Análise de Sobrevivência em Problemas de Riscos Competitivos

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida pela Srta. Cecília Cândolo e aprovada pela Comissão Julgadora.

Campinas, 27 de dezembro de 1988.

Prof. Dr. José Ferreira de Carvalho

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

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA
DA COMPUTAÇÃO
DEPARTAMENTO DE ESTATÍSTICA

Análise de Sobrevivência em Problemas de Riscos Competitivos

aluna: Cecília Cândolo

orientador: Prof. Dr. José Ferreira de Carvalho

Aos meus pais,
Gaspar e Lúcia.

Agradecimentos

- ao Prof. Dr. José Ferreira de Carvalho pela orientação e confiança,
- à CAPES e FAPESP pelo auxílio financeiro,
- ao CEMEQ pela atenção e disponibilidade,
- aos Professores Dr. Armando Infante, Dr. Carlos Alberto de Bragança Pereira e Dr. Manuel Folledo pelas sugestões,
- à minha família pelo apoio e incentivo,
- à todos os colegas do IMECC e da FEL pela ajuda e amizade
- e à todos aqueles que de alguma forma contribuíram para a realização deste trabalho.

Índice

Capítulo 1

Introdução

Análise de Sobrevivência é o conjunto de técnicas e modelos estatísticos usados para analisar a variável tempo até a ocorrência de um certo evento. Este evento tanto pode ser a falha de um dispositivo eletrônico ou mecânico, como a morte de um ser vivo, ou ainda a execução de uma tarefa.

A importância de análise de sobrevivência em medicina e em estudos de confiabilidade de sistemas é notadamente reconhecida. Como exemplo, podemos considerar o estudo comparativo do efeito de vários tratamentos nos tempos de vida de pacientes portadores de certo tipo de tumor. Outro exemplo: o estudo comparativo do tempo até falha de motores fabricados segundo diversas técnicas, para escolher a mais adequada. Às vezes, em um estudo de análise de sobrevivência ou confiabilidade o objetivo pode ser estimar a função de distribuição do tempo até falha em um único grupo de estudo. Em outras situações, o interesse é comparar os tempos até falha em dois ou mais grupos. Uma outra situação importante, e alvo de muita atenção, é quando dispomos de variáveis explicativas medidas para cada indivíduo em estudo, e queremos estudar a influência destas variáveis no tempo até falha. Por exemplo, o tempo até falha de um componente de certa máquina pode ser influenciado pela temperatura da máquina. A idade no diagnóstico e a contagem de células brancas no sangue são covariáveis candidatas a entrarem num estudo do tempo até cura em crianças com leucemia. Especialmente em ensaios clínicos, é comum ser levado em consideração um número grande de covariáveis. A mesma técnica empregada para análise do efeito de variáveis na sobrevivência pode ser empregada para comparar grupos ou tratamentos, pela introdução de variáveis indica-

doras.

Uma dificuldade especial que surge no tratamento de dados de sobrevivência é a possibilidade de não ser observado, para alguns indivíduos, o tempo completo até o evento de interesse, pois pode acontecer que ao término do tempo estipulado para o experimento nem todos indivíduos, ou componentes, tenham falhado. Ou ainda, em ensaios clínicos, um paciente pode simplesmente abandonar o acompanhamento médico ou falhar por alguma outra causa diferente da que está sendo estudada. Estas observações incompletas do tempo até a falha são chamadas de *censura*. A ocorrência de observações censuradas diferencia a análise de sobrevivência de outras técnicas estatísticas, pois estas observações contém uma informação parcial sobre a variável de interesse, e devem permanecer no estudo.

Vamos supor que Y_i é a variável aleatória que denota o tempo até falha do indivíduo i em uma amostra de tamanho n, e que c_i é o tempo até censura deste mesmo indivíduo. Na prática, o que será observado é

$$Z_i = \min(Y_i, c_i) \ .$$

A censura pode ser do Tipo I quando c_i é um valor constante, c, prédeterminado; do Tipo II quando o estudo termina depois de um número prédeterminado de falhas e assim, c torna-se aleatório. A censura também pode ser aleatória quando em aplicações na medicina, por exemplo, ocorrem: perda de acompanhamento do paciente, desistência do tratamento, morte ou falha por uma causa distinta da que está sendo estudada. Quando a censura é aleatória, uma suposição crucial é que as variáveis Y_i e c_i sejam independentes.

Como vimos até aqui, estamos considerando apenas um fator na determinação da falha, e, quando a falha ocorre por algum outro motivo, a observação é censurada. Entretanto, parece natural em muitos casos estudar o efeito de várias causas concorrentes na determinação da falha.Por exemplo, em um estudo sobre má formação congênita como causa de morte infantil, algumas crianças morrerão por outras causas, tais como tuberculose, insuficiência cardíaca ou diarréia. Estas crianças nem sobreviverão ao primeiro ano de vida e nem morrerão por má formação congênita. Há um efeito competitivo das outras causas na atribuição de má formação congênita como causa de morte. O uso de certos medicamentos em pacientes com câncer pode aumentar o risco de morte por outras causas como

doenças cardiovasculares. Podemos também considerar como causas distintas de falha, em um estudo de pacientes com câncer submetidos a um tratamento: morte pela doença, resposta favorável ao tratamento (cura) e vivos no final do estudo, ainda com a doença. Na análise do tempo até falha de motores, podem ser consideradas várias causas ou riscos competitivos para a falha. É interessante notar que podemos considerar a censura como um risco competitivo, quando temos censura do tipo aleatória.

Uma maneira de analisar estas situações é através da teoria de riscos competitivos, onde consideramos a existência de uma variável aleatória Y_i associada a cada causa de falha i, que denota o tempo de vida de um indivíduo que falha pela causa i, e a variável de interesse é

$$Z = \min_i Y_i$$

O estudo de riscos competitivos começou com o trabalho de Daniel Bernoulli em 1760 sobre o efeito que a erradicação da varíola causaria na estrutura de mortalidade da população. Atualmente podemos extender esta idéia se substituirmos a varíola por câncer, doenças do coração ou, AIDS, entre outras.

Num outro exemplo, vamos abordar o problema da seguinte maneira: supor que um indivíduo está relacionado com um sistema de k componentes ligados em série, onde a falha de um ou mais componentes causa a falha do sistema. Se um destes componentes for melhorado, que efeito isto provocaria no desempenho do sistema? Como podemos estimar a taxa de falha de um componente em particular, na presença de todos outros? Como estimar a taxa de falha de um componente, se quisermos considerá-lo como o único componente do sistema?

Outra questão que surge é a estimação da relação entre covariáveis e taxas de falha por causas específicas.

O estudo da teoria de riscos competitivos já tem resultados genéricos de envergadura suficiente para fazê-los úteis em problemas aplicados. O objetivo deste trabalho é expor a teoria de maneira clara para disseminar seu conhecimento e mostrar sua aplicação em problemas práticos. Para tanto, no Capítulo 2 foram colocadas as definições e notações básicas de análise de sobrevivência e riscos competitivos. O Capítulo 3 trata do problema de estimação de probabilidades de falha e sobrevivência nos casos paramétrico e não-paramétrico, considerando que os dados possam ou não estar agrupados em tabelas de vida. No Capítulo 4 é estudada a influência

de covariáveis no tempo até a falha, através do modelo proposto por Cox em 1972. O Capítulo 5 dá uma abordagem de análise através de modelos lineares tanto para estimação de probabilidades como para verificar relação com covariáveis. Neste trabalho estaremos considerando que os riscos agem independentemente.

O Capítulo 6 traz um exemplo de aplicação com dados reais cuja variável de interesse é o tempo até falha de microcomputadores de 8 bits, considerando 4 causas de falha. Estes dados se restringem a um estudo observacional dos microcomputadores 8 bits da UNICAMP, com objetivo didático de estudar a teoria que está sendo abordada. É nossa intenção extender esta análise para microcomputadores de 16 bits, e, para tanto, já foi iniciada a coleta dos dados.

Capítulo 2

Notação e Definições

2.1 Análise de Sobrevivência

Seja Z a variável aleatória que denota o tempo decorrido entre a entrada de um indivíduo em estudo, até o evento de interesse. Para denotar esse evento, usaremos o termo *falha* , que pode ser, por exemplo, morte, cura ou quebra de um equipamento.

A variável Z é maior ou igual a zero, com função de densidade de probabilidade, f.d.p., $f_Z(x)$ e função de distribuição acumulada, FDA, $F_Z(x)$. Assim, a *função de sobrevivência* é definida por

$$\overline{F}_Z(x) = 1 - F_Z(x) = P(Z > x). \tag{2.1}$$

A *Taxa de Falha* é

$$\lambda_Z(x) = \frac{f_Z(x)}{\overline{F}_Z(x)}, \tag{2.2}$$

ou seja, a probabilidade instantânea de falha em x, dado que sobreviveu até x:

$$\lambda_Z(x) = \lim_{\Delta \to 0} \frac{P(x < Z < x + \Delta | Z > x)}{\Delta} .$$

2.2 Riscos Competitivos

Suponha-se que K causas de falha estejam agindo simultaneamente sobre uma população ψ. Cada indivíduo desta população falhará devido a uma

das K causas de falha. Os termos *riscos* e *causas* determinam a mesma condição; a diferença é que *risco* é a condição antes de ocorrer a falha, e *causa* é a condição depois de ocorrida a falha.

Seja Y_l a variável aleatória que denota o tempo de vida de um indivíduo sujeito a um único risco C_l, $l = 1, 2, \ldots, K$. Na presença simultânea de K riscos, temos K valores hipotéticos para Y, isto é, para cada indivíduo, criamos valores hipotéticos de seu tempo de vida, relativos a cada um dos K riscos, mas será observado apenas o menor destes valores.

Temos então $Y_1, \ldots, Y_K$ variáveis aleatórias para cada indivíduo, mas observaremos apenas a menor delas, $\min_l Y_l$, e a causa de falha correspondente. Assim, o tempo de vida observável será $\min_l Y_l$, ou seja, a variável aleatória Z, que denota o tempo de vida de um indivíduo até a ocorrência da falha, será

$$Z = \min_l Y_l \ ,$$

e sejam $P_i(x)$ e $p_i(x)$, a FDA de Y_i e a fdp de Y_i, respectivamente, para $i = 1, \ldots, K$ e, F_Z e f_Z o análogo para Z.

A Função de Sobrevivência de Z é

$$\overline{F}_Z(x) = 1 - F_Z(x) = P(Z > x).$$

Como $Z = \min_l Y_l$, temos:

$$\overline{F}_Z(x) = P(Z > x) = P(\min_l Y_l > x) = P(Y_1 > x, \ldots, Y_K > x). \qquad (2.3)$$

A probabilidade instantânea de falha em $(x, x + \Delta)$ ou função de falha, ou ainda, taxa de falha de Z, é $\lambda_Z(x)$:

$$\lambda_Z(x) = \lim_{\Delta \to 0} \frac{P(x < Z < x + \Delta | Z > x)}{\Delta} = \frac{f_Z(x)}{\overline{F}_Z(x)}. \qquad (2.4)$$

Integrando $\lambda_Z(x)$ temos:

$$\int_0^x \lambda_Z(t)dt = \int_0^x \frac{f_Z(t)dt}{\overline{F}_Z(t)} dt = [-\log \overline{F}_Z(t)]_0^x = -\log \overline{F}_Z(x). \qquad (2.5)$$

Logo

$$\overline{F}_Z(x) = \exp[-\int_0^x \lambda_Z(t)dt] \ . \qquad (2.6)$$

Vamos denotar a taxa de falha pela causa C_i, na presença dos K riscos, por $\lambda_i(x)$. Como $\lambda_Z(x)$ é a taxa de falha por qualquer uma das K causas, supondo $Y_1, \ldots, Y_k$ independentes, temos que

$$\lambda_Z(x) = \sum_{i=1}^{K} \lambda_i(x) \ , \tag{2.7}$$

e ainda

$$P(Z > x) = P(Y_1 > x) \ldots P(Y_K > x) \ ,$$

isto é,

$$\bar{F}_Z(x) = \prod_{i=1}^{K} \bar{P}_i(x) \tag{2.8}$$

e a taxa de falha pela causa C_i fica

$$\lambda_i(x) = \frac{p_i(x)}{\bar{P}_i(x)} \tag{2.9}$$

pois,

$$\begin{aligned}
\lambda_i(x) &= \lim_{\Delta \to 0} \frac{P(x < Y_i < x + \Delta | Y_1 > x \ldots Y_K > x)}{\Delta} \\
&= \frac{p_i(x) \prod_{\substack{j=1 \\ j \neq i}}^{K} P(Y_j > x)}{\bar{F}_Z(x)} \\
&= \frac{p_i(x) \prod_{\substack{j=1 \\ j \neq 1}}^{K} P(Y_j > x)}{\prod_{j=1}^{K} P(Y_j > x)} \\
&= \frac{p_i(x)}{\bar{P}_i(x)} \ .
\end{aligned}$$

Podemos chamar as variáveis aleatórias $Y_1, \ldots, Y_K$ de tempos de vida potenciais ou tempos de vida *líquidos*.Definamos variáveis aleatórias $X_1, \ldots,$ X_K , onde $X_i = Y_i | Y_i = \min_l Y_l$ para i e $l = 1, \ldots, K$, que chamaremos de tempos de vida *brutos*, e sejam F_i a FDA e f_i a f.d.p. de X_i . A variável Y_i é o tempo de vida para o risco i, supondo que este é o único risco agindo na população, que na prática é não observ'avel, e a variável

X_i é o tempo de vida para o risco i quando todos os riscos estão agindo na população, ou seja, a variável observável. Temos então, os seguintes tipos de probabilidades de falha a considerar:

Probabilidade Bruta Probabilidade de falha por uma causa específica, na presença de todos riscos agindo na população:

$Q_i(a,b)$ = P(um indivíduo vivo no tempo a falhar no intervalo (a,b) pela causa C_i, na presença de todos outros riscos)=

$$=\mathrm{P}(a < X_i < b | X_i > a).$$

Probabilidade Líquida Probabilidade de falha por uma causa específica se esta é a única presente na população:

$q_i(a,b)$ = P(um indivíduo vivo em a falhar no intervalo (a,b) se C_i é o único risco)=

$$=\mathrm{P}(a < Y_i < b | Y_i > a).$$

E também,

Probabilidade Bruta Parcial Probabilidade de falhar por uma causa específica quando outro risco (ou riscos) é eliminado como risco de falha:

$Q_{i,j}(a,b)$ = P(um indivíduo vivo em a falhar no intervalo (a,b) por C_i se C_j é eliminado como risco de falha).

Quando a causa de falha não é especificada, temos:

q(a,b) = P(um indivíduo vivo em a falhar no intervalo (a,b))=

$$=\mathrm{P}(a < Z < b | Z > a).$$

As probabilidades de sobrevivência respectivas são:

$$P_i(a,b) = 1 - Q_i(a,b)$$

$$p_i(a,b) = 1 - q_i(a,b)$$

$$p(a,b) = 1 - q(a,b).$$

As probabilidades Bruta, Líquida e Bruta Parcial podem ser escritas em termos das funções $\lambda_Z(x)$ e $\lambda_i(x)$. Considerando falha sem especificação de causa, a probabilidade $p(a,b)$ é, por (2.6),

$$p(a,b) = \exp\left[- \int_a^b \lambda_Z(t)dt\right].$$

Para deduzirmos $Q_i(a,b)$, consideramos um ponto x dentro do intervalo (a,b) e temos

$$P(\text{indivíduo vivo em } a \text{ falhar por } C_i \text{ no intervalo } (x, x+\Delta)) =$$

$$= \exp\left[- \int_a^x \lambda_Z(t)dt\right]\lambda_i(x)dx \ ,$$

onde a função exponencial é a probabilidade de sobreviver de a até x, quando todos os riscos estão agindo, e o fator $\lambda_i(x)dx$ é a probabilidade de falha instantânea pela causa C_i no intervalo $(x, x+\Delta)$. Somando a expressão acima para todos valores possíveis de x, com $a < x < b$, temos

$$Q_i(a,b) = \int_a^b \exp\left[- \int_a^x \lambda_Z(t)dt\right]\lambda_i(x)dx. \tag{2.10}$$

A probabilidade líquida de falha é

$$q_i(a,b) = 1 - \exp\left[- \int_a^b \lambda_i(t)dt\right] \tag{2.11}$$

e a probabilidade bruta parcial de falha quando o risco C_j é eliminado, é dada por

$$Q_{i,j}(a,b) = \int_a^b \exp\left[- \int_a^x \lambda_Z^j(t)dt\right]\lambda_i^j(x)dx \tag{2.12}$$

onde $\lambda_i^j(x)$ e $\lambda_Z^j(t)$ são, respectivamente, a taxa de falha por C_i e a taxa de falha de Z depois da eliminação de C_j como risco de falha.

Quando os riscos são independentes, temos

$$\lambda_i^j(x) = \lambda_i(x)$$

$$\lambda_Z^j(t) = \lambda_Z(t) - \lambda_j(t),$$

e então

$$Q_{i,j}(a,b) = \int_a^b \exp\left[- \int_a^x (\lambda_Z(t) - \lambda_j(t))dt\right]\lambda_i(x)dx \ . \tag{2.13}$$

Capítulo 3

Estimação

Neste capítulo, trataremos do problema da estimação das probabilidades vistas no capítulo anterior. Abordaremos o caso paramétrico e o caso não-paramétrico. No primeiro, o problema é a estimação dos parâmetros da função de probabilidade do tempo de vida através do método da máxima verossimilhança. No segundo estudaremos métodos não paramétricos via taxas de falha proporcionais e o estimador Kaplan-Meier adaptado a riscos competitivos.

3.1 Estimação em Modelos Paramétricos

3.1.1 Riscos Independentes

Nossas variáveis são

$$\begin{aligned} Y_i &= \text{tempo de vida líquido para a causa } C_i\,, \quad i = 1, \ldots, K \\ \text{e} & \\ X_i &= Y_i | \, Y_i = \min_l Y_l, \quad l = 1 \ldots, K\,. \end{aligned}$$

Estamos considerando que cada Y_i tem uma f.d.p. $p_i(x)$ conhecida, mas não conhecemos a f.d.p. da variável observável X_i , $f_i(x)$, e vamos determiná-la através das $p_i(x)$'s. Nesta seção trataremos o caso de $Y_i's$ independentes, ou seja, vamos supor que as causas de falha independam entre si.

Seja π_i a probabilidade de falha pela causa C_i:

$$\pi_i = P(Y_i = \min_l Y_l),$$

com $\pi_i > 0$ e $\sum_{i=1}^{k} \pi_i = 1$.

De David e Moeschberger (1978) temos:

$$f_i(x)dx = \frac{1}{\pi_i} P(x - dx < Y_i \leq x) \prod_{\substack{l=1 \\ l \neq i}}^{k} P(Y_l > x) \ .$$

Então, a função densidade de probabilidade da variável X_i fica:

$$f_i(x) = \frac{1}{\pi_i} p_i(x) \prod_{\substack{l=1 \\ l \neq i}}^{k} \bar{P}_l(x) \ . \tag{3.1}$$

Vamos, agora, construir a função de verossimilhança. A amostra consiste de n indivíduos onde n_1 falham por C_1,..., n_k falham por C_k. Então temos:

N_i = número de indivíduos que falham pela causa C_i

e

X_{ij} = tempo de vida do j-ésimo indivíduo que falha pela causa C_i ,

com $j = 1, \ldots, n_i$ e $i = 1, \ldots, k$.

Seja $f(\mathbf{X}, \boldsymbol{\Theta})$ a distribuição conjunta dos X_{ij}, onde

$\mathbf{X}$ = $(x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k})$ = vetor de observações,

e

$\boldsymbol{\Theta}$ = $(\Theta_1, \ldots, \Theta_k)$ = vetor de parâmetros a serem estimados.

Note que Θ_i pode ter dimensão maior que um, se a f.d.p. do risco i tiver mais que um parâmetro.

Se $\mathbf{X}$ é composto por uma amostra aleatória, temos, por (3.1)

$$f(\mathbf{X}, \boldsymbol{\Theta}) = \prod_{i=1}^{k} \frac{1}{\pi_i^{n_i}} \prod_{j=1}^{n_i} p_i(x_{ij}, \Theta_i) \prod_{\substack{l=1 \\ l \neq i}}^{k} \bar{P}_l(x_{ij}, \Theta_i) \ . \tag{3.2}$$

Como $\mathbf{N} = (N_1, ..., N_k)$ tem distribuição multinomial com

$$f(n_1, ..., n_k; n) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} \pi_i^{n_i} \ ,$$

a função de verossimilhança de interesse fica

$$L(\Theta) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} \prod_{j=1}^{n_i} p_i(x_{ij}, \Theta_i) \prod_{\substack{l=1 \\ l \neq i}} \bar{P}_l(x_{ij}, \Theta_i) \ . \tag{3.3}$$

Para facilitar a estimação dos Θ_i , particionamos a função L da seguinte maneira:

$$L(\Theta) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} L_i \ , \tag{3.4}$$

$$L_i = (\prod_{j=1}^{n_i} p_i(x_{ij}, \Theta_i))(\prod_{\substack{l=1 \\ l \neq i}}^{k} \prod_{j=1}^{n_l} \bar{P}_i(x_{lj}, \Theta_i)) \ . \tag{3.5}$$

Desta forma, a estimação pode ser realizada individualmente para cada risco, ao maximizarmos L_i com respeito a Θ_i .

Como ilustração, vejamos alguns exemplos.

