ~ Normal */
    X3 = quantile("Normal", U[,3],0,1); /* X3 ~ Normal */
    X = X1||X2||X3;
/* Se Z ~ MVN(0,Sigma), corr(X) geralmente é proxima de Sigma,
em que X=(X1,X2,...,Xm) andX_i = F_i^{-1}(Phi(Z_i)) */
    varNames = ("X1":"X3");
    create XX from X[c=varNames];
    append from X;
    close XX;

run;
quit;

dataFrailty&centro.;
    callstreaminit(&seed2.);

```

```

        frailty = rand("Lognormal");
run;
dataCENTRO&centro.(keep=ID CENTRO GRUPO X2 X3 FRAILITY);
    retain ID CENTRO GRUPO X2 X3 FRAILITY;
    if _n_ = 1 then set Frailty&centro.;
    set XX;
    U = uniform(76849);
    GRUPO = (X1 <= U);
    CENTRO = &centro.;
    ID = CENTRO*100 + _N_;
run;
%mend;
%centros(centro=1, ncentro=65, seed1=31, seed2=33, corr1=0.7, corr2=0.6,
corr3=0.7);
%centros(centro=2, ncentro=45, seed1=32, seed2=36, corr1=0.8, corr2=0.5,
corr3=0.7);
%centros(centro=3, ncentro=45, seed1=33, seed2=63, corr1=0.7, corr2=0.5,
corr3=0.7);
%centros(centro=4, ncentro=50, seed1=34, seed2=43, corr1=0.4, corr2=0.6,
corr3=0.6);
%centros(centro=5, ncentro=61, seed1=35, seed2=73, corr1=0.8, corr2=0.6,
corr3=0.7);
%centros(centro=6, ncentro=60, seed1=36, seed2=93, corr1=0.8, corr2=0.7,
corr3=0.5);
%centros(centro=7, ncentro=55, seed1=37, seed2=31, corr1=0.7, corr2=0.5,
corr3=0.5);
%centros(centro=8, ncentro=35, seed1=21, seed2=34, corr1=0.8, corr2=0.8,
corr3=0.7);
%centros(centro=9, ncentro=60, seed1=51, seed2=37, corr1=0.7, corr2=0.6,
corr3=0.4);
%centros(centro=10, ncentro=40, seed1=351, seed2=133, corr1=0.8, corr2=0.6,
corr3=0.7);
%centros(centro=11, ncentro=50, seed1=371, seed2=233, corr1=0.7, corr2=0.6,
corr3=0.7);

```

```

%centros(centro=12, ncentro=37, seed1=231, seed2=333, corr1=0.5, corr2=0.6,
corr3=0.7);
%centros(centro=13, ncentro=50, seed1=131, seed2=533, corr1=0.8, corr2=0.6,
corr3=0.7);
%centros(centro=14, ncentro=50, seed1=312, seed2=363, corr1=0.7, corr2=0.5,
corr3=0.5);
%centros(centro=15, ncentro=43, seed1=317, seed2=353, corr1=0.7, corr2=0.8,
corr3=0.7);
%centros(centro=16, ncentro=70, seed1=391, seed2=343, corr1=0.7, corr2=0.6,
corr3=0.7);
%centros(centro=17, ncentro=60, seed1=831, seed2=3333, corr1=0.6, corr2=0.6,
corr3=0.7);
%centros(centro=18, ncentro=34, seed1=391, seed2=3243, corr1=0.7, corr2=0.6,
corr3=0.7);
%centros(centro=19, ncentro=40, seed1=441, seed2=83, corr1=0.7, corr2=0.8,
corr3=0.7);
%centros(centro=20, ncentro=50, seed1=231, seed2=8883, corr1=0.5, corr2=0.6,
corr3=0.7);

```

```

data XXF;
    set CENTRO1-CENTRO20;
run;

```

*Com censura;

```

data TEMPO;
    set XXF;
    callstreaminit(10);
    HazardRate = 5;          * Rate at which subject experiences event;
    CensorRate = 10;        * Rate at which subject drops out;
    EndTime = 20;          * End of study period (days);
    linpred = exp(1.2*Grupo - 0.2*X2 + 0.8*X3 - Frailty);
    tEvent = rand("WEIBULL", 1, HazardRate*linpred);      * Time of event;
    rate = HazardRate*linpred;
    c = rand("WEIBULL", 1, CensorRate); * Time of censoring;

```

