INFLUÊNCIA LOCAL EM MODELOS DE SOBREVIVÊNCIA COM FRAÇÃO DE CURA

Marcia Fumi Mizoi

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para a
obtenção do grau
de
Doutor em Estatística

Área de Concentração: ESTATÍSTICA
Orientador: PROF. DR. ANTONIO CARLOS PEDROSO DE LIMA

São Paulo - Novembro - 2004

Abstract

In this work we consider a diagnostic analysis in a survival model with cure fraction introduced by Chen et al. (1999), using the local influence technique (Cook 1986). The model has the structure of competing risks and has been used for modeling survival data of nonhomogeneous populations, where a subpopulation does not present the event of interest, despite a long period of follow-up . We investigate the application of the technique under different perturbation schemes, considering first the case of covariates measured without error and also extending the study to the case where one of the covariates is subject to measurement errors. In order to obtain consistent estimators in the model with error in variable, we consider the use of the corrected score method (Nakamura 1990, Gimenez & Bolfarine 1997). A simulation study is presented in order to assess the behavior of these estimators. Real and simulated data sets are considered to illustrate the technique.

Resumo

Neste trabalho consideramos um estudo de diagnóstico em um modelo de sobrevivência com fração de cura introduzido por Chen et al. (1999), utilizando a técnica de influência local (Cook 1986). O modelo abordado tem a estrutura de riscos competitivos e possibilita modelar dados de sobrevivência de populações não-homogêneas, em que parte da população não apresenta o evento de interesse, mesmo após um longo período de acompanhamento. Investigamos a aplicação da técnica sob diferentes esquemas de perturbação, considerando inicialmente o modelo com covariáveis medidas sem erro e, na seqüência, estendemos a investigação para o caso em que uma das covariáveis é sujeita a erros de medida. Para a obtenção de estimadores consistentes no modelo com erro nas variáveis, propomos a utilização do método do escore corrigido (Nakamura 1990, Gimenez & Bolfarine 1997). Um estudo de simulação é apresentado a fim de avaliar o comportamento destes estimadores. Como ilustração, apresentamos a aplicação dos resultados desenvolvidos em conjuntos de dados reais e simulados.

Agradecimentos

Sobretudo a Deus pelo amparo e pela permissão de realizar este trabalho.

Aos profs. Antonio Carlos Pedroso de Lima e Heleno Bolfarine pela inestimável orientação e também pelo apoio e atenção dispensados que possibilitaram a realização deste projeto.

Aos meus pais e irmãos.

À todos os amigos do IME pela força, amizade e pelas dicas.

Aos amigos que estão longe mas sempre presentes, pelas mensagens de carinho e pela constante torcida.

Ao Eduardo pela eterna preocupação em cuidar de mim e pela atenção e incentivo nos momentos difíceis.

À todos os professores do IME que de alguma forma me ajudaram a chegar aqui.

Aos funcionários da USP pela atenção dispensada e pelos ótimos serviços prestados.

À CAPES e ao CNPq pelo suporte financeiro concedido.

Sumário

Capítulo 1

Introdução

1.1 Modelos de Sobrevivência com Fração de Cura

Em análise de dados de sobrevivência, determinados estudos caracterizam-se por apresentar uma fração significativa de sobreviventes (isto é, unidades experimentais que não apresentam o evento de interesse), mesmo após um longo período de acompanhamento. Situações em que tais casos podem ocorrer são, por exemplo, ensaios clínicos sobre reincidência de câncer, estudos relacionados ao desenvolvimento de AIDS em pacientes HIV-positivos, testes de durabilidade de componentes eletrônicos, etc...

A Figura 1.1 a seguir, apresenta o gráfico da distribuição do tempo de sobrevivência estimada (estimador Kaplan-Meier) para um conjunto de dados reais referente a um ensaio clínico sobre reincidência de melanoma maligno (Ibrahim et al. 2001b). Para este caso, é plausível admitirmos que entre as observações censuradas, há a possibilidade de existirem pacientes que não sofrerão a reincidência da doença, ou seja, que estão curados. Observamos que o gráfico apresenta a cauda direita em um nível aproximadamente constante e estritamente maior do que zero por um período considerável. Considerar para esta situação modelos de sobrevivência usuais, que assumem que a função de sobrevivência converge para zero quando a variável tempo tende a infinito (função de sobrevivência