Exemplo 1

Considere que os tempos de vida teóricos $Y_1, \ldots, Y_k$, ou seja, os tempos de vida supondo cada risco como o único agindo na população, têm distribuição exponencial:

$$Y_i \sim \exp(1/\Theta_i).$$

Então

$$p_i(x_{ij}, \theta_i) = (1/\theta_i) \exp(-x_{ij}/\theta_i) \ ,$$

$$P_i(x_{ij}, \theta_i) = 1 - \exp(-x_{ij}/\theta_i) \quad \text{e} \quad \bar{P}_i(x_{ij}, \theta_i) = \exp(-x_{ij}/\theta_i) \ .$$

A função de verossimilhança, L_i, fica:

$$\begin{aligned} L_i &= \Big(\prod_{j=1}^{n_i}(1/\theta_i)\exp(-x_{ij}/\theta_i)\Big)\Big(\prod_{\substack{l=1\\ l\neq i}}^{k}\prod_{j=1}^{n_l}\exp(-x_{lj}/\theta_i)\Big) \\ &= \theta_i^{-n_i}\exp\Big(\sum_{j=1}^{n_i}x_{ij}/\theta_i\Big)\exp\Big(-\sum_{\substack{l=1\\ l\neq i}}^{k}\sum_{j=1}^{n_l}x_{ij}/\theta_i\Big) \\ &= \theta_i^{-n_i}\exp\Big(-\sum_{l=1}^{k}\sum_{j=1}^{n_l}x_{ij}/\theta_i\Big) \ . \end{aligned}$$

Queremos o valor de θ_i que maximiza a função L_i. A obtenção deste valor está a seguir:

$$\log L_i = n_i \log\theta_i - \frac{1}{\theta_i}\sum_{l=1}^{k}\sum_{j=1}^{n_l}x_{lj}$$

$$\frac{\partial \log L_i}{\partial\theta_i} = \frac{-n_i}{\theta_i} + \frac{1}{\theta_i^2}\sum_{l=1}^{k}\sum_{j=1}^{n_l}x_{lj} = 0$$

$$\Longrightarrow \hat{\theta}_i = \sum_{l=1}^{k}\sum_{j=1}^{n_l}\frac{x_{lj}}{n_i} \ .$$

Então, $\hat{\theta}_i$ é o estimador do parâmetro θ_i para a causa C_i, com $i = 1,\ldots,k$.

Exemplo 2

Seja $Y_1,\ldots,Y_k$ com distribuição de Weibull de dois parâmetros:

$$Y_i \sim W(\Theta_i), \quad \text{onde } \Theta_i = (\theta_{i1},\theta_{i2}).$$

Então:

$$p_i(x_{ij},\Theta_i) = \frac{\theta_{i2}x_{ij}^{\theta_{i2}-1}}{\theta_{i1}}\exp\Big(\frac{-x_{ij}^{\theta_{i2}}}{\theta_{i1}}\Big), \ \text{para } x_{ij} > 0 \ ,$$

$$P_i(x_{ij},\Theta_i) = 1 - \exp\Big(\frac{-x_{ij}^{\theta_{i2}}}{\theta_{i1}}\Big)$$

e

$$\bar{P}_i(x_{ij}, \Theta_i) = \exp(\frac{-x_{ij}^{\theta_{i2}}}{\theta_{i1}}) \ .$$

A função de verossimilhança fica:

$$\begin{aligned} L_i &= \prod_{j=1}^{n_i} \frac{\theta_{i2} x_{ij}^{\theta_{i2}-1}}{\theta_{i1}} \exp(\frac{-x_{ij}^{\theta_{i2}}}{\theta_{i1}}) \prod_{\substack{l=1 \\ l\neq i}}^{k} \prod_{j=1}^{n_l} \exp(\frac{-x_{ij}^{\theta_{i2}}}{\theta_{i1}}) \\ &= (\frac{\theta_{i2}}{\theta_{i1}})^{n_i} (\prod_{j=1}^{n_i} x_{ij})^{\theta_{i2}-1} \exp(-\frac{1}{\theta_{i1}} \sum_{j=1}^{n_i} x_{ij}^{\theta_{i2}}) \exp(-\frac{1}{\theta_{ij}} \sum_{\substack{l=1 \\ l\neq i}}^{k} \sum_{l=1}^{k} x_{lj}^{\theta_{i2}}) \\ &= (\frac{\theta_{i2}}{\theta_{i1}})^{n_i} (\prod_{j=1}^{n_i} x_{ij})^{\theta_{i2}-1} \exp(-\frac{1}{\theta_{i1}} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{ij}^{\theta_{i2}}) \end{aligned}$$

$$\log L_i = n_i \log \theta_{i2} - n_i \log \theta_{i1} + (\theta_{i2} - 1) \sum_{j=1}^{n_i} \log x_{ij} - \frac{1}{\theta_i} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}}$$

$$\begin{aligned} \frac{\partial \log L_i}{\partial \theta_{i1}} &= -\frac{n_i}{\theta_{i1}} + \frac{1}{\theta_{i1}^2} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \\ \frac{\partial \log L_i}{\partial \theta_{i2}} &= \frac{n_i}{\theta_{i2}} + \sum_{j=1}^{n_i} \log x_{ij} - \frac{1}{\theta_{i1}} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \log x_{lj} \ . \end{aligned}$$

O estimador de máxima verossimilhança de $\Theta_i = (\theta_{i1}, \theta_{i2})$ é obtido a partir da resolução do sistema:

$$\begin{cases} -\frac{n_i}{\theta_{i1}} + \frac{1}{\theta_{i1}^2} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} = 0 \\ \frac{n_i}{\theta_{i2}} + \sum_{j=1}^{n_i} \log x_{ij} - \frac{1}{\theta_{i1}} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \log x_{lj} = 0 \ . \end{cases}$$

Para acharmos a solução deste sistema aplicamos o método iterativo de Newton-Raphson:

$$\boldsymbol{\Theta}_{n+1} = \boldsymbol{\Theta}_n - \mathbf{J}_{\boldsymbol{\Theta}_n}^{-1} \mathbf{F}_{\boldsymbol{\Theta}_n}$$

onde:

$\boldsymbol{\Theta}$: vetor dos parâmetros a serem estimados

$\mathbf{F}$: vetor composto pelas duas equações do sistema

J: matriz das derivadas parciais de **F** com respeito a Θ.

No nosso exemplo:

$$\Theta = \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix},$$

$$\mathbf{F} = \begin{pmatrix} -\frac{n_i}{\theta_{i1}} + \frac{1}{\theta_{i1}^2} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \\ \frac{n_i}{\theta_{i2}} + \sum_{j=1}^{n_i} \log x_{ij} - \frac{1}{\theta_{i1}} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \log x_{lj} \end{pmatrix}$$

e

$$\mathbf{J} = \begin{pmatrix} \frac{n_i}{\theta_{i1}^2} - \frac{1}{\theta_{i1}^4} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} & \frac{1}{\theta_{i1}^2} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \log x_{lj} \\ \frac{1}{\theta_{i1}^2} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} \log x_{lj} & -\frac{n_i}{\theta_{i2}^2} - \frac{1}{\theta_{i1}} \sum_{l=1}^{k} \sum_{j=1}^{n_l} x_{lj}^{\theta_{i2}} (\log x_{lj})^2 \end{pmatrix}.$$

Um programa SAS que resolve este sistema está no Apêndice A.

3.1.2 Riscos Dependentes

Nem sempre os riscos aos quais a população está exposta agem independentemente. Uma maneira de obter independência é através do agrupamento de riscos em categorias, dentro das quais os riscos são relacionados e assim provavelmente dependentes, mas, entre as quais, espera-se pouca ou nenhuma dependência. Contudo, dependendo do problema, este agrupamento de informações pode não ser de interesse. Então, é necessário estudar os riscos, sem a suposição de independência.

Vamos considerar, novamente, os $Y_1, \ldots, Y_K$ tempos de vida teóricos e

$$\begin{aligned} X_i &= Y_i | Y_i = \min_l Y_l \ , \\ \pi_i &= P(Y_i = \min_l Y_l), \ \text{com } i = 1, \ldots, k \ . \end{aligned}$$

Seja $p(y_1, ..., y_k)$ a função distribuição conjunta absolutamente contínua dos Y_i. A partir dela, obtemos a função densidade de probabilidade de X_i, $f_i(x)$:

$$f_i(x) = \frac{1}{\pi_i} p_i(x) \int_x^\infty \cdots \int_x^\infty p(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_k | Y_i = x) \prod_{\substack{l=1 \\ l \neq i}}^{k} dy_l \ ,$$

pois

$$p(y_1,\ldots,y_{i-1},x,y_{i+1},\ldots,y_k) =$$
$$p_i(x)p(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_k|Y_i=x) \ .$$

Se N_i indivíduos falham pela causa C_i, e X_{ij} denota o tempo de vida do j-ésimo indivíduo que falha pela causa C_i, $j=1\ldots,n_i$ e $i=1,\ldots,K$, temos a seguinte f.d.p. conjunta para X_{ij}:

$$\begin{aligned} f(\mathbf{X},\Theta) &= \prod_{i=1}^{k}\frac{1}{\pi_i}\prod_{j=1}^{n_i}p_i(x_{ij},\Theta_i) \ \times \\ &\times \int_{x_{ij}}^{\infty}\cdots\int_{x_{ij}}^{\infty}p(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_k|Y_i=x_{ij})\prod_{\substack{l=1\\l\neq i}}dy_l \ , \end{aligned}$$

onde $\mathbf{X}$ e Θ são como na seção 3.1.1. Então a função de verossimilhança fica

$$\begin{aligned} L &= \frac{n!}{\prod_{i=1}^{k}n_i!}\prod_{i=1}^{k}\prod_{j=1}^{n_i}p_i(x_{ij},\Theta_i) \times \\ &\times \int_{x_{ij}}^{\infty}\cdots\int_{x_{ij}}^{\infty}p(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_k \mid Y_i=x_{ij})\prod_{\substack{l=1\\l\neq i}}^{k}dy_l \ , \end{aligned}$$

ou

$$L \propto \prod_{i=1}^{k}L_i \ , \quad L_i=\prod_{j=1}^{n_i}\pi_i f_i(x_{ij},\Theta_i) \ .$$

Deste modo, o estimador de máxima verossimilhança de Θ_i é aquele que maximiza L_i .

Moeschberger (1974) trata o problema de riscos dependentes e apresenta dois exemplos de estimação de parâmetros através de Máxima Verossimilhança: distribuição Weibull bivariada e distribuição Normal bivariada.

3.2 Taxas de Falha Proporcionais: Estimação de Tabelas de Vida

Na seção anterior vimos o caso paramétrico onde, dado o modelo probabilístico, podemos estimar as probabilidades bruta e líquida estando os dados agrupados ou não. Nesta seção veremos a estimação destas probabilidades sem conhecer o modelo probabilístico e quando os dados estão grupados em intervalos, ou seja a estimação de tabelas de vida. A técnica que estudaremos supõe riscos independentes e supõe também que as taxas de falha sejam proporcionais.

3.2.1 Taxas de falha Proporcionais

Temos Y_i, $i = 1, \ldots, k$ tempos de vida líquidos independentes com

$$Y_i \sim p_i(x), \quad p_i(x) \text{ desconhecida.}$$

Como definido no Capítulo 2, seja

$$Z = \min_l Y_l, \quad X_i = Y_i | Y_i = \min_l Y_l \quad \text{e} \quad \pi_i = P(Y_i = \min_l Y_l) .$$

Vimos que

$$\begin{aligned} f_i(x) &= \frac{1}{\pi_i} p_i(x) \prod_{\substack{l=1 \\ l \neq i}}^{k} \bar{P}_l(x) \\ &= \frac{1}{\pi_i} \lambda_i(x) \bar{F}_Z(x) \end{aligned}$$

e também

$$\lambda_Z(x) = \frac{f_Z(x)}{\bar{F}_Z(x)}, \quad \lambda_i(x) = \frac{p_i(x)}{\bar{P}_i(x)}$$

$$\text{com } \lambda_Z(x) = \sum_{i=1}^{k} \lambda_i(x).$$

Dizemos que as taxas de falha são proporcionais se existem constantes γ_i, tais que

$$\frac{\lambda_i(x)}{\lambda_Z(x)} = \gamma_i \ , \tag{3.6}$$

ou seja, a razão entre as taxas de falha devidas a cada risco i e a taxa de risco global depende apenas de i e não de x.

Esta proporcionalidade entre as taxas tem as seguintes consequências:

1. Se as taxas são proporcionais, a constante de proporcionalidade, γ_i, é igual à probabilidade de falha pela causa i, π_i, isto é

$$\gamma_i = \pi_i$$

Prova

$$\begin{aligned} \pi_i &= \int_o^\infty \bar{F}_Z(x)\lambda_i(x)dx = \int_o^\infty \bar{F}_Z(x)\lambda_Z(x)\gamma_i dx = \\ &= \gamma_i \int_o^\infty \lambda_Z \bar{F}_Z dx = \gamma_i \int_o^\infty f_Z(x)dx = \\ &= \gamma_i \ . \ \diamond \end{aligned}$$

2. Se as taxas são proporcionais, a probabilidade de falha pela causa i, π_i, é constante no tempo, ou seja,

$$P(\text{falha por } C_i | \text{falha no intervalo } (a,b)) = \pi_i \quad ,$$

independe do intervalo.

Prova

$$P(\text{falha por } C_i | \text{falha em } (a,b)) = \frac{\int_a^b \bar{F}_Z(x)\lambda_i(x)dx}{\int_a^b f_Z(x)dx} =$$

$$= \gamma_i \frac{\int_a^b f_Z(x)dx}{\int_a^b f_Z(x)dx} = \gamma_i = \pi_i \ . \ \diamond$$

A volta também vale: Se a probabilidade de falha por C_i é constante no tempo, então as taxas de risco são proporcionais.

Prova

Se P(falha por C_i|falha em (a,b))=π_i temos

$$\frac{\int_a^b f_Z(x)\lambda_i(x)dx}{\int_a^b f_Z(x)dx} = \pi_i$$

$$\int_a^b \bar{F}_Z(x)\lambda_i(x)dx = \pi_i \int_a^b f_Z(x)dx$$

$$\frac{1}{\pi_i}\int_a^b \bar{F}_Z(x)\lambda_i(x)dx = \int_a^b f_Z(x)dx$$

$$\text{mas } f_i(x) = \frac{1}{\pi_i}\lambda_i(x)\bar{F}_Z(x) \text{ então}$$

$$\frac{1}{\pi_i}\int_a^b \bar{F}_Z(x)\lambda_i(x)dx = \int_a^b f_i(x)dx = \int_a^b f_Z(x)dx \ ,$$

$$\text{então} \quad F_i(b) - F_i(a) = F_Z(b) - F_Z(a) \ .$$

Como a e b podem ser quaisquer, dentro dos valores possíveis de Z, temos que

$$F_i(x) = F_Z(x) \text{ e portanto } \frac{dF_i(x)}{dx} = \frac{dF_Z(x)}{dx}$$

$$\Rightarrow \quad f_i(x) = f_Z(x) \ ,$$

$$\text{então } \frac{1}{\pi_i}\lambda_i(x)\bar{F}_Z(x) = \lambda_Z(x)\bar{F}_Z(x)$$

$$\Rightarrow \quad \frac{\lambda_i(x)}{\lambda_Z(x)} = \pi_i \ ,$$

ou seja, as taxas são proporcionais. $\diamondsuit$

3. Se as taxas são proporcionais, a f.d.p. de X_i é a mesma de Z, isto é, não depende de i:

$$f_i(x) = f_Z(x)$$

Prova

$$f_i(x) = \underbrace{\frac{1}{\pi_i}\lambda_i(x)}_{\lambda_Z(x)} \bar{F}_Z(x) = \lambda_Z(x)\bar{F}_Z(x) = f_Z(x) \ . \ \diamondsuit$$

A volta também vale: se as f.d.p's dos X_i's não dependem de i, então as taxas de falha são proporcionais.

Prova

Como estamos considerando riscos independentes, a f.d.p. de Z pode ser escrita em função das fdp's dos X_i's da seguinte maneira:

$$f_Z(x) = \sum_{i=1}^{k} \pi_i f_i(x),$$

e então, se $f_1(x) = \cdots = f_k(x)$, temos

$$f_Z(x) = f_i(x), \ \text{ pois } \sum_{i=1}^{k} \pi_i = 1.$$

Como

$$f_i(x) = \frac{1}{\pi_i}\lambda_i(x)\bar{F}_Z(x) \ \text{ e } f_Z(x) = \lambda_Z(x)\bar{F}_Z(x) \ ,$$

temos

$$\frac{1}{\pi_i}\lambda_i(x)\bar{F}_Z(x) = \lambda_Z(x)\bar{F}_Z(x)$$

$$\Rightarrow \quad \frac{\lambda_i(x)}{\lambda_Z(x)} = \pi_i,$$

ou seja, as taxas são proporcionais. $\diamondsuit$

4. Se as variáveis Y_i são independentes, temos taxas de falha proporcionais se, e só se, existem constantes $\gamma_i's > 0$ tais que

$$\bar{F}_Z(x) = [\bar{P}_i(x)]^{1/\gamma_i} \ .$$

Prova

($\Longrightarrow$) Se as taxas de falha são proporcionais temos

$$\frac{\lambda_i(x)}{\lambda_Z(x)} = \pi_i.$$

Pela independência dos Y_i's,

$$\lambda_i(x) = \frac{p_i(x)}{\bar{P}_i(x)} \ .$$

Então,

$$\frac{\lambda_i(x)}{\lambda_Z(x)} = \frac{p_i(x)/\bar{P}_i(x)}{f_Z(x)/\bar{F}_Z(x)} = \pi_i$$

$$\Rightarrow \quad \frac{p_i(x)}{\bar{P}_i(x)} = \pi_i \frac{f_Z(x)}{\bar{F}_Z(x)}.$$

Calculando a integral nos dois lados, temos

$$\ln \bar{P}_i(x) = \pi_i \ln \bar{F}_Z(x)$$

$$\Rightarrow \quad [\bar{P}_i(x)]^{1/\pi_i} = \bar{F}_Z(x).$$

($\Longleftarrow$) Se

$$\bar{F}_Z(x) = [\bar{P}_i(x)]^{1/\pi_i}$$

temos

$$\ln \bar{P}_i(x) = \pi_i \ln \bar{F}_Z(x),$$

derivando, fica:

$$\frac{p_i(x)}{\bar{P}_i(x)} = \pi_i \frac{f_Z(x)}{\bar{F}_Z(x)}$$

$$\Rightarrow \quad \frac{\lambda_i(x)}{\lambda_Z(x)} = \pi_i \ . \quad \diamondsuit$$

Verificação de Taxas de Falha Proporcionais

A validade do suposto de proporcionalidade pode ser avaliada por verificar uma das consequências vistas acima. No caso paramétrico, isto é, quando conhecemos as $p_i(x)$'s, a verificação é direta. Vamos ver dois exemplos:

Exemplo 1

Sejam $Y_1, \ldots, Y_k$ independentes com distribuição exponencial:

$$Y_i \sim \exp(1/\theta_i),$$

então

$$p_i(x) = \frac{1}{\theta_i} \exp(-x/\theta_i) \ \text{ e } \ \bar{P}_i(x) = \exp(-x/\theta_i) \ .$$

$$\lambda_i(x) = \frac{p_i(x)}{\bar{P}_i(x)} = \frac{\frac{1}{\theta_i} \exp(-x/\theta_i)}{\exp(-x/\theta_i)} = \frac{1}{\theta_i} \ ,$$

$$\lambda_Z(x) = \sum_{l=1}^{k} \frac{1}{\theta_l} \ .$$

Então,

$$\frac{\lambda_i(x)}{\lambda_Z(x)} = \frac{\frac{1}{\theta_i}}{\sum_{l=1}^{k} \frac{1}{\theta_l}} = \gamma_i \ ,$$

não depende de x, apenas de i. Logo, quando $Y_1, \ldots, Y_k$ têm distribuição exponencial, as taxas de falha são proporcionais.

Exemplo 2

Sejam $Y_1, \ldots, Y_k$ independentes com distribuição Weibull:

$$Y_i \sim W(\theta_{i1}, \theta_{i2}),$$

então

$$p_i(x) = \frac{\theta_{i2} x^{\theta_{i2}-1}}{\theta_{i1}} \exp(-x^{\theta_{i2}}/\theta_{i1}) \text{ e } \bar{P}_i(x) = \exp(-x^{\theta_{i2}}/\theta_{i1}) ,$$

$$\lambda_i(x) = \frac{\theta_{i2} x^{\theta_{i2}-1}}{\theta_{i1}} ,$$

$$\lambda_Z(x) = \sum_{l=1}^{k} \frac{\theta_{l2} x^{\theta_{l2}-1}}{\theta_{l1}} .$$

Então

$$\frac{\lambda_i(x)}{\lambda_Z(x)} = \frac{\theta_{i2} x^{\theta_{i2}-1}}{\theta_{i1} \sum_{l=1}^{k} \frac{\theta_{l2} x^{\theta_{l2}-1}}{\theta_{l1}}} ,$$

que depende de x.

Se o parâmetro da forma for igual para todo i, isto é,

$$Y_i \sim W(\theta_i, \alpha),$$

temos

$$\lambda_i(x) = \frac{\alpha x^{\alpha-1}}{\theta_i} \text{ e } \lambda_Z(x) = \alpha x^{\alpha-1} \sum_{l=1}^{k} \frac{1}{\theta_l} ,$$

então

$$\frac{\lambda_i(x)}{\lambda_Z(x)} = \frac{1}{\theta_i \sum_{l=1}^{k} \frac{1}{\theta_l}} = \gamma_i .$$

Portanto, quando $Y_1, \ldots, Y_k$ têm distribuição Weibull do tipo $W(\theta_i, \alpha)$, as taxas de falha são proporcionais.

No caso não paramétrico podemos verificar se as taxas são proporcionais através da consequência 3, realizando um teste não paramétrico para testar se $f_i(x) = f_z(x)$, isto é, se as f.d.p.'s dos X_i's não dependem de i. Uma maneira simples de fazer isto é através de gráficos probabilísticos. Vamos apresentar aqui o gráfico probabilístico tipo P-P.