```
    TEMPO = min(tEvent, c, EndTime);  
    CENSURA = (c <tEvent | tEvent>EndTime);  
run;  
procsgplot data=TEMPO;  
    histogram TEMPO;  
run;  
    procsgplot data=TEMPO;  
    histogramtEvent;  
run;  
    procsgplot data=TEMPO;  
    histogram c;  
run;  
    procfreq data=TEMPO;  
    table CENSURA;  
run;  
    proc phreg data=TEMPO;  
    title1 "Modelo de Cox (com fragilidade)";  
    class CENTRO;  
    model TEMPO*CENSURA(1) = Grupo X2 X3 / rl;  
    random CENTRO / dist=lognormal;  
run;
```

Apêndice B – Ajuste dos modelos usando o programa SAS

```

options ps=58 ls=80 nocenternodatenonumber! formchar='|----|+|----+=|-\<>*' ;
options ls=100 PSIZE=100;
* N POR CENTRO;
proc means data=Exemplo1 n maxdec=0 ;
    class CENTRO;
    var CENSURA;
run;
proc freq data=Exemplo1;
    table CENTRO*CENSURA / nopercntnocol;
run;
proc means data=Exemplo1 (where=(CENSURA=0)) n maxdec=0 ;
    class CENTRO;
    var GRUPO;
run;
proc means data=Exemplo1 meanstdmaxdec=4 ;
    class CENTRO;
    var X2;
run;
proc means data=Exemplo1 mean stdmaxdec=4 ;
    class CENTRO;
    var X3;
run;
proc means data=Exemplo1 n mean stdmaxdec=4;
    class CENTRO;
    var CENSURA Grupo X2 X3 FRAILITY;
run;
proc means data=Exemplo1 mean maxdec=4 ;
    class CENTRO;
    var FRAILITY;
run;
proc means data=Exemplo1 (where=(GRUPO=1)) N maxdec=0;
    class CENTRO;

```

```

var Grupo;

run;
proclifetest data=Exemplo1 method=km;
    time TEMPO*CENSURA(1);
run;
proclifetest data=Exemplo1 method=km;
    time TEMPO*CENSURA(1);
    strata CENTRO;
run;
proclifetest data=Exemplo1 method=km;
    ods select Quartiles;
    ods output Quartiles=teste;
    time TEMPO*CENSURA(1);
run;
%macro Quartiles;
    proc sort data=Exemplo1;
        by CENTRO;
    run;
    %do j=1 %to 20;
    proclifetest data=Exemplo1 method=km;
        where CENTRO = &j.;
        ods select Quartiles;
        ods output Quartiles=QCentro&j.;
        time TEMPO*CENSURA(1);
    run;
    dataMediana&j.(keep=CENTRO MEDIANA LowerLimitUpperLimit);
        retain CENTRO MEDIANA LowerLimitUpperLimit;
        setQCentro&j.;
        CENTRO = &j.;
        MEDIANA = Estimate;
        if Percent = 50;
    run;
    %end;
dataMedianas_Centro;

```

```

        set Mediana1-Mediana20;

run;

%mend;

%Quartiles;

procprint data=Medianas_centro;
    title1 "Medianas do Tempo por Centro";
run;
title1;
procsgplot data=Exemplo1;
    xaxislabel = "Fragilidade";
    yaxis label = "Tempo";
    scatter x = Frailty y = Tempo / group=Grupo;
run;
procsgplot data=Exemplo1;
    xaxislabel = "Fragilidade";
    yaxis label = "Tempo";
    scatter x = Frailty y = Tempo / group=Censura;
run;

proclifetest data=Exemplo1 method=km;
    time TEMPO*CENSURA(1);
run;
proclifetest data=Exemplo1 method=km;
    time TEMPO*CENSURA(1);
    strata CENTRO;
run;
proc phreg data=Exemplo1;
    title1 "Modelo de Cox (sem fragilidade)";
    model TEMPO*CENSURA(1) = Grupo X2 X3 / rl;
run;