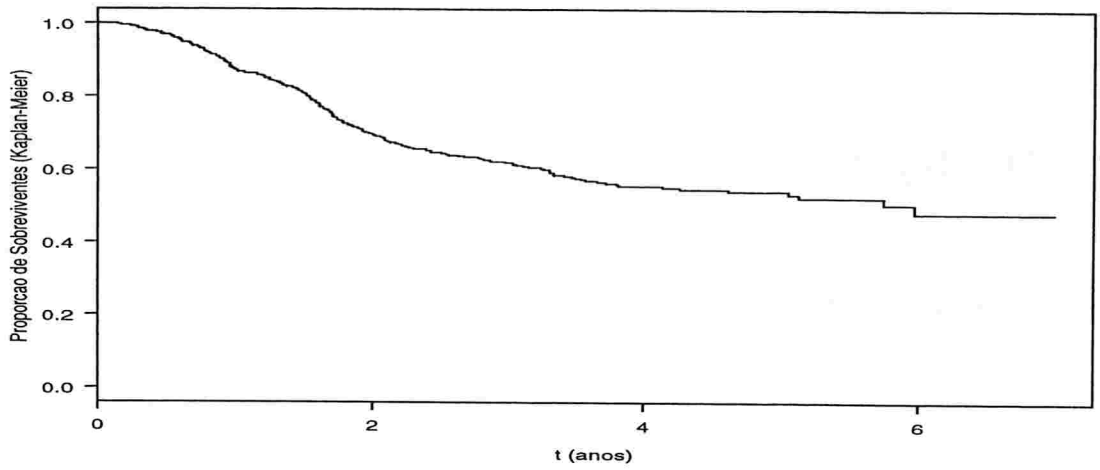

Figura 1.1: *Estimativas Kaplan-Meier para os dados de melanoma.*

própria), pode não ser adequado.

Para modelar este tipo de conjunto de dados, tem se tornado crescente o uso de modelos de sobrevivência com fração de cura.

Uma possível abordagem tem como base modelos de mistura em que considera-se uma função de sobrevivência imprópria para a população total (curados e não curados) em conjunto com uma função de sobrevivência própria para a parte da população composta por não curados. Neste contexto, Berkson & Gage (1952) introduzem um modelo de mistura de distribuições paramétricas, o qual pressupõe que a população total contém uma porcentagem π de indivíduos curados e assim, sua função de sobrevivência é escrita como

$$S(t) = \pi + (1 - \pi)S^*(t), \quad t \geq 0,$$

em que S^* é a função de sobrevivência própria associada com os indivíduos não curados. Este modelo tem sido estudado por vários autores, como por exemplo, Farewell (1982), Goldman(1984, 1991), Greenhouse & Wolfe (1984), Halpern & Brown (1987) e Sposto

et al. (1992).

Alternativamente, Yakovlev et al. (1993) propuseram uma nova classe de modelos de mistura com a estrutura de riscos competitivos. Para este modelo, a função de sobrevivência para a população total e a função de sobrevivência para a parte da população não curada são dadas, respectivamente, por

$$S(t) = \exp\left(-\theta F(t)\right) \tag{1.1}$$

e

$$S^*(t) = \frac{\exp\left(-\theta F(t)\right) - \exp(-\theta)}{1 - \exp(-\theta)}, \quad t \geq 0, \theta > 0,$$

sendo os tempos de sobrevivência definidos como o mínimo de um conjunto de N tempos com distribuição F, N uma variável com distribuição de Poisson com parâmetro θ. A fração de cura, neste caso, é dada por $S(\infty) = \exp(-\theta)$ e a função risco acumulado $H(\cdot) = -\ln(S(\cdot))$ é tal que $\lim_{t\to\infty} H(t) = \theta$.

Quando o parâmetro θ é definido como uma função de covariáveis observadas, este modelo apresenta a estrutura de riscos proporcionais, fato que não se verifica no modelo de mistura anterior.

A investigação de modelos baseados em (1.1) e a aplicação em diversos conjuntos de dados reais, pode ser vista em trabalhos como Klebanov et al. (1993), Yakovlev & Tsodikov (1996), Asselain et al. (1996), Tsodikov et al. (1995), Tsodikov et al. (1998) e Tsodikov(1998a, 1998b). Em trabalhos mais recentes, citamos Chen et al. (1999), Chen & Ibrahim (2001), Chen, Harrington & Ibrahim (2002) e Ibrahim et al.(2001a, 2001b). Nestes últimos, assume-se que, em (1.1), F segue uma distribuição Weibull ou exponencial particionada.

1.2 Objetivos

Este trabalho discute alguns procedimentos de diagnóstico aplicados ao modelo apresentado em (1.1), considerando que F segue uma distribuição Weibull. Para tanto, utilizamos técnicas que possibilitam medir o quanto pequenas alterações nos dados ou no modelo podem influir nos resultados inferenciais do problema em estudo.

Técnicas simples e bastante utilizadas para tal propósito se baseiam na retirada individual de casos, onde medidas de influência para cada observação da amostra são construídas através da comparação de estimativas calculadas para o cojunto de dados completo e para o conjunto de dados obtido eliminando-se a observação correspondente. Neste contexto, Cook (1977) sugere uma medida de influência desenvolvida inicialmente para modelos de regressão linear com erros normais.