Gráfico P-P

O gráfico P-P consiste em traçar o gráfico da função distribuição acumulada empírica de X_i , F_i^E , contra a de X_j , F_j^E , para $i \neq j = 1, \ldots, k$. Se X_i e X_j são identicamente distribuídas, para qualquer quantil q temos $F_i^E(q) = F_j^E(q)$, e portanto o gráfico de F_i^E contra F_j^E se aproximará de uma reta com coeficiente angular 1 passando pela origem.

A função distribuição acumulada empírica é calculada por

$$F_i^E(x_{il}) = (v_{il} - 0.5)/N \ , \ \text{para} \ i = 1, \ldots, k \ e \ l = 1, \ldots, n_i \ , \qquad (3.7)$$

onde v_{il} é o posto correspondente a x_{il} quando todas as observações estão em ordem crescente. Uma outra opção é calcular a função distribuição acumulada empírica por

$$F_i^E(x_{il}) = \frac{v_{il}}{N+1} \qquad (3.8)$$

que é o valor esperado de F_i^E (Kimball (1960)).

Uma discussão sobre esta técnica e outras técnicas, como o gráfico Q-Q, pode ser encontrada em Wilk e Gnanadesikan (1968).

3.2.2 Estimação de tabelas de Vida

Vamos agora considerar que os dados estão agrupados em intervalos $I_a = (t_a \ , \ t_{a+1})$, $a = 1, \ldots, A$, como é mais frequente em estudos do tipo tabela de vida.

Sejam

$$N_{ia} = \text{número de falhas no intervalo } I_a \text{ por } C_i \ ,$$
$$S_a = \text{número de sobreviventes em } t_a \ .$$

Dado que um elemento sobreviveu até t_a, as probabilidades de que ele falhe devido a $C_1, \ldots, C_k$ são, respectivamente, $Q_{1a}, \ldots, Q_{Ka}$, e a probabilidade de que ele sobreviva no intervalo I_a é p_a (estas probabilidades estão

definidas no capítulo 2). Então, a distribuição de probabilidade conjunta de $N_{1a},\ldots,N_{Ka},S_{a+1}$, dado $S_a = s_a$, é a multinomial

$$(N_{1a},\ldots,N_{Ka},S_{a+1}) \sim M(Q_{1a},\ldots,Q_{Ka},p_a) \ ,$$

com f.p. conjunta

$$\prod_{a=1}^{A} \frac{s_a}{N_{1a}!\cdots N_{Ka}!S_{a+1}!} Q_{1a}^{N_{1a}}\cdots Q_{Ka}^{N_{Ka}} p_a^{S_{a+1}} \ . \tag{3.9}$$

Desta forma, com as frequências observadas $n_{1a},\ldots,n_{Ka},s_a$ no intervalo I_a, os estimadores de máxima verossimilhança para as probabilidades Q_{ia} e p_a, são

$$\hat{Q}_{ia} = \frac{n_{ia}}{s_a} \ , \ i = 1,\ldots,k \ , \tag{3.10}$$

e

$$\hat{p}_a = \frac{s_{a+1}}{s_a} \ . \tag{3.11}$$

Com isso, temos os estimadores das probabilidades brutas. A estimação das probabilidades líquidas, q_{ia}, não é tão direta. Para chegarmos a estes estimadores usaremos a suposição de proporcionalidade, que, para dados grupados em intervalos, fica:

$$\frac{\lambda_i(x,a)}{\lambda_Z(x,a)} = \gamma_{ia} \ , \tag{3.12}$$

que depende apenas de i e a .

As consequências (1) a (4), vistas na seção anterior, continuam valendo:

1.

$$\gamma_{ia} = \frac{\pi_{ia}}{\pi_{.a}} \ ,$$

pois:

$$\begin{aligned} \pi_{ia} &= P(Y_i = Z \ , \ Z' \in I_a) \\ &= \int_{I_a} \bar{F}_Z(x)\lambda_i(x)dx \\ &= \gamma_{ia}\int_{I_a} f_Z(x)dx \\ &= \gamma_{ia}\pi_{.a} \ , \end{aligned}$$

$$\Rightarrow \quad \gamma_{ia} = \frac{\pi_{ia}}{\pi_{.a}} \ , \ \text{com } \pi_{.a} = P(\text{falha em} I_a).$$

2. P(falha por C_i| falha em $(c,b) \subset I_a) = \gamma_{ia}$, independe de (c,d) .

3. $f_i(x) \propto f_Z(x)$,pois

$$f_i(x) = \frac{1}{\pi_i}\lambda_i(x)\bar{F}_Z(x) = \frac{\gamma_{ia}}{\pi_i}\lambda_Z(x)\bar{F}_Z(x) = \frac{\gamma_{ia}}{\pi_i}f_Z(x) \ .$$

4.

$$\frac{\bar{P}_i(t_{a+1})}{\bar{P}_i(t_a)} = \left[\frac{\bar{F}_Z(t_{a+1})}{\bar{F}_Z(t_a)}\right]^{\gamma_{ia}} \ ,$$

onde

$$\frac{\bar{P}_i(t_{a+1})}{\bar{P}_i(t_a)} = P(\text{sobreviver a } C_i \text{ em } I_a | \text{ vive em } t_a).$$

Se a suposição de proporcionalidade é satisfeita, as probabilidades q_{ia} podem ser estimadas através da seguinte relação

$$\hat{q}_{ia} = 1 - \hat{p}_a^{\hat{Q}_{ia}/\hat{q}_a} \ , \tag{3.13}$$

onde

$$\hat{q}_a = 1 - \hat{p}_a \ ,$$

e as probabilidades brutas parciais, $Q_{ia\,j}$, podem ser estimadas por:

$$\hat{Q}_{ia.j} = \frac{\hat{Q}_{ia}}{\hat{q}_a - \hat{Q}_{ja}}(1 - \hat{p}_a^{(\hat{q}_a - \hat{Q}_{ja})/\hat{p}_a}). \tag{3.14}$$

A primeira relação é obtida como segue:

$$q_{ia} = 1 - \frac{\bar{P}_i(t_{a+1})}{\bar{P}_i(t_a)}$$

e

$$Q_{ia} = P(Y_i = Z \ , \ Z \in I_a \mid Z > t_a) = \frac{\pi_{ia}}{\bar{F}_Z(t_a)} \ .$$

Aplicando a consequência 4, temos:

$$q_{ia} = 1 - \left[\frac{\bar{F}_Z(t_{a+1})}{\bar{F}_Z(t_a)}\right]^{\gamma_{ia}} = 1 - p_a^{\gamma_{ia}} \quad ,$$

$$q_a = \frac{\pi_{.a}}{\bar{F}_Z(t_a)} = \frac{\pi_{ia}}{\gamma_{ia}\bar{F}_Z(t_a)} = \frac{Q_{ia}}{\gamma_{ia}} \quad ,$$

$$\Rightarrow \quad \gamma_{ia} = \frac{Q_{ia}}{q_a} \quad ,$$

portanto

$$q_{ia} = 1 - p_a^{Q_{ia}/q_a} \quad .$$

A segunda relação é obtida da seguinte maneira:

$$Q_{ia.j} = P(\, Y_i = Z',\ Z' \in I_a |\, Z' > t_a) = \gamma'_{ia} q_{a.j} \quad ,$$

onde:

$$Z' = \min_{l \neq j} Y_l \quad ,$$

$$\gamma'_{ia} = \frac{\lambda_i(x,a)}{\lambda_{Z'}(x,a)} = \frac{\lambda_i(x,a)}{\lambda_Z(x,a) - \lambda_j(x,a)} = \frac{Q_{ia}}{q_a - Q_{ja}} \quad ,$$

$$\left(\text{pois } \gamma_{ia} = \frac{Q_{ia}}{q_a}\right) \quad ,$$

e

$$\begin{aligned} q_{a.j} &= P(\text{ falha em } I_a \text{ quando } C_j \text{ é eliminado } |\text{ vivo em } t_a) \\ &= 1 - p_a^{(q_a - q_{ja})/q_a} \quad . \end{aligned} \tag{3.15}$$

Então

$$Q_{ia.j} = \gamma'_{ia} q_{a.j} = \frac{Q_{ia}}{q_a - Q_{ja}}\left(1 - p_a^{(q_a - Q_{ja})/q_a}\right) \quad .$$

Temos os estimadores de máxima verossimilhança das probabilidades de falha Q_{ia} , q_{ia} e $Q_{ia.j}$, e das probabilidades de sobrevivência respectivas. As esperanças e variâncias destes estimadores estão a seguir.

Temos que

$$E(\hat{Q}_{ia}|\, s_a) = Q_{ia} \quad \text{e} \tag{3.16}$$
$$E(\hat{p}_{ia}|\, s_a) = p_a \quad . \tag{3.17}$$

Então

$$E(\hat{Q}_{ia}) = E(E(\hat{Q}_{ia}|\, s_a)) = E(Q_{ia}) = Q_{ia} \quad \text{e}$$
$$E(\hat{p}_a) = E(E(\hat{p}_a|\, sa)) = E(p_a) = p_a \quad .$$

logo, $\hat{Q}_{ia}$ e $\hat{p}_a$ são estimadores não viciados.

Variâncias e Covariâncias

Vamos estimar a variância dos estimadores vistos até agora. Como em (3.16) e (3.17) ,temos

$$V(\hat{Q}_{ia} \mid s_a) = \frac{Q_{ia}(1 - Q_{ia})}{s_a} \ ,$$

$$V(\hat{p}_a|s_a) = \frac{p_a(1 - p_a)}{s_a} \ ,$$

e

$$Cov[(\hat{Q}_{ia} \ , \ \hat{Q}_{ja}) \mid s_a] = -\frac{Q_{ia}Q_{ja}}{s_a} \ , \ i \neq j \ .$$

Para obtermos as variâncias e covariâncias não condicionais, usamos a seguinte propriedade de covariância:

$$Cov(Y_i \ , \ Y_j) = E[Cov((Y_i \ , \ Y_j) \mid X)] + Cov[E(Y_i \mid X) \ , \ E(Y_j \mid X)],$$

cuja demonstração pode ser vista em Rao (1973).

Como

$$E(\hat{Q}_{ia} \mid s_a) = Q_{ia} \quad e$$
$$E(\hat{p}_a \mid s_a) = p_a \ ,$$

temos

$$Cov[E(\hat{Q}_{ia} \mid s_a)\ ,\ E(\hat{Q}_{ia} \mid s_a)] = 0 \quad e$$
$$Cov[E(\hat{Q}_{ia} \mid s_a)\ ,\ E(\hat{p}_a \mid s_a)] = 0\ .$$

Então, as variâncias e covariâncias não condicionais ficam:

$$V(\hat{Q}_{ia}) = E(\frac{1}{S_a})Q_{ia}(1 - Q_{ia}), \tag{3.18}$$

$$V(\hat{p}_a) = E(\frac{1}{S_a})p_a q_a\ , \tag{3.19}$$

$$Cov(\hat{Q}_{ia}\ ,\ \hat{Q}_{ja}) = -E(\frac{1}{S_a})Q_{ia}Q_{ja}\ ,\ i \neq j\ \text{e} \tag{3.20}$$

$$Cov(\hat{Q}_{ia}\ ,\ \hat{p}_a) = -E(\frac{1}{S_a})Q_{ia}p_a\ , \tag{3.21}$$

onde

$$E(\frac{1}{S_a}) \approx \frac{1}{E(S_a)} \tag{3.22}$$

(ver apêndice B) e

$$E(S_a) = s_0 \times p(\text{sobreviver até } t_a) = s_0 \prod_{l=1}^{a} p_l\ .$$

A variância de $\hat{q}_{ia}$ pode ser obtida através da aproximação

$$V[\phi(X,Y)] \approx \sigma_X^2(\frac{\partial \phi}{\partial X}\mid_{X=\mu_X})^2 + \sigma_Y^2(\frac{\partial \phi}{\partial Y}\mid_{Y=\mu_Y})^2 +$$
$$+ 2\sigma_{XY}(\frac{\partial \phi}{\partial X})(\frac{\partial \phi}{\partial Y}) \tag{3.23}$$

(ver Apêndice B).

No nosso caso

$$\hat{q}_{ia} = \phi(\hat{Q}_{ia}\ ,\ \hat{p}_a) = 1 - \hat{p}_a^{\frac{Q_{ia}}{q_a}}\ ,$$

então

$$V(\hat{q}_{ia}) \approx E(\frac{1}{S_a})\frac{(1-q_{ia})^2}{p_a q_a} \times$$

$$\times \left[p_a \ln(1-q_{ia})\ln(1-q_{a.i}) + Q_{ia}^2\right] , \qquad (3.24)$$

onde $q_{a.i}$ é dada por (3.15).

A variância de $\hat{Q}_{ia.j}$ pode ser obtida através do mesmo método.

3.3 Estimador de Kaplan-Meier

O estimador da probabilidade de sobrevivência desenvolvido por Kaplan e Meier em 1958 considera dois riscos competivos: falha e perda de observação, ou censura. Esta técnica supõe falhas e censuras independentes e não assume forma paramétrica para as distribuições de probabilidade dos riscos. Nesta seção, veremos a expansão desta técnica para K riscos competitivos independentes, como foi feito por Hoel (1972) além de David e Moeschberger (1978).

3.3.1 Construção do Estimador

Sejam $Y_1, \ldots, Y_K$ os tempos de vida teóricos associados aos K riscos competitivos. Vamos supor que estas variáveis aleatórias são independentes. Os valores observados são:

$$x_{11}, \ldots, x_{1n_1}, \ldots, x_{K1}, \ldots, x_{Kn_K} .$$

A idéia do estimador Kaplan-Meier é criar intervalos aleatórios para cada risco, onde os limites superiores dos intervalos são os tempos de vida observados, ou seja, para cada C_i , criam-se $n_i + 1$ intervalos I_{ia} , $a = ,1, \ldots, n_i$, da seguinte maneira:

$$[0\ ,\ x_{i1}),\ [x_{i1}\ ,\ x_{i2}), \ldots, [x_{in_i}\ ,\ \infty) ,$$

para $i = 1, \ldots, k$.

Agora definimos $v_{i1}, \ldots, v_{in_i}$, os postos correspondentes a $x_{i1}, \ldots, x_{in_i}$ entre

todas as observações em ordem crescente $X_1, \ldots, X_n$. Assim, os estimadores das probabilidades líquidas de falha, q_{ij} , $j = 0, 1, \ldots, n_i$, são

$$\begin{aligned} \hat{q}_{i0} &= 0 \quad \text{para } [0\,, x_{i1}) \ , \\ \hat{q}_{i1} &= \frac{1}{n - v_{i1} + 1} \quad \text{para } [x_{i1}\,,\ x_{i2}) \ , \\ &\vdots \qquad\quad \vdots \ . \end{aligned}$$

Generalizando

$$\hat{q}_{ij} = \frac{1}{n - v_{ij} + 1} \ . \tag{3.25}$$

A probabilidade de sobreviver ao risco C_i para cada intervalo é estimada por $1 - \hat{q}_{ij}$:

$$\hat{p}_{ij} = \frac{n - v_{ij}}{n - v_{ij} + 1} \ , \tag{3.26}$$

e, para o primeiro intervalo, esta probabilidade é 1 (note que esta probabilidade é uma probabilidade líquida). Então, a probabilidade de um indivíduo sobreviver a C_i , até o tempo x, quando C_i é o único risco agindo na população, é estimada por

$$\hat{\bar{P}}_i(x) = \prod_{l=1}^{L} \frac{n - v_{il}}{n - v_{il} + 1} \quad \text{para } x \geq x_{i1} \ , \tag{3.27}$$

onde

$$L = \{j |\ x_{ij} \leq x\} \ , \ j = 1, \ldots, n_i \ ,$$

e

$$\hat{\bar{P}}_i = 1 \quad \text{para } \ x < x_{i1} \ .$$

O estimador Kaplan-Meier nos dá, então, uma estimativa não paramétrica da probabilidade de sobrevivência da variável Y_i , o tempo de vida teórico correspondente ao risco C_i . Podemos, de maneira análoga e bem simples, estimar as probabilidades brutas $Q_i(x)$. Vamos considerar os tempos de vida ordenados $X_1, \ldots, X_n$ e seja

r_i = número de mortes pela causa i até o tempo x .

Então,

$$\hat{Q}_i(x) = \frac{r_i}{n} \ .$$

A probabilidade de sobrevivência fica

$$\hat{\bar{Q}}_i(x) = \frac{n_i - r_i}{n} \ .$$

3.3.2 Propriedades e Variância do Estimador de Kaplan-Meier

Além de $\hat{\bar{P}}_i(x)$ ser um estimador não paramétrico, ele é um estimador de máxima verossimilhança de $\bar{P}_i$. Veremos a seguir que $\hat{\bar{P}}_i(x)$ maximiza a função de verossimilhança dos valores observados, dentro da classe de possíveis funções $\bar{P}_i(x)$. A função de verossimilhança é dada por

$$L = \prod_{i=1}^{K} L_i \ , \text{ onde}$$

$$\begin{aligned} L_i &= [\prod_{m=1}^{v_{ij}-1} (X_m)] \times \prod_{j=1}^{n_i} \{[\bar{P}_i(x_{ij} - o) - \bar{P}_i(x_{ij})] \times \\ &\times \prod_{m=v_{ij}+1}^{v_{ij+1}} \bar{P}_i(X_m)\} \ , \end{aligned}$$

onde $v_{ij+1} = n + 1$.

Para que L_i seja máxima, $\bar{P}_i(X_m)$ e $\bar{P}_i(X_{ij} - o)$ devem ser grandes, enquanto $\bar{P}_i(x_{ij})$ deve ser pequena. Para tanto, fazemos:

$$\begin{aligned} \bar{P}_i(X_m) &= \bar{P}_i(x_{i1} - o) = 1 \ , m = 1, \ldots, v_{i1} - 1 \ , \\ \bar{P}_i(x_{ij}) &= \bar{P}_i(X_m) = \bar{P}_i(x_{ij+1} - o) \ , \qquad (3.28) \\ m &= v_{ij} + 1, \ldots, v_{ij+1} - 1 \ \ e \ \ j = 1, \ldots, n_i - 1 \ , \\ \bar{P}_i(x_{in_i}) &= \bar{P}_i(X_m) \quad m = v_{in_i} + 1, \ldots, n \ . \qquad (3.29) \end{aligned}$$

Vamos chamar (3.29) e (3.30) de $\bar{P}_{ij}$ para $j = 1, \ldots, n_i$ e substituir em L_i:

$$L_i = \prod_{j=1}^{n_i} (\bar{P}_{ij-1} - \bar{P}_{ij}) \bar{P}_{ij}^{v_{ij+1} - v_{ij} - 1} \ .$$

Como

$$p_{ij} = 1 - q_{ij} = \frac{\bar{P}_{ij}}{\bar{P}_{ij-1}} \ ,$$

temos

$$\bar{P}_{ij} = p_{i1}p_{i2}\cdots\ p_{ij} \ \text{ e}$$
$$\bar{P}_{ij-1} - \bar{P}_{ij} = p_{i1}p_{i2}\cdots\ p_{ij-1}q_{ij} \ .$$

Substituindo em L_i:

$$\begin{aligned} L_i &= q_{i1}p_{i1}^{v_{i2}-v_{i1}-1} \times p_{i1}q_{i2}p_{i2}^{v_{i3}-v_{i2}-1} \times \cdots \times \\ &\quad \times p_{i1}p_{i2}\cdots p_{in_i-1}q_{in_i}p_{in_i}^{v_{in_i+1}-v_{in_i}-1} = \\ &= \prod_{j=1}^{n_i} p_{ij}^{B}q_{ij} \ , \end{aligned}$$

onde

$$\begin{aligned} B &= \sum_{a=j}^{n_i}(v_{ia+1} - v_{ia-1}) + n - j = \\ &= v_{ij+1} + \cdots + v_{in_i} + v_{in_i+1} - \\ &\quad -(v_{ij} + \cdots + v_{in_i}) - (n_i - j + 1) + n - j = \\ &= v_{in_i+1} - v_{ij} - 1 = n + 1 - v_{ij} - 1 = \\ &= n - v_{ij} \ . \end{aligned}$$

Logo

$$L_i = \prod_{j=1}^{n_i} p_{ij}^{n-v_{ij}}q_{ij} \ .$$

Seja

$$L_{ij} = p_{ij}^{n-v_{ij}}(1 - p_{ij}) \ .$$

Queremos o valor de p_{ij} que maximiza L_{ij} :

$$\log L_{ij} = (n - v_{ij})\log p_{ij} + \log(1 - p_{ij})$$

$$\frac{\partial \log L_{ij}}{\partial p_{ij}} = \frac{n - v_{ij}}{p_{ij}} - \frac{1}{1 - p_{ij}} = 0$$

$$\Rightarrow \quad \hat{p}_{ij} = \frac{n - v_{ij}}{n - v_{ij} + 1} \ .$$

Portanto

$$\hat{\bar{P}}_i(x) = \prod \hat{p}_{ij}$$

é um estimador de máxima verossimilhança para $\bar{P}_i(x)$.