*Modelo ajustado por centro (sem fragilidade);
proc phreg data=Exemplo1;

```

```

class CENTRO / PARAM=REF REF=FIRST;
title1 "Modelo de Cox, ajustado por centro (sem fragilidade)";
model TEMPO*CENSURA(1) = Grupo X2 X3 CENTRO / rl;

run;

proc phreg data=Exemplo1;
title1 "Modelo de Cox (com fragilidade lognormal)";
class CENTRO;
model TEMPO*CENSURA(1) = Grupo X2 X3 / rl;
random CENTRO / dist=lognormal solution;
assess PH / resample;

run;

title1;

procnlmixed data=Exemplo1;
bounds gamma > 0; /* Gamma = Weibull shape parameter */
linp = b0 + b1*Grupo + b2*X2 + b3*X3 + z;
alpha = exp(-linp);
G_t = exp(-(alpha*Tempo)**gamma);
g = gamma*alpha*((alpha*Tempo)**(gamma-1))*G_t;
ll = (Censura=0)*log(g) + (Censura=1)*log(G_t);
model Tempo~ general(ll);
random z ~ normal(0,exp(2*logsig)) subject=CENTRO out=EB;
predictG_t out=Gt;
id CENTRO;
estimate 'Grupo: HR (1 x 0)' exp(-b1);
estimate 'Grupo: IC95 HR_Inf' exp(-b1-2.093024054*0.02572);
estimate 'Grupo: IC95 HR_Sup' exp(-b1+2.093024054*0.02572);
estimate 'HR: X2 (Aumento de 1 unidade)' exp(-b2);
estimate 'X2: IC95 HR_Inf' exp(-b2-2.093024054*0.01564);
estimate 'X2: IC95 HR_Sup' exp(-b2+2.093024054*0.01564);
estimate 'HR: X3 (Aumento de 1 unidade)' exp(-b3);
estimate 'X3: IC95 HR_Inf' exp(-b3-2.093024054*0.01605);
estimate 'X3: IC95 HR_Sup' exp(-b3+2.093024054*0.01605);

run;

```

Apêndice C – Ajuste dos modelos usando o programa R

Utilizando o software R para modelar os dados simulados

```

simulacao<- read.table("Dados_Exemplo1_Lognormal.txt", header = T)

simulacao[1:10, ]

require(survival)

# Histogramas e Tabela #

hist(TEMPO, main="Histograma do Tempo", ylab="Frequência")

hist(tEvent)

hist(c)

table(CENSURA)

# Modelo de Cox - Sem Fragilidade#

modelo1<- coxph(Surv(TEMPO, 1- CENSURA) ~ GRUPO + X2 + X3,
data=simulacao)

summary(modelo1)

# Modelo de Cox - Com Fragilidade #

modelo2<- coxph(Surv(TEMPO, 1-CENSURA) ~ GRUPO + X2 + X3 +
frailty(CENTRO, dist = "gauss"), data = simulacao)

summary(modelo2)

## Efeito Aleatório:

# Visualizando, através de um histograma, os efeitos aleatórios estimados
para o modelo com fragilidade.

hist(modelo2$frail, xlim = c(-2,6), prob=T, breaks=6, main = "Fragilidade
Lognormal", xlab="Efeito aleatório", ylab="Frequência")

y=density(modelo2$frail)

```

```
lines(y)

# Dispersão do tempo de acordo com os valores de fragilidade, por grupo.

plot(FRAILITY, TEMPO, xlab="Fragilidade", ylab="Tempo", main="Gráfico do
Tempo vs Fragilidade")

# Risco Gauss

modelo2$frail

exp(modelo2$frail)

## Carregar o item g source("Rfun.r")

# Para o gráfico o modelo precisa ter sido gerado com sparse=T

source("Rfun.r")

modelo2<- coxph(Surv(TEMPO, 1-CENSURA) ~ GRUPO + X2 + X3 +
frailty(CENTRO, sparse = T, dist = "gauss"), data = simulacao)

plot.frail(simulacao$CENTRO,modelo2)

title("Fragilidades estimadas – Lognormal")

### Modelo Paramétrico de sobrevivência ##

model.weibull <- survreg(Surv(TEMPO, 1-CENSURA) ~ GRUPO + X2 + X3 +
frailty(CENTRO), data=simulacao, dist="weibull")

summary(model.weibull)
```