Utilizando uma versão mais geral da estatística proposta por Cook (1977) onde, ao invés de retirar uma dada observação, atribui-se um peso para a mesma, Cook (1986) apresenta a técnica denominada *influência local*. Nesta técnica, são introduzidas perturbações em cada um dos casos, simultaneamente, e a medida de influência é construída a partir da função de log-verossimilhança. Aqui, diferentes esquemas de perturbação podem ser aplicados, de acordo com qual elemento da análise o pesquisador deseja monitorar. Esta técnica permite detectar observações conjuntamente influentes, o que constitui uma vantagem em relação ao esquema de retirada de casos, visto que, neste último, possíveis observações influentes podem não ser detectadas devido à presença de outras observações (*masking effect*). Um exemplo claro disto pode ser visto em Weissfeld & Schneider (1990) em um estudo de influência em modelos de sobrevivência com distribuição Weibull. A presença de observações influentes na amostra pode levar a resultados inferenciais completamente diferentes, sendo importante ao pesquisador conhecer e analisar estes casos

para decidir pelas suas retiradas, ou não, do estudo.

Vários trabalhos apresentam o desenvolvimento da aplicação da técnica de influência local em classes de modelos específicos. Em particular, para dados de sobrevivência, destacam-se os trabalhos de Escobar & Meeker (1992) em modelos de vida acelerada paramétricos, Pettitt & Bin Daud (1989) e Weissfeld (1990) em modelos de riscos proporcionais de Cox e Ortega et al. (2003) em modelos de regressão log-gama generalizados.

O objetivo do presente projeto é pesquisar a utilização da metodologia de influência local na classe de modelos com fração de cura (1.1), assumindo que F segue uma distribuição Weibull.

Consideramos o modelo sob duas perspectivas. Primeiramente, investigamos o modelo com covariáveis medidas sem erro e, em seguida, estendemos o estudo para o caso em que uma das covariáveis do modelo é sujeita a erros de medida. Para a estimação no modelo com erro nas variáveis, propomos a utilização do método do escore corrigido (Nakamura 1990, Gimenez & Bolfarine 1997) o qual possibilita a obtenção de estimadores consistentes. Realizamos, também, várias simulações a fim de observar o comportamento destes estimadores.

Este trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada a formulação do modelo e o processo de estimação dos parâmetros quando as covariáveis são medidas sem erro. A extensão para o caso com erro nas variáveis e um estudo de simulação são considerados no Capítulo 3. Na seqüência, no Capítulo 4, a metodologia de influência local é descrita e aplicada aos modelos discutidos nos capítulos anteriores. Diferentes esquemas de perturbação são considerados. Os resultados da aplicação desta teoria são investigados no Capítulo 5, mediante dados simulados e um conjunto de dados de sobrevivência reais. A conclusão do trabalho é apresentada no Capítulo 6, com uma

discussão dos resultados obtidos e a proposição de possíveis pesquisas futuras.

Capítulo 2

Modelo com Fração de Cura

2.1 Introdução

Em análise de dados envolvendo tempos de sobrevivência (ou tempos de falha), normalmente assume-se que o evento de interesse ocorrerá, em algum instante, para todo indivíduo da população alvo, contanto que o tempo de acompanhamento seja suficientemente grande. Ou seja, assumindo-se que o tempo de sobrevivência pode ser representado por uma variável aleatória não negativa T, caracterizada por uma função de sobrevivência $S(t) = P(T > t)$, esta função, em geral é tal que

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

Conseqüentemente, a função de risco acumulado $H(t) = -\ln(S(t))$ é tal que

$$H(\infty) = \lim_{t \to \infty} H(t) = \infty,$$

ou seja, esta função não é limitada.

Contudo, em determinados experimentos, para uma certa parcela da população em estudo o evento especificado pode nunca ocorrer. Denominamos esta parcela de curada ou imune. Por exemplo, em estudos médicos muitas vezes se tem interesse no tempo de recorrência de determinada doença para pacientes que recebem algum tipo de tratamento.

Parte destes pacientes pode se manter livre da doença após o tratamento e ser considerada curada após um período suficientemente longo de acompanhamento (esta suficiência do tempo de acompanhamento pode, muitas vezes, ser aproximadamente estabelecida pela experiência do pesquisador ou conhecimento prévio do comportamento e evolução de determinadas doenças). Em certos casos de tratamento de câncer, por exemplo, considera-se que a não recorrência da doença em um intervalo de 5 a 10 anos é indicativo da cura do paciente. Um dos interesses do pesquisador, nestes casos, pode ser a estimação da proporção de indivíduos curados, o que pode auxiliar na escolha de tratamentos a serem indicados.