Variância

Temos

$$\hat{P}_i(x) = \prod_{j=1}^{L} \hat{p}_{ij} \ ,$$

então

$$\log \hat{P}_i(x) = \sum_{j=1}^{L} \log \hat{p}_{ij} \ .$$

Aplicando a aproximação vista na seção 3.2.2 em (3.23), temos

$$\begin{aligned} V(\log \hat{p}_{ij}) &\approx V(\hat{p}_{ij})\left(\frac{\partial}{\partial \hat{p}_{ij}} \log \hat{p}_{ij}\right)^2 \\ &\approx \frac{\hat{p}_{ij}\hat{q}_{ij}}{n^*} \frac{1}{\hat{p}_{ij}^2} = \frac{\hat{q}_{ij}}{n^* \hat{p}_{ij}} \ , \end{aligned}$$

onde $n^* = n - v_{ij} + 1$. Assumindo independência entre $\log \hat{p}_{ij}$'s :

$$\begin{aligned} V(\log \hat{\bar{P}}_i(x)) &\approx \sum V(\log \hat{p}_{ij}) = \sum \frac{\hat{q}_{ij}}{n^* \hat{p}_{ij}} \\ V(\hat{\bar{P}}_i) &\approx V(\log \hat{\bar{P}}_i(x))\left(\frac{\partial}{\partial \log \hat{\bar{P}}_i} \exp \log \hat{\bar{P}}_i\right)^2 \\ &\approx \sum_{j=1}^{L} \frac{\hat{q}_{ij}}{n^* \hat{p}_{ij}} \hat{\bar{P}}_i^{\,2} \ . \end{aligned}$$

Substituindo os valores de $\hat{q}_{ij}$ e $\hat{p}_{ij}$ dados por (3.25) e (3.26), temos

$$V(\hat{\bar{P}}_i(x)) \approx \hat{\bar{P}}_i^{\,2}(x) \sum_{j=1}^{L} \frac{1}{(n - v_{ij} + 1)(n - v_{ij})} \quad . \tag{3.30}$$

Propriedades assintóticas do estimador Kaplan-Meier podem ser vistas em Breslow e Crowley (1974).

Capítulo 4

Uso de Covariáveis: Modelo de Cox

No capítulo anterior, estudamos o problema de estimação levando em consideração apenas uma variável em estudo: o tempo até falha. Entretanto, frequentemente são introduzidas no problema covariáveis, ou variáveis explicativas. Por exemplo, em uma comparação simples de dois tratamentos, podemos considerar uma covariável binária igual a zero para um dos tratamentos e igual a um, para o outro tratamento. Dependendo do desenho experimental utilizado, podemos ter várias covariáveis desse tipo. Covariáveis podem, também, conter informações sobre características dos indivíduos em estudo, como por exemplo, sexo e idade. Desta forma, para cada indivíduo estará associado um vetor de covariáveis, $\mathbf{Z} = (Z_1, \ldots, Z_m)$, e queremos analisar o efeito dessas covariáveis sobre o tempo até falha, ou seja, na função que estivermos estimando. Para tanto, temos que ajustar modelos nos quais o efeito das covariáveis é representado por parâmetros desconhecidos.

Vamos restringir nosso estudo, neste capítulo, ao modelo proposto por Cox (1972), que tem uma interpretação simples e pode acomodar riscos competitivos.

Este modelo, conhecido como *Modelo de Taxas Proporcionais de Cox*, supõe que a taxa de falha, λ, de cada indivíduo é dada por

$$\lambda(x; \mathbf{Z}) = \lambda_0(x) \exp(\mathbf{Z}'\beta) \ , \qquad (4.1)$$

onde

$\mathbf{Z}$ = vetor das covariáveis observadas para um indivíduo com dimensão $mx1$

β = vetor de parâmetros desconhecidos $(mx1)$

$\lambda_0(x)$ = taxa de falha para um indivíduo, quando $\mathbf{Z} = \mathbf{0}$.

A função $\exp(\mathbf{Z}'\beta)$ poderia ser qualquer função $f(\mathbf{Z}'\beta)$ positiva e satisfazendo $f(\mathbf{0}) = 1$. Cox usou $f(\mathbf{Z}'\beta) = \exp(\mathbf{Z}'\beta)$, pois , além de satisfazer as duas condições acima, esta função facilita a computação dos parâmetros desconhecidos.

4.1 O Modelo de Cox com K Riscos Competitivos

No caso de Riscos Competitivos, para cada indivíduo temos $Y_1, \ldots, Y_K$ tempos de vida teóricos independentes, e o vetor de covariáveis $\mathbf{Z}$, com

$$Y_i \sim p_i(x; \mathbf{Z}) \quad \text{e}$$

$$\lambda_i(x; \mathbf{Z}) = \frac{p_i(x; \mathbf{Z})}{\bar{P}_i(x; \mathbf{Z})}$$

correspondendo a cada risco. A taxa global fica

$$\lambda(x; \mathbf{Z}) = \sum_{i=1}^{K} \lambda_i(x; \mathbf{Z}) \ .$$

Em termos do modelo de Cox temos:

$$\lambda_i(x; \mathbf{Z}) = \lambda_{0i}(x) \exp(\mathbf{Z}'\beta_i) \ , \tag{4.2}$$

onde β_i é o vetor de parâmetros desconhecidos correspondente ao risco i , $\beta_i = (\beta_{i1}, \ldots, \beta_{im})$. Um caso particular é quando β_i pode ser substituído por um β comum a todos riscos.

A partir de (4.2) temos

$$p_i(x; \mathbf{Z}) = \underbrace{\lambda_{0i}(x) \exp(\mathbf{Z}'\beta)}_{\lambda_i(x;\mathbf{Z})} \underbrace{\exp(-\Lambda_{0i}(x) \exp(\mathbf{Z}'\beta))}_{\bar{P}_i(x;\mathbf{Z})} \ ,$$

pois, por (2.6),

$$\begin{aligned}\bar{P}_i(x;\mathbf{Z}) &= \exp(-\int_0^x \lambda_i(t;\mathbf{Z})dt) \\ &= \exp(-\int_0^x \lambda_{0i}(t)\exp(\mathbf{Z}'\beta_i)dt) \\ &= \exp(-\exp(\mathbf{Z}'\beta_i)\underbrace{\int_0^x \lambda_{0i}(t)dt}_{\Lambda_{0i}(x)}) \ .\end{aligned}$$

4.1.1 Estimação de Máxima Verossimilhança de β

Cox construiu a estimação de β através de uma verossimilhança condicional. Esta verossimilhança baseia-se em que, nos intervalos em que não ocorre falha, não podemos tirar informação alguma sobre β, pois λ_{0i} pode, teoricamente, ser igual a zero nestes intervalos. Cox, então, argumenta condicionalmente ao conjunto de instantes onde as falhas ocorrem. Então, para um tempo x_{ij}, condicionalmente ao conjunto de risco $\Re(x_{ij})$, isto é, os indivíduos que estão vivos no tempo x_{ij}, a probabilidade do individuo j falhar pela causa i no tempo x_{ij} dado que a falha é em x_{ij}, é, para $i = 1,\ldots,K$ e $j = 1,\ldots,n_i$,

$$\begin{aligned}&\frac{\lambda_{0i}(x_{ij})\exp(\mathbf{Z}'_{x_{ij}}\beta_i)}{\sum_{l=1}^{k}\sum_{h\in\Re}\lambda_{0l}(x_{ij})\exp(\mathbf{Z}'_h\beta_i)} = \\ &= \frac{\lambda_{0i}(x_{ij})\exp(\mathbf{Z}'_{x_{ij}}\beta_i)}{\sum_{l=1}^{k}\lambda_{0l}(x_{ij})\sum_{h\in\Re}\exp(\mathbf{Z}'_h\beta_i)} \\ &= \frac{\lambda_{0i}(x_{ij})}{\lambda_0(x_{ij})}\frac{\exp(\mathbf{Z}'_{x_{ij}}\beta_i)}{\sum_{h\in\Re}\exp((\mathbf{Z}'_h\beta_i)} \ ,\end{aligned}$$

onde

$\mathbf{Z}_{x_{ij}}$ = vetor covariado $mx1$ correspondente ao indivíduo ij
$\mathbf{Z}_h$ = vetor covariado $mx1$ correspondente ao indivíduo h do conjunto $\Re$.

Se as taxas são proporcionais, a razão $\lambda_{0i}(x_{ij})/\lambda_0(x_{ij})$ não depende do tempo de falha. Então a probabilidade condicional não depende do tempo de falha dado.

A Função de Verossimilhança Condicional fica :

$$L(\beta_1, \beta_2, \ldots, \beta_k) = \prod_{i=1}^{k} \frac{\lambda_{0i}}{\lambda_0} \prod_{j=1}^{n_i} \frac{\exp(\mathbf{Z}'_{x_{ij}}\beta_i)}{\sum_{h\in\Re} \exp(\mathbf{Z}'_h\beta_i)} \ . \tag{4.3}$$

Podemos notar que λ_{0i} e β_i variam para $i = 1, \ldots, k$, ou seja, para todos os riscos competitivos, e também que os vetores $\beta'_i s$ podem ser estimados separadamente para cada risco:

$$L_i(\beta_i) = \frac{\lambda_{0i}}{\lambda_0} \prod_{j=1}^{n_i} \frac{\exp(\mathbf{Z}'_{x_{ij}}\beta_i)}{\sum_{h\in\Re} \exp(\mathbf{Z}'_h\beta_i)} \tag{4.4}$$

$$\ln L_i(\beta_i) = \ln \frac{\lambda_{0i}}{\lambda_0} + \sum_{j=1}^{n_i} \mathbf{Z}'_{x_{ij}}\beta_i - \sum_{j=1}^{n_i} \ln\Big(\sum_{h\in\Re} \exp(\mathbf{Z}'_h\beta_i)\Big)$$

$$\frac{\partial \ln L_i(\beta_i)}{\partial \beta_{ir}} = \sum_{j=1}^{n_i} \Big(Z_{ijr} - \frac{\sum_{h\in\Re} Z_{hr} \exp(\mathbf{Z}'_h\beta_i)}{\sum_{h\in\Re} \exp(\mathbf{Z}'_h\beta_i)}\Big) = D_{ir} \ , \tag{4.5}$$

para $r = 1, \ldots, m$.

Assim, fazendo $D_{ir} = 0$, através de métodos iterativos, obtemos os estimadores de Máxima Verossimilhança para os vetores de parâmetros, $\hat{\beta}_1, \ldots, \hat{\beta}_k$.

A partir dos estimadores $\hat{\beta}_1, \ldots, \hat{\beta}_k$ podemos obter os estimadores das funções $\bar{P}_i(x; \mathbf{Z})'s$, para um certo valorde $\mathbf{Z} = \tilde{\mathbf{Z}}$:

$$\begin{aligned}
\hat{\bar{P}}_i(x; \tilde{\mathbf{Z}}) &= \exp\Big(-\int_0^x \hat{\lambda_{0i}}(t) \exp(\tilde{\mathbf{Z}}'\hat{\beta}_i) dt\Big) \\
&= \exp\Big(-\exp(\tilde{\mathbf{Z}}'\hat{\beta}_i) \int_0^x \hat{\lambda_{0i}}(t) dt\Big) \\
&= \Big[\exp\Big(-\int_0^x \hat{\lambda_{0i}}(t) dt\Big)\Big]^{\exp(\tilde{\mathbf{Z}}'\hat{\beta}_i)} \\
&= \hat{\bar{P}}_{0i}(x)^{\exp(\tilde{\mathbf{Z}}'\hat{\beta}_i)} \ ,
\end{aligned} \tag{4.6}$$

onde $\hat{\bar{P}}_{0i}(x)$ é o estimador da função de sobrevivência líquida do risco i quando $\mathbf{Z} = \mathbf{0}$. Este estimador pode ser obtido através de (3.27), o estimador Kaplan-Meier.

4.2 Verificação do Modelo

Para testar a hipótese $H_0 : \beta_i = 0$, Cox usa a estatística

$$\mathbf{D}_i' \mathbf{I}_i^{-1} \mathbf{D}_i \ , \tag{4.7}$$

onde

$$\begin{aligned} \mathbf{D}_i &= \left(\frac{\partial \ln L_i(\beta_i)}{\partial \beta_{i1}}|_{\beta_i=\mathbf{0}}, \ldots, \frac{\partial \ln L_i(\beta_i)}{\partial \beta_{im}}|_{\beta_i=\mathbf{0}}\right) \\ &= \left(\sum_{j=1}^{n_i}\left(Z_{ij1} - \frac{\sum_{h\in\Re} Z_{h1}}{\eta_{ij}}\right), \ldots, \sum_{j=1}^{n_i}\left(Z_{ijm} - \frac{\sum_{h\in\Re} Z_{hm}}{\eta_{ij}}\right)\right) , \end{aligned}$$

com η_{ij} = número de indivíduos em $\Re$ para o indivíduo x_{ij}, e $\mathbf{I}_i^{-1}$ é o inverso da matriz de Informação de Fisher, $\mathbf{I}_i$, calculada em $\beta_i = \mathbf{0}$:

$$\mathbf{I}_i|_{\beta_i=\mathbf{0}} = \begin{pmatrix} \frac{-\partial^2 \ln L_i(\beta_i)}{\partial \beta_{i1}\partial \beta_{i1}}|_{\beta_i=\mathbf{0}} & \cdots & \frac{-\partial^2 \ln L_i(\beta_i)}{\partial \beta_{i1}\partial \beta_{im}}|_{\beta_i=\mathbf{0}} \\ \vdots & \ddots & \vdots \\ \frac{-\partial^2 \ln L_i(\beta_i)}{\partial \beta_{im}\partial \beta_{i1}}|_{\beta_i=\mathbf{0}} & \cdots & \frac{-\partial^2 \ln L_i(\beta_i)}{\partial \beta_{im}\partial \beta_{im}}|_{\beta_i=\mathbf{0}} \end{pmatrix} ,$$

onde

$$\begin{aligned} -\frac{\partial^2 \ln L_i(\beta_i)}{\partial \beta_{ir}\partial \beta_{is}} &= \frac{\sum_{h\in\Re} Z_{hr} Z_{hs} \exp(\mathbf{Z}_h'\beta_i)}{\sum_{h\in\Re} \exp(\mathbf{Z}_h'\beta_i)} - \\ &\quad - \frac{\sum_{h\in\Re} Z_{hr} \exp(\mathbf{Z}_h'\beta_i) \sum_{h\in\Re} Z_{hs} \exp(\mathbf{Z}_h'\beta_i)}{(\sum_{h\in\Re} \exp(\mathbf{Z}_h'\beta_i))^2} , \end{aligned}$$

e quando $\beta_i = \mathbf{0}$:

$$-\frac{\partial^2 \ln L_i(\beta_i)}{\partial \beta_{ir}\partial \beta_{is}}|_{\beta_i=\mathbf{0}} = \frac{\sum_{h\in\Re} Z_{hR} Z_{hs}}{\eta_{ij}} - \frac{\sum_{h\in\Re} Z_{hr} \times \sum_{h\in\Re} Z_{hs}}{\eta_{ij}^2} .$$

Sob Ho, a estatística dada por (4.7) tem distribuição assintótica χ^2 com m graus de liberdade.

Capítulo 5

Análise de Riscos Competitivos através de Modelos Lineares

Johnson e Koch (1978), em uma extensão do procedimento proposto por Grizzle, Starmer e Koch (1969), abordaram o tratamento de dados de sobrevivência com riscos competitivos quando os dados estão agrupados, através de modelos lineares. Esta abordagem supõe que as taxas de falha são proporcionais, além de supor riscos independentes, e formula o problema de riscos competitivos em termos de matrizes e operações com matrizes sobre uma tabela de contingência. Assim, a probabilidade líquida de sobrevivência para cada causa é estimada pelo Método de Mínimos Quadrados Ponderados e podemos analisar subpopulações diferentes e testar essas diferenças.

5.1 Formulação

Vamos considerar os dados em uma tabela de contingência multidimensional com S subpopulações, A intervalos de tempo e K riscos competitivos:

Tabela 5.1: Tabela de Contingência Multidimensional

subpopulação	intervalo	causa de falha				vivos no fim do intervalo
		C_1	C_2	...	C_k	
1	1	n_{111}	n_{112}	...	n_{11k}	n_{110}
1	2	n_{121}	n_{122}	...	n_{12k}	n_{120}
⋮	⋮	⋮	⋮	...	⋮	⋮
1	A	n_{1A1}	n_{1A2}	...	n_{1Ak}	n_{1A0}
2	1	n_{211}	n_{212}	...	n_{21k}	n_{210}
2	2	n_{221}	n_{222}	...	n_{22k}	n_{220}
⋮	⋮	⋮	⋮	...	⋮	⋮
2	A	n_{2A1}	n_{2A2}	...	n_{2Ak}	n_{2A0}
⋮	⋮	⋮	⋮	...	⋮	⋮
S	1	n_{S11}	n_{S12}	...	n_{S1K}	n_{S10}
S	2	n_{S21}	n_{S22}	...	n_{S2K}	n_{S20}
⋮	⋮	⋮	⋮	...	⋮	⋮
S	A	n_{SA1}	n_{SA2}	...	n_{SAK}	n_{SA0}

Em cada linha hj , $h = 1, \ldots, S$ e $j = 1, \ldots, A$,temos:

$$P(\mathbf{n}_{hj}) = \frac{n_{hj+}!}{n_{hj1}! \cdots n_{hjk}! n_{hj0}!} Q_{hj1}^{n_{hj1}} \cdots Q_{hjk}^{n_{hjk}} p_{hj}^{n_{hj0}} \quad , \tag{5.1}$$

onde, para $i = 1, \ldots, k$,

$\mathbf{n}_{hj}$ = vetor de frequências em cada linha
= $(n_{hj1}\ ,\ n_{hj2}\ ,\ \ldots\ ,\ n_{hjk}\ ,\ n_{hj0})$,
n_{hji} = número de falhas por C_i ,no intervalo j ,
de um indivíduo da subpopulação h ,
$n_{hj+} = \sum_{i=0}^{k} n_{hji}$,
Q_{hji} = P(falha por C_i , no intervalo j, de um indivíduo da subpopulação h, na presença de todas causas

| indivíduo vivo no início do intervalo)
= probabilidade bruta,

p_{hj} = P(sobrevivência no intervalo j, de um indivíduo da subpopulação h | indivíduo vivo no início do intervalo).

Vamos definir (como foi visto no capítulo 2):

q_{hji} = P(falha por C_i , no intervalo j, de um indivíduo da subpopulação h, quando i é o único risco agindo na população | indivíduo vivo no início do intervalo)= probabilidade líquida,

$p_{hji} = 1 - q_{hji}$ = probabilidade líquida de sobrevivência,

$q_{hj} = 1 - p_{hj}$.

5.2 Estimação

As probabilidades Q_{hji} e p_{hj} podem ser estimadas diretamente por

$$\hat{Q}_{hji} = \frac{n_{hji}}{n_{hj+}} \quad \text{e} \quad \hat{p}_{hj} = \frac{n_{hj0}}{n_{hj+}} \ . \tag{5.2}$$

Se a suposição de proporcionalidade é satisfeita, pela relação dada em (3.13) temos

$$q_{hji} = 1 - p_{hj}^{(Q_{hji}/q_{hj})} \ ,$$

então

$$p_{hji} = 1 - q_{hji} = p_{hj}^{(Q_{hji}/q_{hj})} \ .$$

Seja

$$U_{hj} = \frac{Q_{hji}}{q_{hj}} \ ,$$

assim temos

$$\hat{p}_{hji} = \hat{p}_{hj}^{\hat{U}_{hj}}$$

onde

$$\hat{U}_{hj} = \frac{\frac{n_{hji}}{n_{hj+}}}{\frac{n_{hj+}-n_{hji}}{n_{hj+}}} = \frac{n_{hji}}{\sum_{i=1}^{k} n_{hji}} \ .$$

A função que queremos estimar é:

$$F_{hji} = \prod_{l=1}^{j} \hat{p}_{hli} \ , \tag{5.3}$$

que é a probabilidade líquida de sobrevivência a C_i até o intervalo j , para a subpopulação h. Colocando em notação matricial, temos:

$$\mathbf{Q}'_{(1\times(SAK+SA))} = (\mathbf{Q}'_{11} \ , \ldots , \mathbf{Q}'_{1A} \ , \ldots , \ \mathbf{Q}'_{S1} \ , \ldots , \ \mathbf{Q}'_{SA}) \ ,$$

$$\mathbf{Q}'_{hj} = (Q_{hj1} \ , \ldots , \ Q_{hjk} \ , \ p_{hj}) \ ,$$

e

$$\hat{\mathbf{Q}}'_{(1\times(SAK+SA))} = (\hat{\mathbf{Q}}'_{11} \ , \ldots , \ \hat{\mathbf{Q}}'_{1A} \ , \ldots , \ \hat{\mathbf{Q}}'_{S1} , \ldots , \ \hat{\mathbf{Q}}'_{SA}) \ .$$

A variância e esperança de $\hat{\mathbf{Q}}$ são:

$$\begin{aligned} E(\hat{\mathbf{Q}}) &= \mathbf{Q} \\ V(\hat{\mathbf{Q}}) &= E\{[\hat{\mathbf{Q}} - E(\hat{\mathbf{Q}})][\hat{\mathbf{Q}} - E(\hat{\mathbf{Q}})]'\} \\ &= E(\hat{\mathbf{Q}}\hat{\mathbf{Q}}') - \mathbf{Q}\mathbf{Q}' \ . \end{aligned}$$

Podemos particionar $V(\hat{\mathbf{Q}})$ da seguinte maneira:

$$V(\hat{\mathbf{Q}}) = \begin{bmatrix} V(\hat{\mathbf{Q}}_{11}) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & V(\hat{\mathbf{Q}}_{SA}) \end{bmatrix} \ ,$$

e sabemos que

$$E(\hat{Q}_{hji}) = Q_{hji} \ , \quad V(\hat{Q}_{hji}) = \frac{Q_{hji}(1 - Q_{hji})}{n_{hj+}} \ ,$$

$$Cov(\hat{Q}_{hji} \ , \ \hat{Q}_{hji'}) = -\frac{Q_{hji}Q_{hji'}}{n_{hj+}} \quad e \quad Cov(\hat{Q}_{hji} \ , \ \hat{Q}_{h'j'i}) = 0 \ .$$

Então

$$V(\hat{\mathbf{Q}}_{hj}) = \frac{1}{n_{hj+}} \begin{bmatrix} Q_{hj1}(\mathbf{1} - Q_{hj1}) & -Q_{hj1}Q_{hj2} & \dots & -Q_{hj1}Q_{hjk} \\ -Q_{hj2}Q_{hj1} & Q_{hj2}(1 - Q_{hj2}) & \dots & -Q_{hj2}Q_{hjk} \\ \vdots & \vdots & \ddots & \vdots \\ -Q_{hjk}Q_{hj1} & -Q_{hjk}Q_{hj2} & \dots & Q_{hjk}(1 - Q_{hjk}) \end{bmatrix} .$$

Assim, escrevemos $\mathbf{V}(\hat{\mathbf{Q}})$, o estimador consistente da matriz de variâncias e covariâncias de $\hat{\mathbf{Q}}$, como uma matriz bloco diagonal com dimensão $(SAK + SA) \times (SAK + SA)$ e cada bloco $V(\hat{\mathbf{Q}}_{hj})$ com dimensão $(K+1) \times (K+1)$:

$$\mathbf{V}(\hat{\mathbf{Q}}_{hj}) = \frac{1}{n_{hj+}} [\mathbf{D}_{\hat{\mathbf{Q}}_{hj}} - \hat{\mathbf{Q}}_{hj}\hat{\mathbf{Q}}'_{hj}] ,$$

onde

$\mathbf{D}_{\hat{\mathbf{Q}}_{hj}}$ = matriz diagonal com os elementos de $\hat{\mathbf{Q}}_{hj}$ na diagonal principal .

e

$$\mathbf{V}(\hat{\mathbf{Q}}) = [\mathbf{D}_{\hat{\mathbf{Q}}} - \hat{\mathbf{Q}}\hat{\mathbf{Q}}']/\mathbf{N} ,$$

onde

$\mathbf{D}_{\hat{\mathbf{Q}}}$ = matriz diagonal com os elementos de $\hat{\mathbf{Q}}$ na diagonal principal,

$\hat{\mathbf{Q}}\hat{\mathbf{Q}}'$ = matriz bloco diagonal formada pelos blocos $\hat{\mathbf{Q}}_{hj}\hat{\mathbf{Q}}'_{hj}$,

$\mathbf{N}$ = matriz bloco diagonal formada pelos blocos $n_{hj+}\mathbf{1}$
($\mathbf{1}$ = matriz de uns com dimensão (k+1)x(k+1)).

Seja

$$\mathbf{F}_{(SAK\times 1)} = (F_{111}, \dots, F_{SAK}) , \tag{5.4}$$

que em função do vetor $\hat{\mathbf{Q}}$, é igual a:

$$\mathbf{F}(\hat{\mathbf{Q}}) = \exp\left\{ (-\mathbf{A}_6) \exp\left([\mathbf{A}_4, \mathbf{A}_5] \left[\log\left\{ \begin{matrix} \mathbf{A}_1\hat{\mathbf{Q}} \\ \mathbf{A}_3[\log(\mathbf{A}_2\hat{\mathbf{Q}})] \end{matrix} \right\} \right] \right) \right\} , \tag{5.5}$$

onde:

- $\mathbf{A}_1$ gera os numeradores e denominadores de $\hat{U}_{hji}$:

$$\mathbf{A}_1 = \mathbf{I}_{SA} \otimes \begin{bmatrix} \mathbf{I}_k & \mathbf{0}_k \\ \mathbf{1}'_k & 0 \end{bmatrix} ,$$

$$\begin{aligned} \otimes &= \text{ produto de Kronecker,} \\ \mathbf{I}_k &= \text{ matriz identidade } k \times k , \\ \mathbf{0}_k &= \text{ vetor de zeros } k \times 1 , \\ \mathbf{1}_k &= \text{ vetor de uns } k \times 1 . \end{aligned}$$

- $\mathbf{A}_2$ gera os numeradores e denominadores de $\hat{p}$:

$$\mathbf{A}_2 = \mathbf{I}_{SA} \otimes [\mathbf{0}'_k, 1] .$$

- $\mathbf{A}_3$, $\mathbf{A}_4$ e $\mathbf{A}_5$ geram, respectivamente, $-\log \hat{p}_{hj}$, $\log \hat{U}_{hji}$ e $\log(-\log \hat{p}_{hj})$:

$$\begin{aligned} \mathbf{A}_3 &= -I_{SA} , \\ \mathbf{A}_4 &= [I_{SA} \otimes [\mathbf{I}_k , \quad \mathbf{1}_k] , \\ \mathbf{A}_5 &= \mathbf{I}_{SA} \otimes \mathbf{1}_k . \end{aligned}$$

- $\mathbf{A}_6$ gera $\log \hat{F}_{hji}$:

$$\mathbf{A}_6 = \mathbf{I}_S \otimes [\mathbf{I}_k \otimes \mathbf{T}'_{1,A}] ,$$

e $\mathbf{T}_{1,A}$ =matriz triangular inferior de uns $A \times A$.

A matriz de covariâncias de $\hat{\mathbf{F}}$ é obtida aplicando a relação vista no capítulo 3, também conhecida como Método Delta:

$$V(\hat{\mathbf{F}}) \approx [\frac{d\hat{\mathbf{F}}(\mathbf{X})}{d\mathbf{X}}|_{X=\hat{\mathbf{Q}}}] V(\hat{\mathbf{Q}}) [\frac{d\hat{\mathbf{F}}(\mathbf{X})}{d\mathbf{X}}|_{\mathbf{X}=\hat{\mathbf{Q}}}]' , \tag{5.6}$$

$$\frac{d\hat{\mathbf{F}}(\mathbf{X})}{d\mathbf{X}}|_{\mathbf{X}=\hat{\mathbf{Q}}} = \mathbf{D}_{\hat{\mathbf{F}}} [\mathbf{A}_6] \mathbf{D}_{A_7} [\mathbf{A}_4 \mathbf{D}_{\mathbf{a}_1}^{-1} , \mathbf{A}_5 \mathbf{D}_{A_9}^{-1}] \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_3 \mathbf{D}_{\mathbf{a}_2}^{-1} \mathbf{A}_2 \end{bmatrix} ,$$

onde $\mathbf{D}_{\hat{\mathbf{F}}}$ e $\mathbf{D}_{A_7}$ são matrizes diagonais com os elementos de $\hat{\mathbf{F}}$ e $\mathbf{A}_7 = \exp(\mathbf{A}_8)$ na diagonal principal, com

$$\mathbf{A}_8 = [\mathbf{A}_4 , \mathbf{A}_2] \begin{bmatrix} \log \begin{Bmatrix} \mathbf{A}_1 \hat{\mathbf{Q}} \\ \mathbf{A}_3 [\log(\mathbf{A}_2 \hat{\mathbf{Q}})] \end{Bmatrix} \end{bmatrix} ,$$

e $\mathbf{D}_{\mathbf{a}_1}$, $\mathbf{D}_{\mathbf{a}_2}$ e $\mathbf{D}_{A_9}$ são matrizes diagonais com os elementos dos vetores

$$\mathbf{a}_1 = \mathbf{A}_1\hat{\mathbf{Q}} \ , \ \mathbf{a}_2 = \mathbf{A}_2\hat{\mathbf{Q}} \ \text{ e } \ \mathbf{A}_9 = \mathbf{A}_3[\log(\mathbf{a}_2)]$$

nas diagonais principais.

5.3 Ajuste de Modelos e Testes de Hipóteses

Vamos tratar agora o problema da comparabilibade de grupos, ou seja, a análise do efeito de tratamentos, e também, a análise do efeito de covariáveis influentes no tempo de vida, e portanto, influentes na função que estamos estimando, $\mathbf{F}(\mathbf{Q})$. Para tanto, ajustamos o modelo:

$$\mathbf{F}(\mathbf{Q}) = \mathbf{X}\beta \ , \tag{5.7}$$

onde

$\mathbf{X}$ = matriz do delineamento experimental ou das covariáveis, com dimensão $SAK \times d$, $d = \text{posto}(\mathbf{X})$,

β = vetor de parâmetros desconhecidos, com dimensão $d \times 1$.

O vetor β é estimado por mínimos quadrados ponderados:

$$\min_{\beta}(\mathbf{F} - \mathbf{X}\beta)'\mathbf{V}_{\mathbf{F}}^{-1}(\mathbf{F} - \mathbf{X}\beta) \ ,$$

ou seja:

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}_{\mathbf{F}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_{\mathbf{F}}^{-1}\mathbf{F}(\hat{\mathbf{Q}}) \ , \tag{5.8}$$

e o estimador consistente para a matriz de covariâncias de $\hat{\beta}$ é

$$\mathbf{V}(\hat{\beta}) = (\mathbf{X}'\mathbf{V}_{\mathbf{F}}^{-1}\mathbf{X})^{-1} \ . \tag{5.9}$$

O ajuste do modelo é testado através da estatística de Wald:

$$W = (\mathbf{F}(\hat{\mathbf{Q}}) - \mathbf{X}\hat{\beta})'\mathbf{V}_{\mathbf{F}}^{-1}(\mathbf{F}(\hat{\mathbf{Q}}) - \mathbf{X}\hat{\beta}) \ , \tag{5.10}$$

que tem distribuição assintótica χ^2 com $(SAK - d)$ graus de liberdade, sob H_0, onde

$$H_0 : E(\mathbf{F}(\mathbf{Q})) = \mathbf{X}\beta.$$

Se o modelo é satisfatório, isto é, se não há evidências para rejeitarmos que $\mathbf{F(Q)}$ tem a forma dada pela equação (5.7), podemos testar hipóteses lineares com respeito a β do tipo:

$$H_0' : \mathbf{C}\beta = \mathbf{0} \ ,$$

onde $\mathbf{C}$ é uma matriz de contrastes com dimensão $c \times d$, $c \leq d$ e posto completo.

A estatística do teste é

$$\hat{\beta}'\mathbf{C}'[\mathbf{C}\mathbf{V}_{\hat{\beta}}\mathbf{C}']^{-1}\mathbf{C}\hat{\beta} \ , \tag{5.11}$$

que, sob H_0' , tem distribuição assintótica χ^2 com c graus de liberdade.

Capítulo 6

Exemplo

Para aplicar as técnicas vistas nos capítulos anteriores vamos utilizar dados reais de um estudo de observação do tempo até falha da CPU de microcomputadores ITAUTEC I7000 de 8 bits, considerando-se 4 riscos competitivos, ou causas de falha:

Causa 1: problemas na interface serial

Causa 2: problemas na fonte ou de fusível

Causa 3: problemas de mau contato

Causa 4: outros (geralmente troca de componentes)

Neste estudo entraram as CPU's adquiridas pela UNICAMP no início de 1985. Os dados foram coletados a partir das fichas de controle de manutenção do CEMEQ (Centro de Manutenção de Equipamentos) da UNICAMP. A variável medida foi o tempo até ocorrência de falha. É razoável supor que este tipo de variável tenha distribuição exponencial (a suposição será testada mais adiante). Considerou-se como *novo*, ou seja, um micro que ainda não falhou, o micro depois de consertado. Assim, foram obtidas 251 observações, das quais 110 foram observações censuradas à direita, isto é, os micros que até a data da coleta dos dados ainda não haviam falhado. Chamamos a censura de Causa 5. Para cada observação foram anotados:

X: tempo até a falha (em dias)

C_i: causa de falha ($i = 1, \ldots, 5$)

M: mês em que ocorreu a falha.

(a variável M será utilizada como covariável).

Uma tabela com estes dados encontra-se no Apêndice C.

6.1 Análise

A Tabela 6.1 nos dá uma distribuição de frequências dos dados:

Tabela 6.1: Distr. de Frequências do tempo até falha

Classes	Causas de falha					Total
	1	2	3	4	5	
0-100	4	6	8	13	0	31
	1.59	2.39	3.19	5.18	0.00	12.35
100-500	10	17	17	20	0	64
	3.98	6.77	6.77	7.97	0.00	25.50
500-+	11	11	9	15	110	156
	4.38	4.38	3.59	5.98	43.82	62.15
Total	25	34	34	48	110	251
	9.96	13.55	13.55	19.12	43.82	100.00

(em cada cela temos frequência e porcentagem).

6.1.1 Análise Descritiva

Vamos chamar de X o tempo até falha para todas as observações menos as censuradas, e Y_i o tempo até falha para cada causa separadamente. A seguir estão as médias e os desvios padrões amostrais calculados para X,Y_1, Y_2, Y_3, Y_4:

$$\begin{aligned}
\bar{X} &= 372.5 \quad DP_X = 295.8 \\
\bar{Y}_1 &= 444.0 \quad DP_{Y_1} = 295.5 \\
\bar{Y}_2 &= 361.1 \quad DP_{Y_2} = 299.8 \\
\bar{Y}_3 &= 372.8 \quad DP_{Y_3} = 268.0 \\
\bar{Y}_4 &= 343.0 \quad DP_{Y_4} = 314.3 \ .
\end{aligned}$$

Se incluirmos as observações censuradas temos:

$$\bar{X} = 642.2 \quad DP_X = 377.7 \ .$$

A Figura 6.1 contém o histograma e um gráfico da função distribuição acumulada empírica de X. A Figura 6.2 contém os histogramas para Y_1, Y_2, Y_3, Y_4 e as Figuras 6.3 a 6.6 contém os respectivos gráficos das funções distribuições acumuladas empíricas.

Analisando a forma dos histogramas e gráficos, podemos verificar alguma semelhança com a distribuição exponencial. Além disso, na maioria dos casos, o valor médio se aproxima do desvio padrão, o que também é uma característica da distribuição exponencial (se $X \sim \exp(\lambda)$, $E(X) = 1/\lambda$ e $V(X) = 1/\lambda^2$). Com base nestes fatos, foi realizado um teste gráfico do tipo P-P, visto na seção 3.2.1, para verificar se a distribuição dos dados é exponencial (com parâmetros estimados pelas médias calculadas). Os gráficos estão nas Figuras 6.7 a 6.11 , e parece que as distribuições são exponenciais. (Vale observar que Y_1, Y_2, Y_3, Y_4 correspondem aos tempos de vida teóricos definidos no Capítulo 2.)

6.1.2 Estimação

Caso Paramétrico

Se assumimos que Y_1, Y_2, Y_3, Y_4 têm distribuição exponencial, podemos estimar $\lambda_1, \ldots, \lambda_4$, as taxas de falha correspondentes a cada causa, como foi feito no Exemplo 1 da seção 3.1. Estas estimativas foram:

$$\hat{\lambda}_1 = 1/2100.6$$

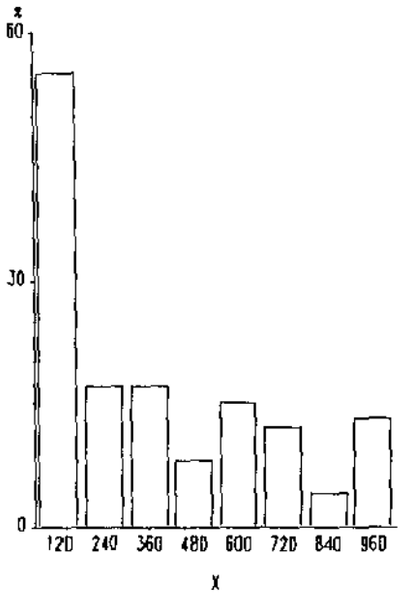
$$\hat{\lambda}_2 = 1/1544.5$$

$$\hat{\lambda}_3 = 1/1544.5$$

$$\hat{\lambda}_4 = 1/1094.1$$

Estas taxas correspondem a considerarmos cada risco na presença de todos outros.

Figura 6.1: Histograma e FDA Empírica de X.



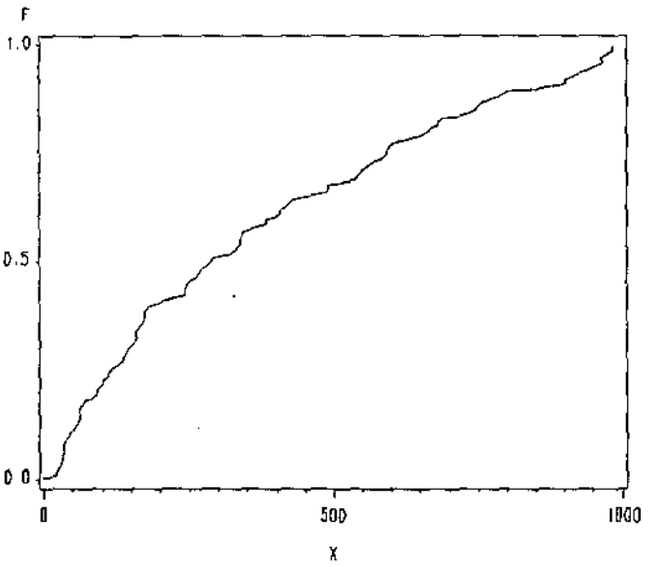


Figura 6.2: Histogramas de Y_1, Y_2, Y_3, Y_4.

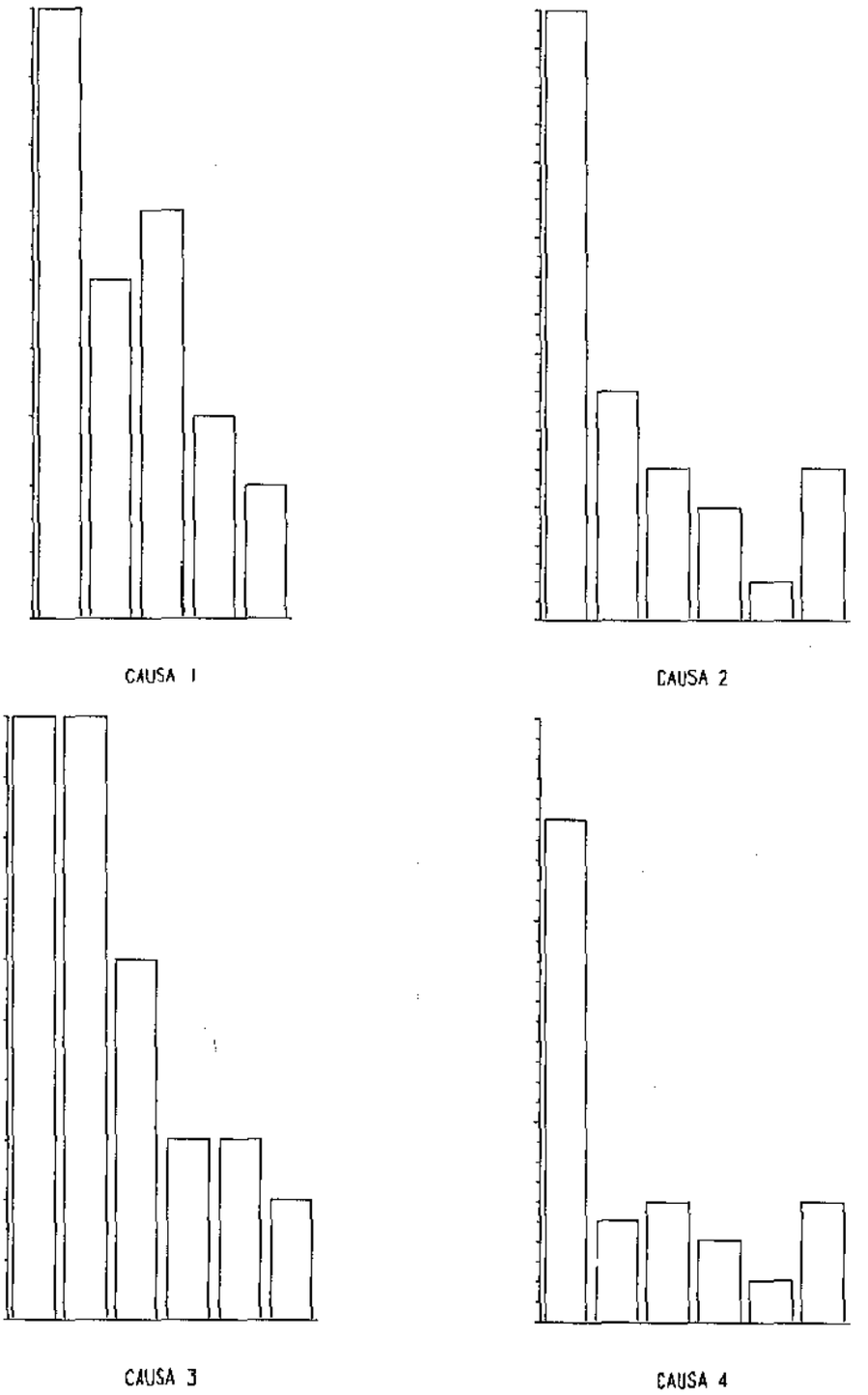


Figura 6.3: FDA Empírica de Y_1.

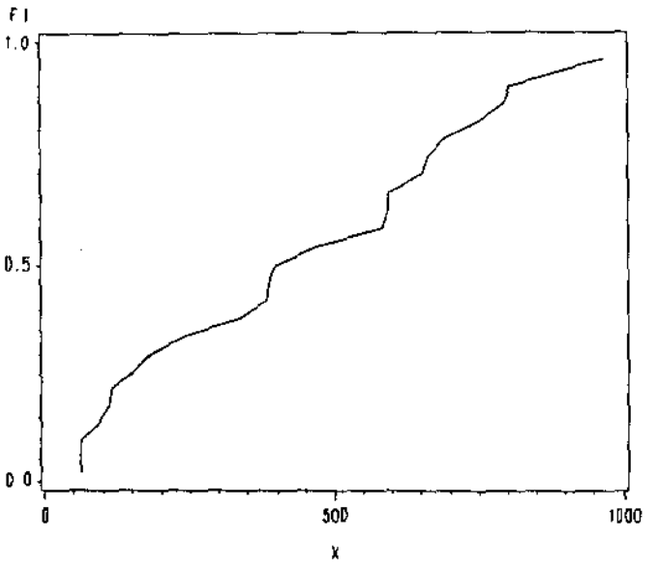


Figura 6.4: FDA Empírica de Y_2.