Formalmente, a existência de uma fração de sobreviventes (a qual denominaremos fração de cura) é caracterizada pelo fato de a função de sobrevivência não convergir para zero quando o tempo aumenta (função de sobrevivência imprópria). A averiguação inicial da existência desta fração em um conjunto de dados pode ser feita através do gráfico da função de sobrevivência estimada de Kaplan-Meier que, neste caso, deve apresentar a cauda direita em um nível constante acima de zero por um período considerado suficientemente grande. Uma discussão mais aprofundada sobre a descrição assim como testes estatísticos sobre a presença de uma fração de cura e da suficiência do tempo de acompanhamento, pode ser vista em Maller & Zhou (1996).

A modelagem de conjuntos de dados onde há a possibilidade de cura (ou a presença de imunes) pode ser realizada considerando uma função de sobrevivência imprópria S_p para a população total (curados e não curados) e uma função de sobrevivência própria S^* para a parte da população de não curados, ou seja, consideramos $S_p(\infty) = \lim_{t\to\infty} S_p(t) > 0$

e $S^*(\infty) = \lim_{t\to\infty} S^*(t) = 0$.

Neste contexto, consideraremos neste trabalho a classe de modelos com fração de cura proposta em Yakovlev et al. (1993), em que tal modelagem é feita impondo um limite superior à função de risco acumulado, fazendo

$$H_p(\infty) = \lim_{t\to\infty} H_p(t) = \theta, \quad \theta > 0,$$

e definindo $H_p(t) = \theta F(t)$, com F função de distribuição de uma variável aleatória não-negativa. Observe que aqui, $S_p(\infty) = \lim_{t\to\infty} \exp(-H_p(t)) = \exp(-\theta) > 0$. Tsodikov (1998b) observa que tomando θ como função de covariáveis observadas, este modelo assume a estrutura de riscos proporcionais (discutiremos esta propriedade mais adiante) e apresenta um estudo sem especificar a distribuição F.

Yakovlev et al. (1993) introduzem uma estrutura de riscos competitivos no modelo considerando as seguintes suposições:

- N: variável aleatória com distribuição de Poisson de média θ;
- $R_1, R_2, \ldots, R_N$: variáveis aleatórias i.i.d., independentes de N, com função de distribuição $F(\cdot)$ e função de sobrevivência $S(\cdot) = 1 - F(\cdot)$;
- T: tempo de ocorrência do evento de interesse, definido como $T = \min\{R_0, R_1, \ldots, R_N\}$, em que $P(R_0 = \infty) = 1$.

Conseqüentemente,

$$\begin{aligned}
S_p(t) &= P\left(\min\left\{R_0, R_1, \ldots, R_N\right\} > t\right) \\
&= \sum_{k=0}^{\infty} P\left(\min\left\{R_0, R_1, \ldots, R_k\right\} > t\right) P(N = k) \\
&= P(N = 0) + \sum_{k=1}^{\infty} P\left(R_0 > t, R_1 > t, \ldots, R_k > t\right) P(N = k) \\
&= \exp(-\theta) + \sum_{k=1}^{\infty} P\left(R_1 > t\right) \ldots P\left(R_k > t\right) P(N = k) \\
&= \exp(-\theta) + \sum_{k=1}^{\infty} \left[S(t)\right]^k \frac{\theta^k}{k!} \exp(-\theta) \\
&= \exp(-\theta) \exp(\theta S(t)),
\end{aligned}$$

e, portanto,

$$S_p(t) = \exp\left(-\theta F(t)\right), \quad t \geq 0. \tag{2.1}$$

Observe que

$$S_p(\infty) \equiv \lim_{t \to \infty} S_p(t) = \exp(-\theta)$$

e assim,

$$S_p(\infty) = P(N = 0) = \exp(-\theta)$$

corresponde à fração de cura induzida pelo modelo. Note ainda que a fração de cura $\exp(-\theta)$ tende a zero quando θ tende a infinito e tende a 1 quando θ tende a zero, o que é intuitivamente esperado, uma vez que se θ cresce, a média de N aumenta, diminuindo a probabilidade de cura. O mesmo argumento vale para θ decrescente.

Supondo que as variáveis $\{R_i, i = 1, \ldots, N\}$ são absolutamente contínuas e denotando sua função densidade de probabilidade por $f(t)$, temos

$$f_p(t) = -\frac{dS_p(t)}{dt} = \theta f(t) \exp\left(-\theta F(t)\right), \quad t \geq 0.$$

Como conseqüência de (2.1), f_p não é uma função densidade de probabilidade própria.