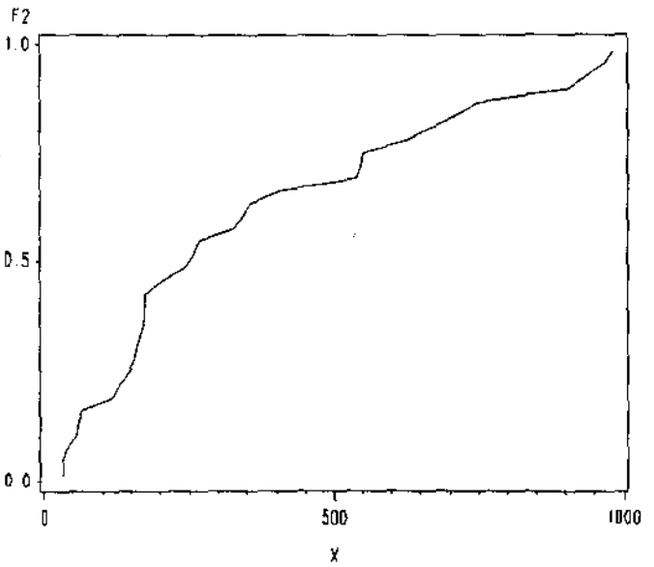


Figura 6.5: FDA Empírica de Y_3.

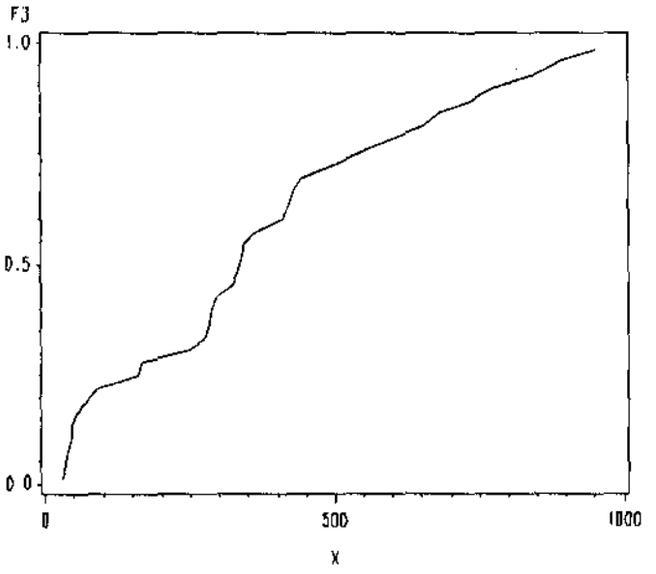


Figura 6.6: FDA Empírica de $Y4$.

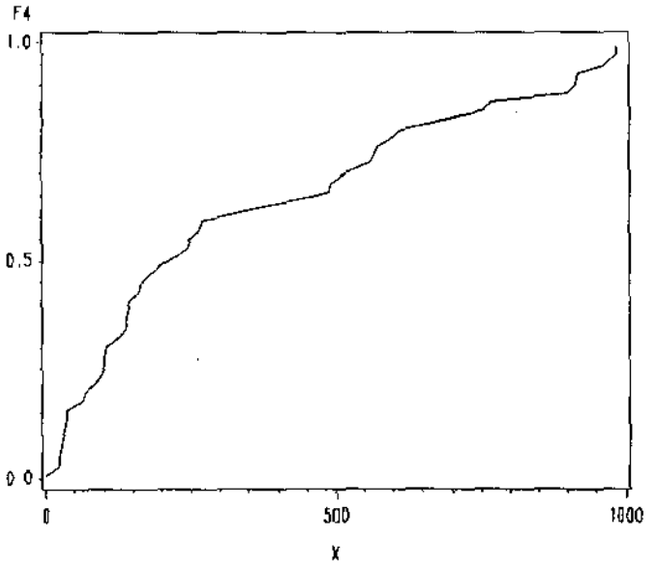


Figura 6.7: Gráfico P-P para X.

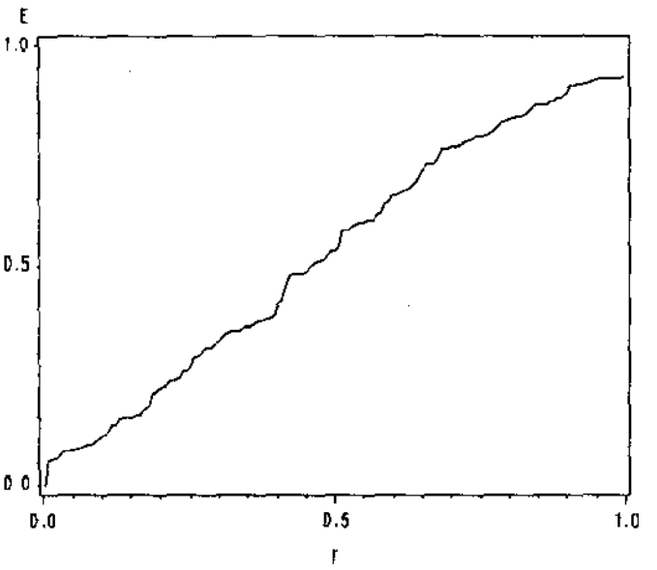


Figura 6.8: Gráfico P-P para Causa 1.

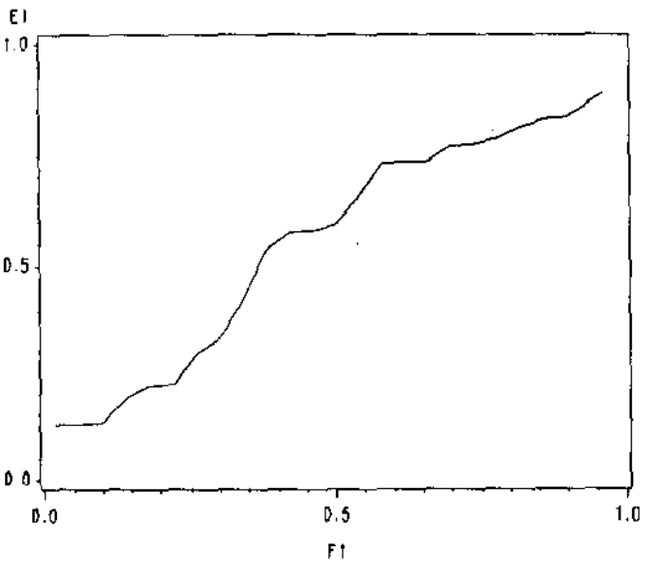


Figura 6.9: Gráfico P-P para Causa 2.

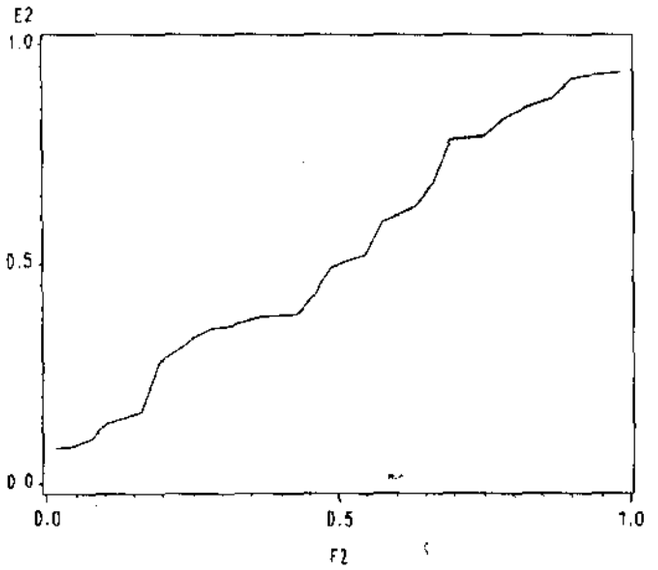


Figura 6.10: Gráfico P-P para Causa 3.

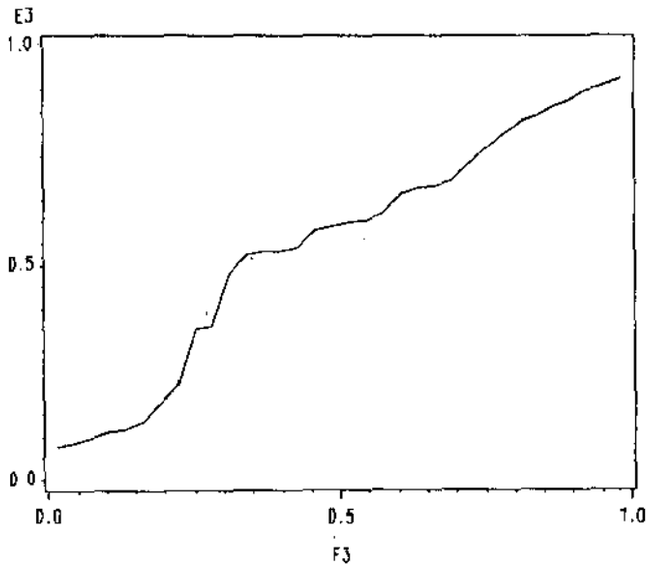
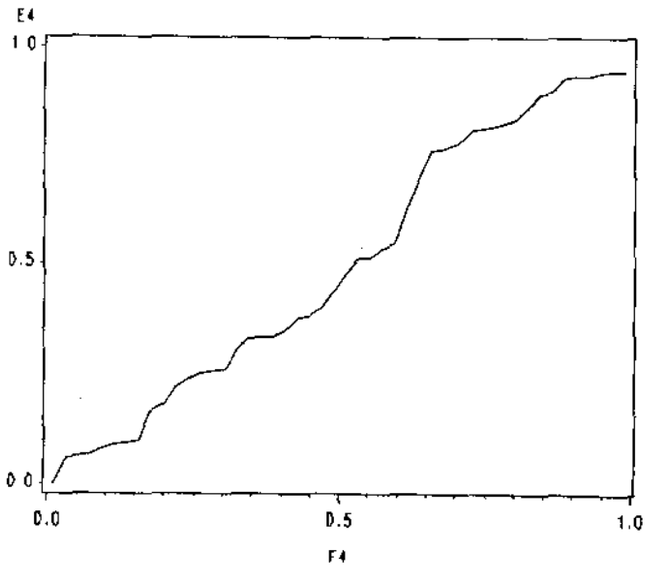


Figura 6.11: Gráfico P-P para Causa 4.



Taxas Proporcionais

Como não estamos trabalhando com dados grupados, iremos apenas testar se as taxas são proporcionais. Esta suposição será necessária mais adiante. Nas Figuras 6.12 a 6.21 temos os gráficos P-P propostos na seção 3.2.1 para efetuar este teste, onde:

F = FDA Empírica de X e

Fi = FDA Empírica de X_i, $i = 1, 2, 3, 4$,

e X_i é o tempo até falha pela Causa i observado, isto é, na presença de todas sa causas. O modelo de taxas proporcionais não é refutado por estes gráficos.

Estimador Kaplan-Meier

Para estimar as curvas de sobrevivência de X e de cada uma das causas, usamos o estimador Kaplan-Meier dado por (3.27). Estas curvas estimadas estão nas Figuras 6.22 a 6.26.

Figura 6.12: Gráfico P-P para X e Causa 1.

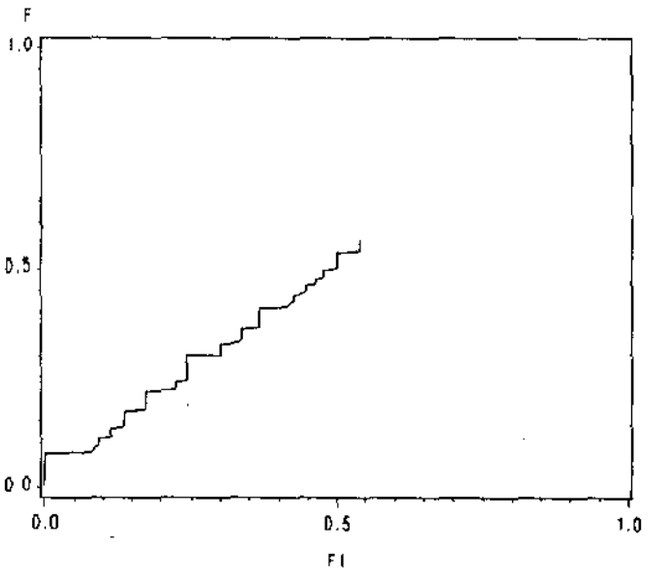


Figura 6.13: Gráfico P-P para X e Causa 2.

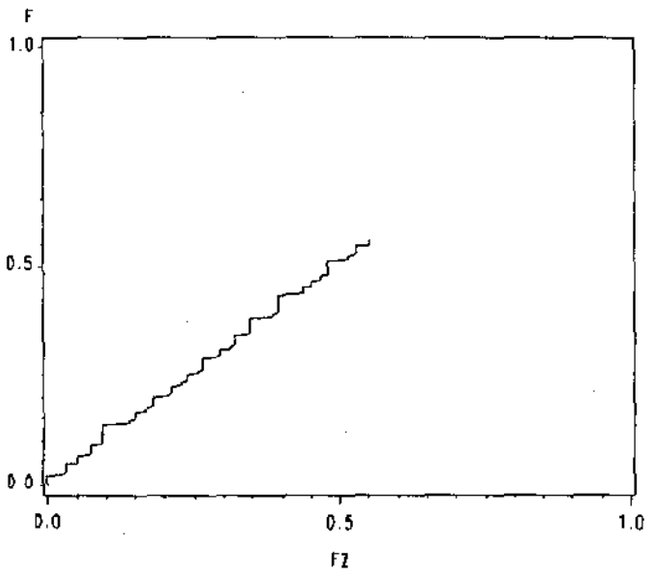


Figura 6.14: Gráfico P-P para X e Causa 3.

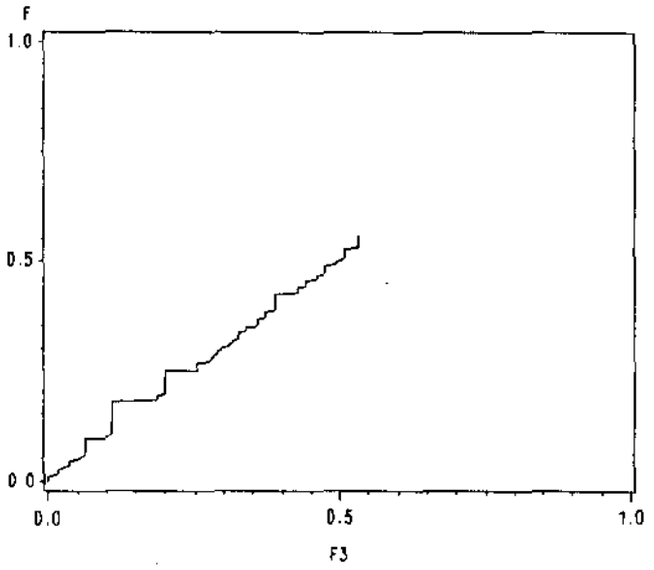


Figura 6.15: Gráfico P-P para X e Causa 4.

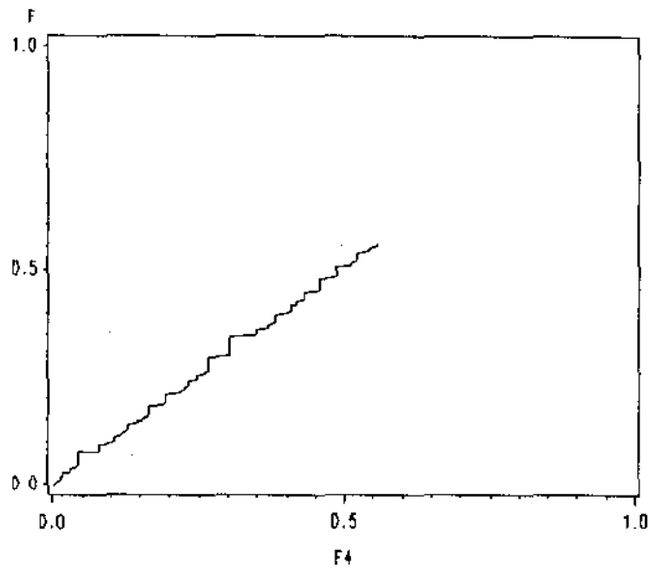


Figura 6.16: Gráfico P-P para Causa 1 e Causa 2.

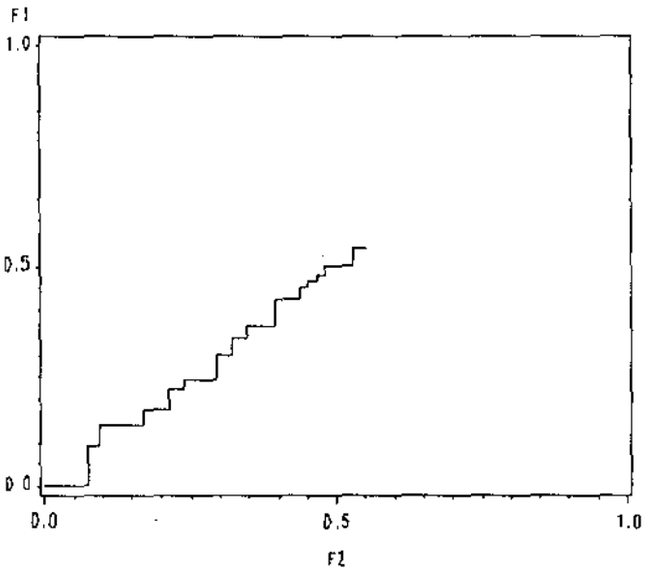


Figura 6.17: Gráfico P-P para Causa 1 e Causa 3.

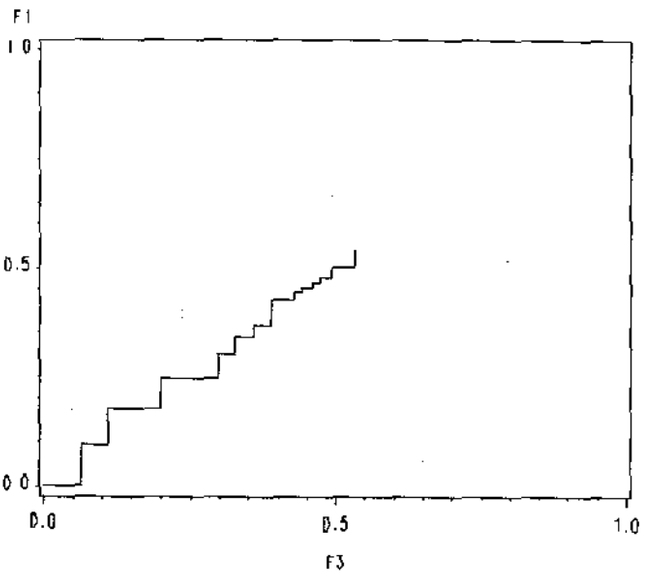


Figura 6.18: Gráfico P-P para Causa 1 e Causa 4.

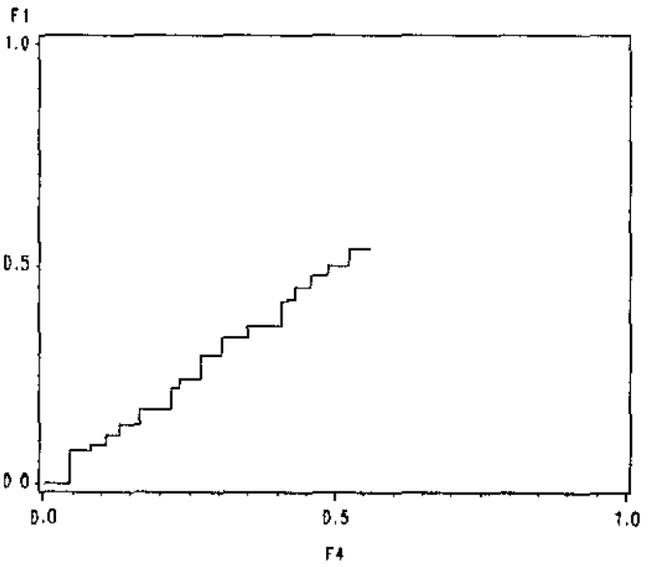


Figura 6.19: Gráfico P-P para Causa 2 e Causa 3.

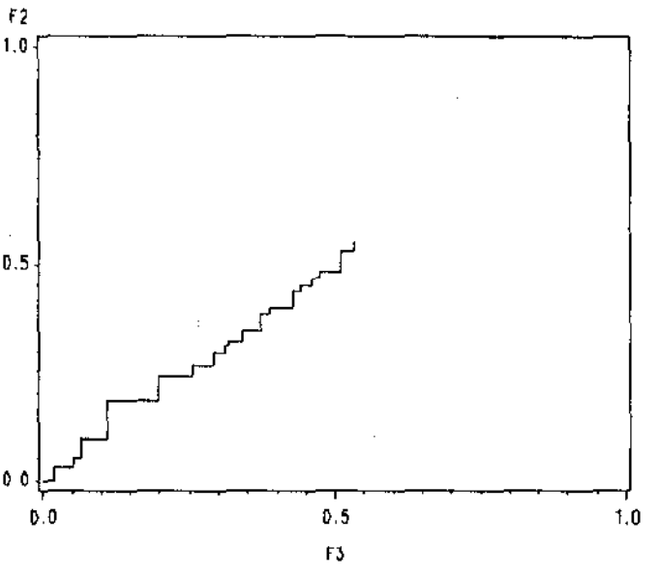


Figura 6.20: Gráfico P-P para Causa 2 e Causa 4.

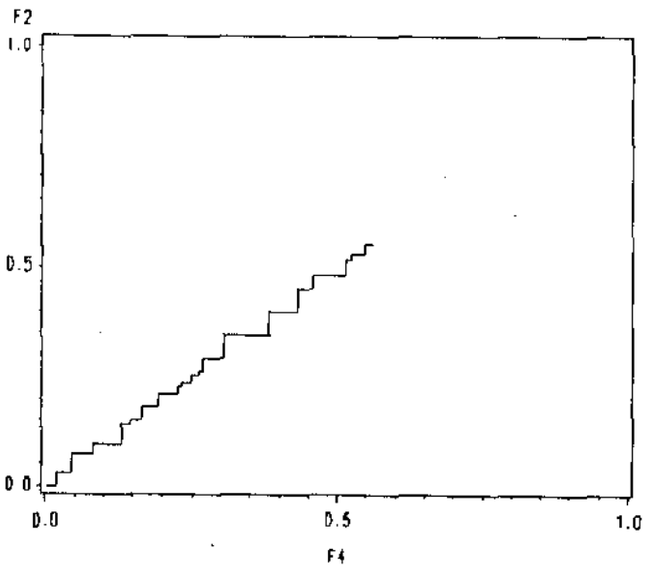


Figura 6.21: Gráfico P-P para Causa 3 e Causa 4.

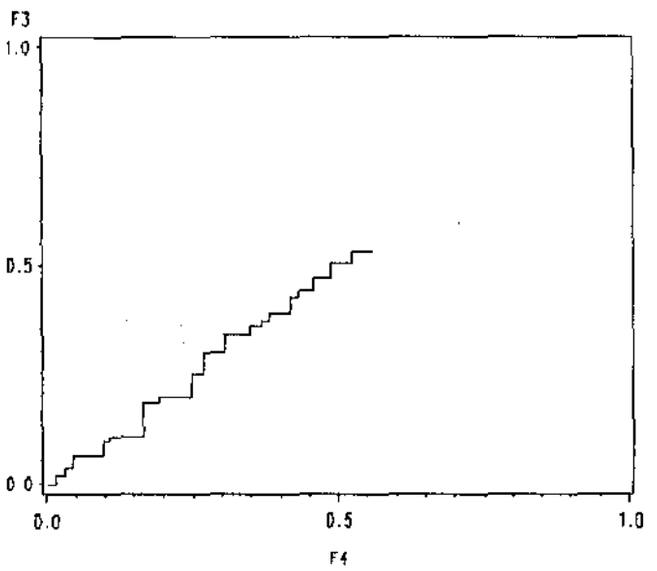


Figura 6.22: Curva de Sobrevivência para X.

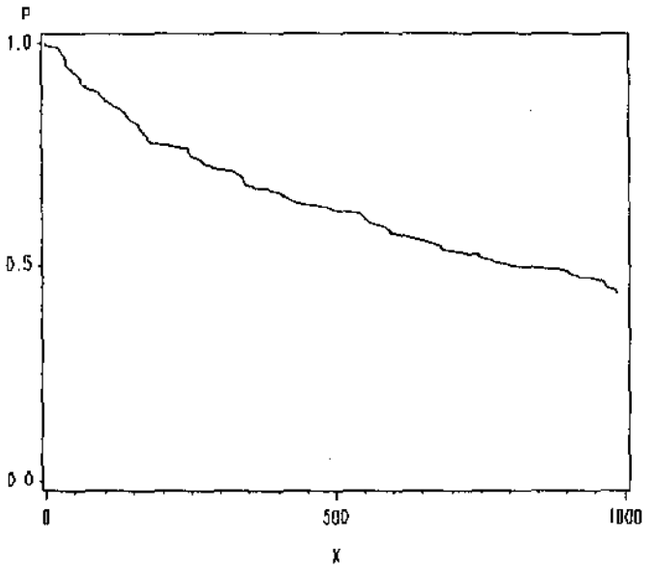


Figura 6.23: Curva de Sobrevivência para Causa 1.

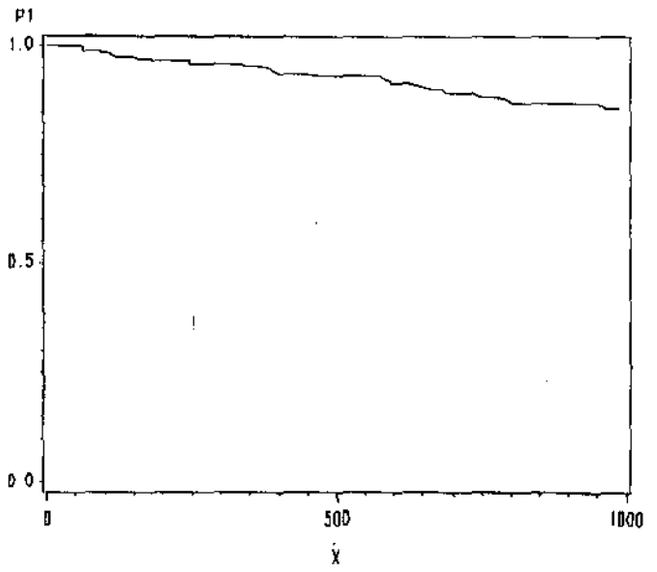


Figura 6.24: Curva de Sobrevivência para Causa 2.

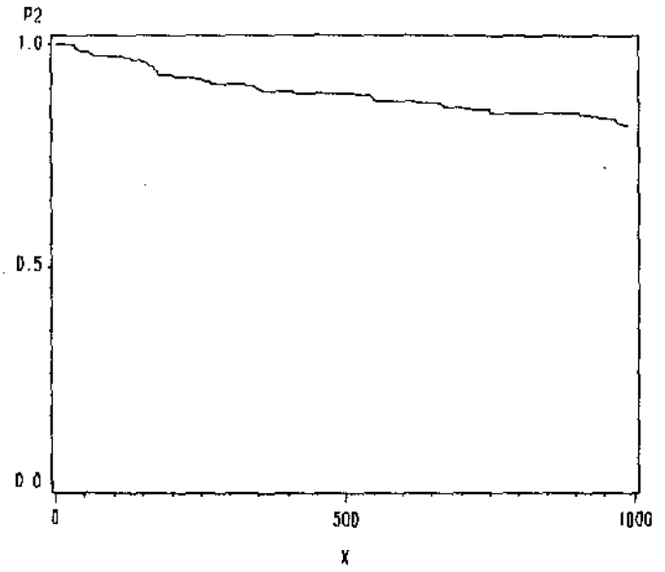


Figura 6.25: Curva de Sobrevivência para Causa 3.

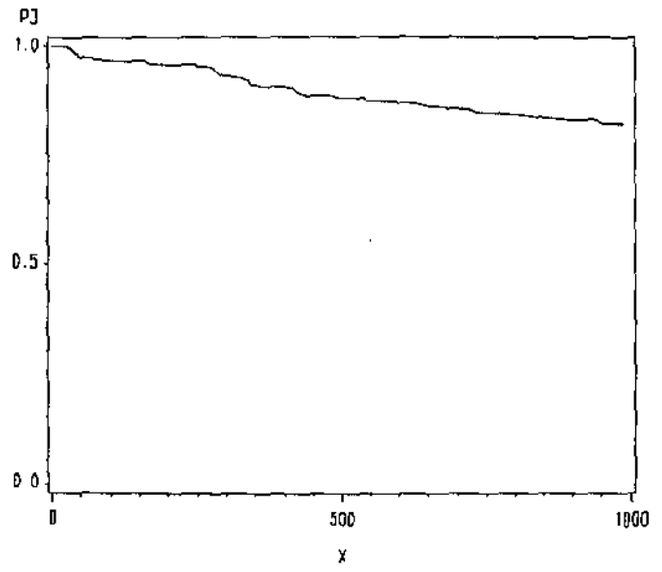
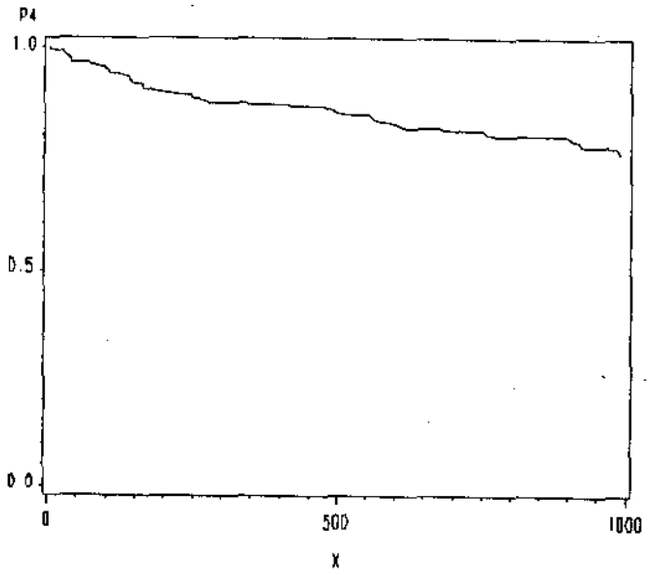


Figura 6.26: Curva de Sobrevivência para Causa 4.



A curva correspondente à Causa 1 (problemas na interface serial) decresce suavemente, ou seja não há indicação de muita ocorrência de problemas na interface serial no início da vida do micro. Já na curva correspondente à Causa 2 (problemas na fonte ou de fusível) percebemos uma declividade mais acentuada para valores pequenos do tempo até falha. Isto também se verifica na curva correspondente à Causa 4 (outros problemas-geralmente troca de componentes). Na curva correspondente à Causa 3 (problemas de mau contato) a declividade se acentua a partir de $x \simeq 250$. Esta causa está, obviamente, associada à qualidade dos contatos das placas, chaves e conectores e dá uma idéia da qualidade da construção dos computadores.

A análise destas curvas, neste exemplo, é importante para o planejamento deste tipo de prestação de serviço no sentido de estar adequadamente capacitado para atender satisfatoriamente à demanda dos diferentes tipos de defeitos, na época em que eles mais ocorrem.

6.1.3 Uso de Covariáveis

Vamos estudar agora a relação entre o tempo até falha e uma covariável. Neste exemplo a covariável considerada é a época do ano em que ocorreu a falha: época de chuvas e época de seca. A idéia é verificar se a taxa de falha está relacionada com a ocorrência de tempestades, onde se verifica a incidência de descargas elétricas. A variável M (mês em que ocorreu a falha) foi codificada como $Z = 0$ para os meses de chuva (outubro a março) e $Z = 1$ para os meses de seca (abril a setembro).

Seguindo a metodologia vista no Capítulo 4, vamos ajustar o modelo dado por (4.2) para $i = 1, 2, 3, 4$. No nosso exemplo, o vetor Z tem dimensão $1x1$, pois temos apenas uma covariável. Este modelo supõe taxas de falha proporcionais, mas como vimos na seção anterior, esta suposição é satisfeita.

Inicialmente ajustamos o modelo para X, isto é, sem considerar causas distintas de falha, e depois para cada uma das causas. Os valores estimados de β e os valores Qui-Quadrado para testar o ajuste dos modelos estão na Tabela 6.2:

Tabela 6.2:

Variável	$\hat{\beta}$	χ^2	$p > \chi^2$
X	1.63	104.09	0.0001
Causa 1	1.80	23.63	0.0001
Causa 2	2.25	47.73	0.0001
Causa 3	0.88	6.40	0.0114
Causa 4	1.69	36.84	0.0001

Vemos que para X, Causa 1, Causa 2 e Causa 4 não há dúvidas em rejeitarmos $\beta = 0$, ou seja, verifica-se que a época do ano está relacionada com a taxa de falha tanto de uma maneira geral (o caso X) como para a Causa 1 (interface serial), Causa 2 (fonte ou fusível) e Causa 4 (outros). No caso da Causa 3 (mau contato), a evidência para rejeitarmos $\beta = 0$ não é tanta ($p > \chi^2 = 0.01$), embora mesmo assim possamos rejeitar a hipótese. É interessante observar que o maior valor de $\hat{\beta}$ ocorreu para problemas de fonte ou fusível. Estes resultados concordam com a idéia intuitiva do assunto, o que pode indicar adequabilidade do modelo.

6.2 Nota Sobre os Cálculos

A ferramenta básica utilizada para os cálculos e gráficos foi o SAS, instalado no computador VAX do Centro de Computação da UNICAMP. No caso da estimação das curvas de sobrevivência foi utilizado o procedimento LIFETEST e, no ajuste do modelo, o COXREGR. Os gráficos foram gerados com o SASGRAPH. Todos os Programas estão no Apêndice A, inclusive um programa SAS-IML (Interactive Matrix Languages) que ajusta todo o procedimento do Capítulo 5, a abordagem de modelos lineares.

6.3 Conclusão

Como pudemos ver neste exemplo, o uso da técnica de riscos competitivos mostrou uma maneira interessante de analisar dados de tempo até falha quando temos várias causas de falha, no sentido de fazermos uma única análise com todas as causas ao invés de analisarmos separadamente cada causa de falha. A análise feita tanto pode ser útil para o planejamento do atendimento do Centro de Manutenção como para indicar caminhos de avanços tecnolólicos para este tipo de equipamento. Já foi iniciada a coleta de dados para extender esta análise para microcomputadores de 16 bits.