A função de taxa de falha associada a S_p é dada por

$$h_p(t) = \frac{f_p(t)}{S_p(t)} = \frac{\theta f(t) \exp\left(-\theta F(t)\right)}{\exp\left(-\theta F(t)\right)} = \theta f(t),$$

Logo, supondo que $\theta = \theta(\mathbf{x})$, $\mathbf{x}$ um vetor com p covariáveis, tem-se que $h_p(t) = \theta(\mathbf{x}) f(t)$, caracterizando um modelo de riscos proporcionais.

Para a parte da população não-curada obtemos a função de sobrevivência própria

$$S^*(t) = P\left(T > t | N \geq 1\right) = \frac{\exp\left(-\theta F(t)\right) - \exp(-\theta)}{1 - \exp(-\theta)}. \tag{2.2}$$

Para tempos absolutamente contínuos, a função densidade de probabilidade e a função de taxa de falha associadas à população não-curada são então dadas respectivamente por

$$f^*(t) = -\frac{dS^*(t)}{dt} = \frac{\exp\left(-\theta F(t)\right)}{1 - \exp(-\theta)} \theta f(t)$$

e

$$h^*(t) = \frac{f^*(t)}{S^*(t)} = \frac{\exp\left(-\theta F(t)\right)}{\exp\left(-\theta F(t)\right) - \exp(-\theta)} h_p(t).$$

Todos os resultados apresentados até o momento assumem no máximo, tempos de falha específicos (isto é, para as variáveis aleatórias $R_i, i = 1, \ldots, N$) absolutamente contínuos e população homogênea. A fim de considerar heterogeneidades populacionais relativamente à fração de curados e de não-curados, é de interesse incluir covariáveis no parâmetro θ. Chen et al. (1999) desenvolvem uma versão deste modelo, introduzindo covariáveis através da relação $\theta \equiv \theta(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, com $\boldsymbol{\beta}$ um vetor p-dimensional com os coeficientes de

regressão associados a $\mathbf{x}$. Desta forma, a fração de cura fica relacionada ao vetor de covariáveis segundo a expressão

$$S_p(\infty) = \exp(-\exp(\mathbf{x}'\boldsymbol{\beta})).$$

Note também que as covariáveis contribuem para a função de sobrevivência (2.2) e, conseqüentemente, para a taxa de falha da população de não-curados.

Baseados nos trabalhos de Chen et al. (1999) e Chen & Ibrahim (2001), desenvolvemos as próximas seções. Na Seção 2.2 apresentamos a construção da função de verossimilhança do modelo com covariáveis. A seguir, na Seção 2.3, o processo de estimação é desenvolvido segundo o algoritmo EM (Dempster et al. 1977) e na Seção 2.4 apresentamos a formulação do modelo apresentado considerando uma distribuição Weibull para R_i's.

2.2 Função de Verossimilhança

Suponhamos agora n indivíduos e para cada indivíduo i, $i = 1, \ldots, n$, as seguintes variáveis:

- N_i: variáveis i.i.d. não observadas com $N_i \sim Poisson(\theta)$;
- $R_{i1}, R_{i2}, \ldots, R_{i,N_i}$: variáveis i.i.d., não observáveis, com função de distribuição $F(\cdot|\boldsymbol{\lambda})$ e função de sobrevivência $S(\cdot|\boldsymbol{\lambda}) = 1 - F(\cdot|\boldsymbol{\lambda})$, com $\boldsymbol{\lambda}$ vetor de parâmetros;
- y_i: tempo observado, dado por $y_i = \min\{T_i, C_i\}$, com $T_i = \min\{R_{i0}, R_{i1}, \ldots, R_{i,N_i}\}$ e C_i tempo de censura do indivíduo;
- ν_i: indicador de censura, com $\nu_i = \begin{cases} 1, & \text{se } y_i \text{ é tempo de falha;} \\ 0, & \text{se } y_i \text{ é tempo de censura.} \end{cases}$

- $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$: vetor de covariáveis associadas, introduzidas no modelo através do parâmetro θ segundo a relação $\theta_i \equiv \theta(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, sendo $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ o vetor correspondente dos coeficientes de regressão.

Definindo os vetores n-dimensionais

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)',$$

$$\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)',$$

$$\mathbf{N} = (N_1, N_2, \ldots, N_n)$$

e a matriz de covariáveis com dimensão $n \times p$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x_1'} \\ \mathbf{x_2'} \\ \vdots \\ \mathbf{x_n'} \end{pmatrix},$$

denotamos o conjunto dos dados completos (com as variáveis latentes) por

$$\mathbf{D_c} = (n, \mathbf{y}, \boldsymbol{\nu}, \mathbf{N}, \mathbf{X}).$$

Seja então $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{\lambda}')'$ o vetor de parâmetros. Mostra-se que a função de verossimilhança dos dados completos para $\boldsymbol{\phi}$ é dada por (vide Apêndice A)

$$L(\boldsymbol{\phi}; \mathbf{D_c}) = \left\{ \prod_{i=1}^{n} S(y_i|\boldsymbol{\lambda})^{N_i - \nu_i} \left[N_i f(y_i|\boldsymbol{\lambda}) \right]^{\nu_i} \right\} \exp\left\{ \sum_{i=1}^{n} \left[N_i \mathbf{x}_i'\boldsymbol{\beta} - \ln(N_i!) - \exp(\mathbf{x_i'}\boldsymbol{\beta}) \right] \right\}. \tag{2.3}$$

Note que esta verossimilhança não é observada, pois depende de $\mathbf{N}$ que é uma variável latente.

Aplicando o logaritmo, vem

$$l(\boldsymbol{\phi}; \mathbf{D_c}) = \sum_{i=1}^{n} \{(N_i - \nu_i) \ln S(y_i|\boldsymbol{\lambda}) + \nu_i \ln N_i + \nu_i \ln f(y_i|\boldsymbol{\lambda})\} + \sum_{i=1}^{n} \left[N_i \mathbf{x}_i' \boldsymbol{\beta} - \ln(N_i!) - \exp(\mathbf{x}_i' \boldsymbol{\beta})\right]. \quad (2.4)$$

Agora, seja $\mathbf{D}$ o conjunto dos dados observados com $\mathbf{D} = (n, \mathbf{y}, \boldsymbol{\nu}, \mathbf{X})$. Uma vez que (2.4) inclui as variáveis latentes $\mathbf{N}$, trabalhamos com a log-verossimilhança marginal (obtida fazendo-se o somatório nas variáveis não observadas $\mathbf{N}$), obtendo

$$l(\boldsymbol{\phi}; \mathbf{D}) = \sum_{i=1}^{n} \left\{\nu_i \mathbf{x}_i' \boldsymbol{\beta} + \nu_i \ln f(y_i|\boldsymbol{\lambda}) - \exp(\mathbf{x}_i' \boldsymbol{\beta}) \left[1 - S(y_i|\boldsymbol{\lambda})\right]\right\}. \quad (2.5)$$

2.3 Estimação dos Parâmetros do Modelo

A maximização da log-verossimilhança dos dados observados $l(\boldsymbol{\phi}|\mathbf{D})$ pode ser obtida utilizando-se o algoritmo EM (Dempster et al. 1977). O passo E do algoritmo consiste em calcular a esperança condicional da log-verossimilhança dos dados completos, ou seja, devemos calcular

$$E\left[l(\boldsymbol{\phi}; \mathbf{D_c}) \,|\, \mathbf{D}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right] \quad (2.6)$$

em que $l(\boldsymbol{\phi}; \mathbf{D_c})$ é dada em (2.4), e $\boldsymbol{\beta}^{(k)}$ e $\boldsymbol{\lambda}^{(k)}$ são as estimativas dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$ na k-ésima iteração do algoritmo.

Observando que os termos $\nu_i \ln N_i$ e $\ln(N_i!)$ em (2.4) não envolvem os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$, temos que para obter a esperança em (2.6), basta calcularmos a esperança condicional

$$E\left[N_i \,|\, \mathbf{D}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right].$$

Calculando a distribuição condicional de N_i dado $\mathbf{D}$, (vide Apêndice B), obtemos que é a mesma distribuição de $V_i+\nu_i$, com V_i variável Poisson tal que $E(V_i) = S(y_i|\boldsymbol{\lambda})\exp(\mathbf{x}_i'\boldsymbol{\beta})$.

Segue então que

$$E\left[N_i \,|\, \mathbf{D}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right] = S(y_i|\boldsymbol{\lambda}^{(k)})\exp(\mathbf{x}_i'\boldsymbol{\beta}^{(k)}) + \nu_i.$$

Assim, denotando

$$N_i^{(k+1)} = E\left[N_i \,|\, \mathbf{D}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right],$$

temos que na $(k+1)$-ésima iteração o algoritmo é formado pelos seguintes passos:

- **Passo E** - Cálculo da esperança condicional da log-verossimilhança dos dados completos:

$$\begin{aligned} E\left[l(\boldsymbol{\phi}; \mathbf{D}_c) \,|\, \mathbf{D}; \boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right] &= \sum_{i=1}^{n}\left\{N_i^{(k+1)}\mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right\} \\ &\quad + \sum_{i=1}^{n}\left\{N_i^{(k+1)}\ln S(y_i|\boldsymbol{\lambda}) + \nu_i \ln h(y_i|\boldsymbol{\lambda})\right\}, \end{aligned}$$

  em que $h(y_i|\boldsymbol{\lambda}) = \frac{f(y_i|\boldsymbol{\lambda})}{S(y_i|\boldsymbol{\lambda})}$ e $N_i^{(k+1)} = S(y_i|\boldsymbol{\lambda}^{(k)})\exp(\mathbf{x}_i'\boldsymbol{\beta}^{(k)}) + \nu_i$.