Apêndice A

Programas

Lista dos programas:

```
/*ESTIMACAO PARAMETRICA WEIBULL COM 2 RISCOS*/
/*GERACAO DE DADOS*/
DATA GER;
  RETAIN N 100 A1 2 B1 4 A2 4 B2 3;
  DO IND=1 TO N;
    X= (-LOG(RANUNI(0)) A1)**(1/B1);
    Y= (-LOG(RANUNI(0))*A2)**(1/B2);
    IF X < Y THEN DO;
                    U=X;    R=1;
                  END;
             ELSE DO;
                    U=Y;    R=2;
                  END;
    OUTPUT;
  END;
  KEEP U R;
DATA GER1;
  SET GER;
  IF R=1 THEN T=U;
         ELSE DELETE;
  OUTPUT;
DATA GER2;
  SET GER;
  IF R=2 THEN V=U;
         ELSE DELETE;
  OUTPUT;
PROC IML;
  USE GER1;
  READ ALL VAR{T} INTO A;
  USE GER2;
  READ ALL VAR{V} INTO B;
/*CALCULO DOS ESTIMADORES VIA NEWTON-RAPHSON*/
  X={ 2 , 4 };
  PRINT A B X;
  Y=X(\1,\);  Z=X(\2,\);
  D=A##Z;
  E=B##Z;
  G=LOG(A);
  H=LOG(B);
  I=NROW(A);
  K=NROW(B);
```

```
  PRINT I K;
  START SUB(X,Y,Z,D,E,G,H,I);
  F=((SUM(D)+SUM(E))/(Y**2) - I/Y) //
    (I/Z+SUM(G)-(SUM(D*G)+SUM(E*H))/Y) ;
  PRINT F;
  DO WHILE (MAX(ABS(F))>.01);
    J=((I/(Y**2)-(SUM(D)+SUM(E))/(Y**4)) \\
      ((SUM(D*G)+SUM(E)/(Y**2)))//
      (((SUM(D*G)+SUM(E*H))/(Y**3)) \\
      (-I/(Z**2)-(SUM(D*(G**2))+SUM(E*(H**2)))/Y)) ;
    DELTA=-SOLVE(J,F);
    X=X + DELTA;
    Y=X(\1,*\); Z=X(\2,*\);
    F=((SUM(D)+SUM(E))/(Y**2) - I/Y) //
      (I/Z+SUM(G)-(SUM(D*G)+SUM(E*H))/Y) ;
  END;
  FINISH;
  RUN SUB(X,Y,Z,D,E,G,H,I);

  PRINT 'SOLUCAO RISCO 1' X 'RESIDUO' F;
  X=[ 2 , 8 ];
  Y=X(\1,\);   Z=X(\2,\);
  RUN SUB(X,Y,Z,D,E,G,H,K);
  PRINT 'SOLUCAO RISCO 2' X 'RESIDUO' F
```

```
/*TESTE P-P PARA DISTRIBUICAO EXPONENCIAL*/
DATA UM;
  INFILE 'DADOS.DAT';
  INPUT T R U Z F 24;
DATA DOIS;
  SET UM;
  IF R>2 THEN IF R=3 THEN R=2;
                ELSE IF R=4 THEN R=3;
                                ELSE IF R=5 THEN R=4;
                                                 ELSE R=5;
  OUTPUT;
DATA TRES;
  SET DOIS;
  IF R=5 THEN DELETE;
PROC SORT;
  BY T;
PROC RANK OUT=QUATRO;
  VAR T;
  RANKS RT;
```

```
/*CALCULO DA DISTRIBUICAO EMPIRICA PARA T*/
DATA CINCO;
  SET QUATRO NOBS=N;
  F=(RT-0.5)/N;
  E=1-EXP(-(1/372.45)*T);
  OUTPUT;
PROC PRINT;
GOPTIONS DEV=TEK4010;
SYMBOL1 V=NONE I=JOIN;
PROC GPLOT DATA=CINCO;
  AXIS1 ORDER=0 TO 1 BY .2 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1 BY .2 LENGTH=10 CM;
  PLOT E*F=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  TITLE;
DATA SEIS;
  SET QUATRO;
  IF R>1 THEN DELETE;
PROC RANK OUT=SETE;
  VAR T;
  RANKS RT1;
/*CALCULO DA DISTRIBUICAO EMPIRICA PARA CAUSA 1*/
DATA OITO;
  SET SETE NOBS=N1;
  F1=(RT1-0.5)/N1;
  E1=1-EXP(-(1/442.96)*T);
  OUTPUT;
PROC GPLOT;
  AXIS1 ORDER=0 TO 1 BY .2 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1 BY .2 LENGTH=10 CM;
  PLOT E1*F1=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  TITLE;
DATA NOVE;
  SET QUATRO;
  IF R>2 OR R<2 THEN DELETE;
PROC RANK OUT=DEZ;

  VAR T;
  RANKS RT2;
/*CALCULO DA DISTRIBUICAO EMPIRICA PARA CAUSA 2*/
DATA ONZE;
  SET DEZ NOBS=N2;
  F2=(RT2-0.5)/N2;
  E2=1-EXP(-(1/361.088)*T);
  OUTPUT;
```

```
PROC GPLOT;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1 BY .5 LENGTH=10 CM;
  PLOT E2*F2=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  TITLE;
DATA DOZE;
  SET QUATRO;
  IF R>3 OR R<3 THEN DELETE;
PROC RANK OUT=TREZE;
  VAR T;
  RANKS RT3;
/*CALCULO DA DISTRIBUICAO EMPIRICA PARA CAUSA 3*/
DATA QUATORZE;
  SET TREZE NOBS=N3;
  F3=(RT3-0.5)/N3;
  E3=1-EXP(-(1/372.824)*T);
  OUTPUT;
PROC GPLOT;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1 BY .5 LENGTH=10 CM;
  PLOT E3*F3=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  TITLE;
DATA QUINZE;
  SET QUATRO;
  IF R<4 \ R>4 THEN DELETE;
PROC RANK OUT=QUINZE1;
  VAR T;
  RANKS RT4;
/*CALCULO DA DISTRIBUICAO EMPIRICA PARA CAUSA 4*/
DATA QUINZE2;
  SET QUINZE1 NOBS=N4;
  F4=(RT4-0.5)/N4;
  E4=1-EXP(-(1/342.979)*T);
  OUTPUT;
PROC GPLOT;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1 BY .5 LENGTH=10 CM;
  PLOT E4*F4=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  TITLE;
-------------------------------------------------------------
/*ESTIMACAO DE L1 L2 L3 L4/DISTRIBUICAO EXPONENCIAL*/
DATA UM;
  INFILE 'DADOS.DAT';
  INPUT T R U Z N &1;
```

```
PROC TRANSPOSE DATA=UM(KEEP=T) OUT=CINCO PREFIX= TT;
PROC TRANSPOSE DATA=UM(KEEP=R) OUT=SEIS PREFIX= RR;
/*CALCULO DOS ESTIMADORES*/
DATA SETE;
  RETAIN K 5 N 251 S 0;
  SET CINCO;
  ARRAY TT[251] TT1-TT251;
  SET SEIS;

  ARRAY RR[251] RR1-RR251;
  ARRAY NU[5] NU1-NU5;
  DO J=1 TO K;
    NU[J]=0;
  END;
  DO I=1 TO N;
    S=S+TT[I];
    DO J=1 TO K;
      IF RR[I]=J THEN  NU[J]=NU[J]+1;
    END;
  END;
  ARRAY L[5] L1-L5;
  ARRAY LL[5] LL1-LL5;
  DO J=1 TO K;
    L[J]=S/NU[J];
    LL[J]=1/L[J];
  END;
PROC PRINT;
DATA _NULL_;
  SET SETE;
  FILE PRINT HEADER=H;
  PUT @3 NU1 @7 L1 @16 LL1 @33 NU2 @37 L2 @46  LL2 @63
  NU3 @67 L3 @76 LL3 @93 NU4 @97 L4 @106 LL4;
  RETURN;
  H: PUT / @10 'RISCO 1' @40 'RISCO 2' @70 'RISCO 3'
           @100 'RISCO 4'
     / @3 'N1' @7 'L1' @16 '1/L1' @33 'N2' @37 'L2'
       @46 '1/L2' @63 'N3' @67 'L3' @76 '1/L3' @93
       'N4' @97 'L4' @106 '1/L4' / ;
  RETURN;
  FILE PRINT;
-------------------------------------------------------------
/*TESTE PARA TAXAS PROPORCIONAIS*/
DATA UM;
  INFILE 'DADOS.CAT';
  INPUT T R U Z $ @@;
```

```
PROC SORT OUT=QUATRO;
  BY T;
PROC RANK OUT=CINCO;
  VAR T; RANKS RT;
/*CALCULO DAS DISTRIBUICOES EMPIRICAS*/
DATA SEIS;
  RETAIN F1 0 F2 0 F3 0 F4 0;
  SET CINCO NOBS=N;
  F=(RT-0.5)/N;
  IF R=1 THEN DO; F1=(RT-0.5)/N; F2=F2; F3=F3; F4=F4;
               END;
         ELSE IF R=2 THEN DO; F1=F1; F2=(RT-0.5)/N;
                              F3=F3; F4=F4;
                         END;
                    ELSE IF R=3 THEN DO; F1=F1; F2=F2;
                                         F3=(RT-0.5)/N;
                                         F4=F4; END;
                               ELSE IF R=4 THEN DO;
                                         F1=F1; F2=F2;
                                         F3=F3; F4=(RT-0.5)/N;
                                                END;
                                    ELSE DELETE;
KEEP T RT F F1 F2 F3 F4;
GOPTIONS DEV=TEK4010;

SYMBOL1 V=NONE I=JOIN;
PROC GPLOT DATA=SEIS;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1 BY .5 LENGTH=10 CM;
  PLOT F*F1=1 F*F2=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  PLOT F*F3=1 F*F4=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  PLOT F1*F2=1 F1*F3=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  PLOT F1*F4=1 F2*F3=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  PLOT F2*F4=1 F3*F4=1 / VAXIS=AXIS1 HAXIS=AXIS2;
  TITLE;
-------------------------------------------------------------------
\*ESTIMADOR KAPLAN-MEIER*/
DATA UM;
  INFILE "DADOS.DAT";
  INPUT T R U I N ;
PROC SORT OUT=DI;
  BY T;
/*CALCULO DO ESTIMADOR PARA T*/
PROC LIFETEST  OUTS=QUATRO;
  TIME T*R(5);
```

```
DATA QUATRO1;
  SET QUATRO;
  P=SURVIVAL; X=T;
  KEEP P X;
GOPTIONS DEVICE=TEK4010;
FOOTNOTE1 "CURVA DE SOBREVIVENCIA PARA X";
SYMBOL1 V=NONE I=JOIN;
PROC GPLOT DATA=QUATRO1;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1000 BY 500 LENGTH=10 CM;
  PLOT P*X=1 / VAXIS=AXIS1 HAXIS=AXIS2 FRAME;
  TITLE;
/*CALCULO DO ESTIMADOR PARA CAUSA 1*/
PROC LIFETEST DATA=UM OUTS=CINCO;
  TIME T*R(2,3,4,5);
DATA CINCO1;
  SET CINCO;
  P1=SURVIVAL; X=T;
  KEEP P1 X;
FOOTNOTE1 "SOBREVIVENCIA PARA CAUSA 1";
PROC GPLOT DATA=CINCO1;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1000 BY 500 LENGTH=10 CM;
  PLOT P1*X=1 / VAXIS=AXIS1 HAXIS=AXIS2 FRAME;
  TITLE;
/*CALCULO DO ESTIMADOR PARA CAUSA 2*/
PROC LIFETEST DATA=UM OUTS=SEIS;
  TIME T*R(1,3,4,5);
DATA SEIS1;
  SET SEIS;
  P2=SURVIVAL; X=T;
  KEEP P2 X;
FOOTNOTE1 "CURVA DE SOBREVIVENCIA PARA RISCO 2";
PROC GPLOT DATA=SEIS1;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1000 BY 500 LENGTH=10 CM;
  PLOT P2*X=1 / VAXIS=AXIS1 HAXIS=AXIS2 FRAME;
  TITLE;
/*CALCULO DO ESTIMADOR PARA CAUSA 3*/
PROC LIFETEST DATA=UM OUTS=SETE;

  TIME T*R(1,2,4,5);
DATA SETE1;
  SET SETE;
  P3=SURVIVAL; X=T;
  KEEP P3 X;
```

```
FOOTNOTE1 "CURVA DE SOBREVIVENCIA PARA RISCO 3";
PROC GPLOT DATA=SETE1;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1000 BY 500 LENGTH=10 CM;
  PLOT P3*X=1 / VAXIS=AXIS1 HAXIS=AXIS2 FRAME;
  TITLE;
/*CALCULO DO ESTIMADOR PARA CAUSA 4*/
PROC LIFETEST DATA=UM NOPRINT OUTS=OITO;
  TIME T*R(1,2,3,5);
DATA OITO1;
  SET OITO;
  P4=SURVIVAL; X=T;
  KEEP P4 X;
FOOTNOTE1 "CURVA DE SOBREVIVENCIA PARA RISCO 4";
PROC GPLOT DATA=OITO1;
  AXIS1 ORDER=0 TO 1 BY .5 LENGTH=7 CM;
  AXIS2 ORDER=0 TO 1000 BY 500 LENGTH=10 CM;
  PLOT P4*X=1 / VAXIS=AXIS1 HAXIS=AXIS2 FRAME;
  TITLE;
-------------------------------------------------------------
/*AJUSTE DO MODELO DE  COX*/
DATA UM;
  INFILE "DADOS.DAT";
  INPUT T R U Z M 20;
PROC SORT OUT=DI;
  BY T;
DATA TRES;
  SET DI;
  IF R=1 THEN R1=2;
         ELSE R1=1;
  IF R=2 THEN R2=2;
         ELSE R2=1;
  IF R=3 THEN R3=2;
         ELSE R3=1;
  IF R=4 THEN R4=2;
         ELSE R4=1;
  IF R=5 THEN R5=1;
         ELSE R5=2;
  IF M>9 OR M<4 THEN Z=0;
                ELSE Z=1;
  KEEP T R1 R2 R3 R4 R5 Z;
PROC PRINT;
/*AJUSTE DO MODELO PARA T*/
PROC COXREGR PRINT=3;
  MODEL T R5=Z;
TITLE3 'MODELO DE COX PARA T';
```

```
/*AJUSTE DO MODELO PARA CAUSA 1*/
PROC COXREGR DATA=TRES PRINT=3;
  MODEL T R1=Z;
TITLE3 'MODELO DE COX PARA RISCO 1';
/*AJUSTE DO MODELO PARA CAUSA 2*/
PROC COXREGR DATA=TRES PRINT=3;
  MODEL T R2=Z;
TITLE3 'MODELO DE COX PARA RISCO 2';
/*AJUSTE DO MODELO PARA CAUSA 3*/

PROC COXREGR DATA=TRES PRINT=3;
  MODEL T R3=Z;
TITLE3 'MODELO DE COX PARA RISCO 3';
/*AJUSTE DO MODELO PARA CAUSA 4*/
PROC COXREGR DATA=TRES PRINT=3;
  MODEL T R4=Z;
TITLE3 "MODELO DE COX PARA RISCO 4";
-------------------------------------------------------------------
/*RISCOS COMPETITIVOS VIA MODELOS LINEARES UTILIZANDO IML*/;
PROC IML;
USE DATA; READ ALL VAR{X1} INTO G;
USE DATA; READ ALL VAR{N1} INTO N2;
/*S=NUMERO DE SUBPOPULACOES (ARGUMENTO DE I3)*/
/*A=NUMERO DE INTERVALOS (ARGUMENTO DE I4,J1)*/
/*K=NUMERO DE RISCOS COMPETITIVOS (ARGUMENTO DE I1,D1,U1)*/
/*SA=ARGUMENTO DE I2*/
SA=S*A; T=(SA*K)+(SA); T2=(A*K)+A; T4=K+1;
I1=I(K);
START;
/*CRIACAO DAS MATRIZES AUXILIARES*/
L=0;
DO I=1 TO SA;
  LL=I+L; L=L+K;
  B1=J(T4,T4,N2(\LL,\));
  B2=J(T4,T4,1);
  IF I>1 THEN DO;
                I3=I-2; B3=B2;
                IF I>2 THEN DO II=1 TO T3;
                              B3=B3\\B2;
                            END;
                B1=B3\\B1;
              END;
  TT3=I+1;
  DO J=TT3 TO SA;
    B1=B1\\B2;
  END;
```

```
  IF I=1 THEN N=B1;
         ELSE N=N//B1;
END;
D0=J(1,1,0);
U0=J(1,1,1);
D1=J(K,1,0);
U1=J(K,1,1);
I2=I(SA);
I3=I(S);
J1=J(A,1,1);
T1=HANKEL(J1);
TT=(T1(\,A\))@;
AA=A-1;
DO I=AA TO 1 BY -1;
  TT=TT//(T1(\,I\))@;
END;
A1=I2@((I1\\D1)//(U1@\\D0));
A2=I2@(D1@\\U0);
A3=-I2;
A4=I2@(I1\\-U1);
A5=I2@U1;
A6=I3@(I1@TT);
QQ=Q*Q@;
I=1; II=I+K;

BL=QQ(\I:II,I:II\);
DO III=2 TO SA;
  I=I+T4; II=I+K;
  BLL=QQ(\I:II,I:II\);
  BL=BLOCK(BL,BLL);
END;
/*CALCULO DA MATRIZ DE COVARIANCIA DE Q*/
VQ=(DIAG(Q)-BL)/K;
PRINT "MATRIZ DE COVARIANCIA DE Q" VQ;
K1=A1*Q;
K2=A2*Q;
K3=LOG(A3*(LOG(K2)));
K4=LOG(K1//K3);
/*CALCULO DA FUNCAO F*/
F=EXP((-A6)*EXP((A4\\A5)*K4));
G=LOG(F);
A8=((A4\\A5)*K4);
A7=EXP(A8);
A9=A3*(LOG(K2));
D1=INV(DIAG(K1));
```

```
D2=INV(DIAG(K2));
D3=INV(DIAG(K3));
D7=INV(DIAG(A7));
DF=DIAG(F);
/*CALCULO DA DERIVADA DE F*/
FQ=DF*A6*D7*((A4*D1)\\(A5*D3))*(A1//(A3*D2*A2));
/*CALCULO DA MATRIZ DE COVARIANCIA DE F*/
VF=FQ*VQ*(FQ@);
IDF=INV(DF);
VG=IDF*VF*IDF;
PRINT "VETOR ESTIMADO F(I)" F;
PRINT "MATRIZ DE COVARIANCIA DE F" VF;
PRINT "VETOR G=LOG(F)" G;
PRINT "MATRIZ DE COVARIANCIA VG" VG;
/*AJUSTE DE MODELOS E TESTES DE HIPOTESES */
X=(-I1//-2*I1//-3*I1//-4*I1//-5*I1)@(I(2));
IVF=GINV(VF);
B=(INV(X@*IVF*X))*X@*IVF*F;
VB=INV(X@*IVF*X);
W=(F-(X*B))@*IVF*(F-(X*B));
GE=X*B;
PRINT "VETOR DE PAR EST B" B
      "MATRIZ DE COVARIANCIA DE B" VB;
PRINT "ESTATISTICA DE WALD" W;
PRINT "G ESTIMADO GE" GE;
IVG=GINV(VG);
B1=(INV(X@*IVG*X))*X@*IVG*G;
VB1=INV(X@*IVG*X);
W1=(G-(X*B1))@*IVG*(G-(X*B1));
GE1=X*B1;
PRINT "VETOR DE PAR EST B1" B1
      "MATRIZ DE COVARIANCIA DE B1" VB1;
PRINT "ESTATISTICA DE WALD" W1;
PRINT "G ESTIMADO GE1" GE1;
FINISH;
RUN;
```

Apêndice B

Método Delta

B.1

Verificação da aproximação dada no capítulo 3 em (3.22):

$$E(\frac{1}{S_a}) \approx \frac{1}{E(S_a)}.$$

Vamos considerar o caso geral onde X é uma variável aleatória com f.d.p. f(x). Assim,

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

e

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx\ .$$

Para g suave, fazendo uma expansão em série em torno do ponto $\eta = E(X)$, temos:

$$g(x) = g(\eta) + g'(\eta)(x-\eta) + g''(\eta)\frac{(x-\eta)^2}{2} + \cdots + g^{(n)}(\eta)\frac{(x-\eta)^n}{n!} + \cdots\ .$$

Aplicando a esperança:

$$\begin{aligned} E(g(x)) &= g(\eta) + g'(\eta)E(x-\eta) + g''(\eta)\frac{E(x-\eta)^2}{2} + \cdots + \\ &\quad + g^{(n)}(\eta)\frac{E(x-\eta)^n}{n!} + \cdots \\ &= g(\eta) + g''(\eta)\frac{\sigma^2}{2} + \cdots + g^{(n)}(\eta)\frac{\mu_n}{n!} + \cdots\ , \end{aligned} \qquad (.1)$$

onde μ_n é o n-ésimo momento de X.

Desta forma, temos aproximadamente

$$E(g(X)) \approx g(\eta)\ ,$$

e fazendo $g(\eta) = 1/\eta$ e $\eta = E(S_a)$ temos:

$$E(\frac{1}{S_a}) \approx \frac{1}{E(S_a)}\ .\diamond$$

B.2

Verificação da aproximação vista em (3.23).

Como anteriormente, seja x uma variável aleatória com f.d.p. f(x). Sabemos que

$$V(g(x)) = E(g^2(x)) - E^2(g(x)) \ .$$

De (1) temos que

$$E(g(x)) \approx g(\eta) + g''(\eta)\frac{\sigma^2}{2} \ .$$

Seja $h(x) = g^2(x)$, então

$$\begin{aligned} h'(x) &= 2g(x)g'(x) \quad \text{e} \\ h''(x) &= 2g(x)g''(x) + 2g'(x)g'(x) \\ &= 2g'(x)^2 + 2g(x)g''(x) \ . \end{aligned}$$

No ponto $x = \eta$, por (1), temos

$$\begin{aligned} E(h(x)) &\approx h(\eta) + h''(\eta)\frac{\sigma^2}{2} \\ &\approx g^2(\eta) + [2g'(\eta)^2 + 2g(\eta)g''(\eta)]\frac{\sigma^2}{2} \ . \end{aligned}$$

Logo

$$\begin{aligned} V(g(x)) &= E(g^2(x)) - E^2(g(x)) \\ &\approx g^2(\eta) + [2g'(\eta)^2 + 2g(\eta)g''(\eta)]\frac{\sigma^2}{2} - \\ &\quad -[g^2(\eta) + 2g(\eta)g''(\eta)\frac{\sigma^2}{2} + g''(\eta)^2\frac{\sigma^4}{4}] \\ &\approx g'(\eta)^2\sigma^2 - g''(\eta)^2\frac{\sigma^4}{4} \\ &\approx [g'(\eta)]^2\sigma^2 \ . \end{aligned}$$

Para duas variáveis, x e y, temos

$$V(g(x,y)) \approx [\frac{\partial g}{\partial x} \ \ \frac{\partial g}{\partial y}] \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix}$$

$$\approx [\frac{\partial g}{\partial x}\sigma_X^2 + dy\sigma_{XY}]\frac{\partial g}{\partial x} + [\frac{\partial g}{\partial x}\sigma_{XY} + dy\sigma_Y^2]dy$$
$$\approx [\frac{\partial g}{\partial x}]^2\sigma_X^2 + [\frac{\partial g}{\partial y}]^2\sigma_Y^2 + 2[\frac{\partial g}{\partial x}][\frac{\partial g}{\partial y}]\sigma_{XY} ,$$

com as derivadas calculadas no ponto

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} E(X) \\ E(Y) \end{bmatrix} . \diamond$$

Apêndice C

Dados do Exemplo

Estes dados foram obtidos no CEMEQ (Centro de Manutenção) da UNICAMP. As variáveis anotadas foram:

X = tempo até falha (em dias),

C = causa de falha e

M = mês em que ocorreu a falha.

OBS	X	C	M	OBS	X	C	M
1	0	4	5	34	111	1	2
2	22	4	4	35	113	1	3
3	23	4	3	36	119	2	10
4	24	4	8	37	127	4	10
5	28	3	2	38	129	2	10
6	30	4	10	39	137	4	5
7	31	2	5	40	139	4	4
8	32	2	10	41	140	4	1
9	32	4	3	42	145	4	8
10	34	3	2	43	146	2	1
11	35	4	2	44	155	1	9
12	36	4	10	45	159	2	1
13	39	3	7	46	160	2	8
14	39	2	6	47	161	3	12
15	45	3	2	48	162	4	7
16	46	3	11	49	165	4	7
17	53	3	4	50	167	3	9
18	54	2	5	51	174	2	7
19	59	2	5	52	174	3	7
20	61	1	6	53	174	2	7
21	61	4	5	54	175	2	7
22	62	1	7	55	176	4	12
23	62	1	4	56	181	1	1
24	64	2	6	57	198	4	8
25	68	4	8	58	202	2	4
26	73	3	4	59	223	4	6
27	86	4	1	60	242	2	9
28	90	3	9	61	243	1	4
29	93	1	2	62	244	4	10
30	94	4	9	63	245	4	2
31	100	4	6	64	246	3	10
32	101	4	6	65	256	2	6
33	104	4	10	66	262	4	9

67	256	2	9
68	270	4	10
69	276	3	10
70	284	3	10
71	285	3	11
72	291	3	11
73	323	3	11
74	325	2	6
75	333	3	3
76	336	1	5
77	339	4	2
78	340	3	12
79	342	2	9
80	343	3	11
81	365	2	1
82	369	3	1
83	381	1	2
84	386	1	1
85	389	1	2
86	407	3	5
87	409	2	2
88	414	4	3
89	429	3	2
90	427	3	4
91	442	3	4
92	471	1	10
93	488	4	6
94	489	3	6
95	490	4	6
96	510	4	8
97	529	2	7
98	542	3	6
99	546	2	7
100	551	2	8
101	559	4	8
102	569	4	8
103	574	4	3
104	581	1	9
105	583	1	9
106	588	4	5
107	594	1	9
108	595	3	9
109	612	4	10
110	625	2	9
111	651	1	11

112	691	3	10
113	693	1	11
114	672	2	11
115	681	4	11
116	683	3	12
117	689	1	10
118	719	2	1
119	725	3	2
120	747	1	1
121	748	2	2
122	754	4	2
123	771	4	10
124	774	3	3
125	795	1	4
126	804	1	4
127	850	3	4
128	891	3	6
129	903	4	7
130	904	2	7
131	914	4	8
132	918	4	8
133	939	2	7
134	951	3	9
135	964	1	9
136	964	1	9
137	966	4	9
138	967	2	9
139	981	1	9
140	985	4	7
141	986	4	10
142	989	5	10
143	990	5	10
144	993	5	10
145	998	5	10
146	999	5	10
147	999	5	10
148	999	5	10
149	999	5	10
150	998	5	10
151	999	5	10
152	999	5	10
153	998	5	10
154	998	5	10
155	998	5	10
156	998	5	10

157	988	5	10	207	988	5	10
158	988	5	10	208	988	5	10
159	988	5	10	209	988	5	10
160	988	5	10	210	988	5	10
161	988	5	10	211	988	5	10
162	988	5	10	212	988	5	10
163	988	5	10	213	988	5	10
164	988	5	10	214	988	5	10
165	988	5	10	215	988	5	10
166	988	5	10	216	988	5	10
167	988	5	10	217	988	5	10
168	988	5	10	218	988	5	10
169	988	5	10	219	988	5	10
170	988	5	10	220	988	5	10
171	988	5	10	221	988	5	10
172	988	5	10	222	988	5	10
173	988	5	10	223	988	5	10
174	988	5	10	224	988	5	10
175	988	5	10	225	988	5	10
176	988	5	10	226	988	5	10
177	988	5	10	227	988	5	10
178	988	5	10	228	988	5	10
179	988	5	10	229	988	5	10
180	988	5	10	230	988	5	10
181	988	5	10	231	988	5	10
182	988	5	10	232	988	5	10
183	988	5	10	233	988	5	10
184	988	5	10	234	988	5	10
185	988	5	10	235	988	5	10
186	988	5	10	236	988	5	10
187	988	5	10	237	988	5	10
188	988	5	10	238	988	5	10
189	988	5	10	239	988	5	10
190	988	5	10	240	988	5	10
191	988	5	10	241	988	5	10
192	988	5	10	242	988	5	10
193	988	5	10	243	988	5	10
194	988	5	10	244	988	5	10
195	988	5	10	245	988	5	10
196	988	5	10	246	988	5	10
197	988	5	10	247	988	5	10
198	988	5	10	248	988	5	10
199	988	5	10	249	988	5	10
200	988	5	10	250	988	5	10
201	988	5	10	251	988	5	10
202	988	5	10	252	988	5	10

249	[illegible]	5	10
250	[illegible]	5	10
251	[illegible]	5	10

Bibliografia

[1] Breslow, N. e Crowley, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship. *Ann. Statist.*, 2, 437-453.

[2] Chiang, C.L. (1968). *Introduction to Stochastic Processes in Biostatistics.* Wiley, New York.

[3] Cox, D.R. (1972). Regression Models and Life-Tables. *J.R. Statist. Soc.* B 34, 189-220.

[4] Cox, D.R. e Oakes, D. (1984).*Analysis of Survival Data.* Chapman and Hall, Londres.

[5] David, H.A. e Moeschberger, M.L. (1978).*The Theory of Competing Risks.* Griffin, Londres.

[6] Elandt-Johnson, R.C. e Johnson, N.L. (1980).*Survival Models and Data Analysis.* Wiley, New York.

[7] Grizzle, J.E., Starmer, C.F. e Koch, G.G. (1969). Analysis of Categorial Data by Linear Models.*Biometrics*, 25, 489-504.

[8] Hoel, D. G. (1972). A Representation of Mortality Data by Competing Risks.*Biometrics*, 28, 475-488.

[9] Johnson, W.D. e Koch, G.G. (1978). Linear Models Analysis of Competing Risks for Grouped Survival Times. *Intern. Statist. Review*, 46, 21-51.

[10] Kalbfleisch, J.D. e Prentice, R.L. (1980).*The Statistical Analysis of Failure Time Data.* Wiley, New York.

[11] Kaplan, E.L. e Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *J. Am. Statist. Assoc.*, 53, 457-481.

[12] Kimball, B.F. (1960). On the Choice of Plotting Positions on Probability Paper.*J. Am. Statist. Assoc.*, 55, 456-560.

[13] Miller, R.G. (1981).*Survival Analysis.* Wiley, New York.

[14] Moeschberger, M.L. (1974). Life Tests Under Dependent Competing Causes of Failure.*Technometrics*, 16, 39-47.

[15] Rao, C. R. (1973).*Linear Statistical Inference and Its Aplications.* 2nd edn. Wiley, New York.

[16] Wilk, M.B. e Gnanadesikan, R. (1968). Probability Plotting Methods for Analysis of Data. *Biometrika* 55, 1-17.