- **Passo M** - Maximização das expressões:

$$Q^{(1)}\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right) \equiv \sum_{i=1}^{n}\left\{N_i^{(k+1)}\mathbf{x}_i'\boldsymbol{\beta} - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right\}$$

  e

$$Q^{(2)}\left(\boldsymbol{\lambda}|\boldsymbol{\beta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right) \equiv \sum_{i=1}^{n}\left\{N_i^{(k+1)}\ln S(y_i|\boldsymbol{\lambda}) + \nu_i \ln h(y_i|\boldsymbol{\lambda})\right\},$$

sendo que a maximização de $Q^{(1)}$ gera o vetor $\boldsymbol{\beta}^{(k+1)}$ e a maximização de $Q^{(2)}$ gera o vetor $\boldsymbol{\lambda}^{(k+1)}$.

As estimativas de máxima verossimilhança $\hat{\boldsymbol{\phi}} = \left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\lambda}}'\right)'$ são então obtidas com a convergência do algoritmo, segundo o critério de parada estabelecido. Consideramos aqui como critério a condição

$$\left|\boldsymbol{\phi}^{(k+1)} - \boldsymbol{\phi}^{(k)}\right| < \varepsilon,$$

com $\boldsymbol{\phi}^{(k)} = \left(\boldsymbol{\beta}^{(k)'}, \boldsymbol{\lambda}^{(k)'}\right)'$.

2.4 Modelo Paramétrico com Distribuição Weibull

Suponhamos agora que F corresponda à função de distribuição de uma variável Weibull, com vetor de parâmetros $\boldsymbol{\lambda} = (\rho, \gamma)$, ou seja, suponhamos que $R_k \sim Weibull(\rho, \gamma)$, $k = 1, \ldots, N_i$, $i = 1, \ldots, n$, com função densidade $f(t) = \rho t^{\rho-1} \exp(\gamma - t^{\rho} e^{\gamma})$ e função de sobrevivência $S(t) = \exp(-t^{\rho} e^{\gamma})$. Neste caso, a log-verossimilhança dada por (2.4) é expressa como

$$\begin{aligned} l(\boldsymbol{\phi}; \mathbf{D_c}) &= \sum_{i=1}^{n} \left\{-N_i y_i^{\rho} e^{\gamma} + \nu_i \ln\left(N_i \rho y_i^{\rho-1} e^{\gamma}\right)\right\} \\ &\quad + \sum_{i=1}^{n} \left\{N_i \mathbf{x}_i' \boldsymbol{\beta} - \ln(N_i!) - \exp(\mathbf{x}_i' \boldsymbol{\beta})\right\}, \end{aligned} \tag{2.7}$$

e a log-verossimilhança marginal em (2.5) por

$$
\begin{aligned}
l(\boldsymbol{\phi};\mathbf{D}) &= \sum_{i=1}^{n}\left\{\nu_i \mathbf{x}_i'\boldsymbol{\beta} + \nu_i \ln\left[\rho y_i^{\rho-1}\exp\left(\gamma - y_i^{\rho}e^{\gamma}\right)\right]\right\} \\
&\quad - \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i'\boldsymbol{\beta})\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\} \\
&= \sum_{i=1}^{n}\left\{\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta} + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right]\right\} \\
&\quad - \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i'\boldsymbol{\beta})\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\}. \qquad (2.8)
\end{aligned}
$$

As estimativas de máxima verossimilhança $\hat{\boldsymbol{\phi}} = \left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\lambda}}'\right)'$ dos parâmetros da log-verossimilhança acima podem ser calculadas através dos passos do algoritmo EM descrito na seção anterior.

Podemos também calcular estimativas para a variância de $\hat{\boldsymbol{\phi}}$ e construir testes de hipóteses para os parâmetros, utilizando o fato que $\hat{\boldsymbol{\phi}}$ tem distribuição assintótica normal multivariada com média $\boldsymbol{\phi}$ e matriz de covariâncias $\mathbf{I}^{-1}(\boldsymbol{\phi})$ com

$$\mathbf{I}(\boldsymbol{\phi}) = -E\left[\ddot{\mathbf{L}}(\boldsymbol{\phi})\right],$$

tal que

$$\ddot{\mathbf{L}}(\boldsymbol{\phi}) = \left\{\frac{\partial^2 l(\boldsymbol{\phi};\mathbf{D})}{\partial\phi_i\partial\phi_j}\right\}_{i,j=1,\ldots,p+2}$$

(ver Yakovlev & Tsodikov (1996)).

Visto que o cálculo de $\mathbf{I}(\boldsymbol{\phi})$ (denominada matriz de informação de Fisher) não é possível devido à presença de censuras, podemos utilizar alternativamente a matriz $-\ddot{\mathbf{L}}(\boldsymbol{\phi})$ avaliada em $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$ (denominada matriz de informação observada), a qual é uma estimativa de $\mathbf{I}(\boldsymbol{\phi})$.

Temos que para o modelo (2.8), a matriz $\ddot{\mathbf{L}}(\boldsymbol{\phi})$ toma a seguinte forma

$$\ddot{\mathbf{L}}(\boldsymbol{\phi}) = \begin{pmatrix} \ddot{\mathbf{L}}^{\boldsymbol{\beta}\boldsymbol{\beta}} & \ddot{\mathbf{L}}^{\boldsymbol{\beta}\rho} & \ddot{\mathbf{L}}^{\boldsymbol{\beta}\gamma} \\ & \ddot{L}^{\rho\rho} & \ddot{L}^{\rho\gamma} \\ & & \ddot{L}^{\gamma\gamma} \end{pmatrix}, \tag{2.9}$$

sendo as submatrizes com elementos

$$\ddot{L}^{\boldsymbol{\beta}\boldsymbol{\beta}}_{k,l} = -\sum_{i=1}^{n} \left\{ x_{ik}x_{il} \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right] \right\}, \; k,l = 1,\ldots,p,$$

$$\ddot{L}^{\boldsymbol{\beta}\rho}_{k} = -\sum_{i=1}^{n} \left\{ x_{ik}y_i^{\rho}e^{\gamma}\ln(y_i) \exp\left(\mathbf{x}_i'\boldsymbol{\beta} - y_i^{\rho}e^{\gamma}\right) \right\}, \; k = 1,\ldots,p,$$

$$\ddot{L}^{\boldsymbol{\beta}\gamma}_{k} = -\sum_{i=1}^{n} \left\{ x_{ik}y_i^{\rho}e^{\gamma} \exp\left(\mathbf{x}_i'\boldsymbol{\beta} - y_i^{\rho}e^{\gamma}\right) \right\}, \; k = 1,\ldots,p,$$

$$\ddot{L}^{\rho\rho} = -\sum_{i=1}^{n} \left\{ \nu_i \left[\frac{1}{\rho^2} + y_i^{\rho}e^{\gamma}(\ln(y_i))^2\right] + y_i^{\rho}e^{\gamma}(\ln(y_i))^2 \exp\left(\mathbf{x}_i'\boldsymbol{\beta} - y_i^{\rho}e^{\gamma}\right)\left[1 - y_i^{\rho}e^{\gamma}\right] \right\},$$

$$\ddot{L}^{\rho\gamma} = -\sum_{i=1}^{n} \left\{ \nu_i y_i^{\rho}e^{\gamma}\ln(y_i) + y_i^{\rho}e^{\gamma}\ln(y_i) \exp\left(\mathbf{x}_i'\boldsymbol{\beta} - y_i^{\rho}e^{\gamma}\right)\left[1 - y_i^{\rho}e^{\gamma}\right] \right\}$$

e

$$\ddot{L}^{\gamma\gamma} = -\sum_{i=1}^{n} \left\{ \nu_i y_i^{\rho}e^{\gamma} + y_i^{\rho}e^{\gamma} \exp\left(\mathbf{x}_i'\boldsymbol{\beta} - y_i^{\rho}e^{\gamma}\right)\left[1 - y_i^{\rho}e^{\gamma}\right] \right\}.$$

Testes estatísticos podem ser realizados para o subvetor de parâmetros $\boldsymbol{\beta}$. Suponhamos que há o interesse em testar a hipótese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Denotemos por $\hat{\boldsymbol{\lambda}}(\boldsymbol{\beta}_0)$ a estimativa obtida maximizando-se a log-verossimilhança $l(\boldsymbol{\phi}; \mathbf{D})$ com relação a $\boldsymbol{\lambda}$ para $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ fixado, e denotemos por

$$\ddot{\mathbf{L}}^{-1}(\phi) = \begin{pmatrix} \ddot{\mathbf{L}}^{\beta\beta} & \ddot{\mathbf{L}}^{\beta\rho} & \ddot{\mathbf{L}}^{\beta\gamma} \\ & \ddot{L}^{\rho\rho} & \ddot{L}^{\rho\gamma} \\ & & \ddot{L}^{\gamma\gamma} \end{pmatrix}^{-1} = \begin{pmatrix} \ddot{\mathbf{L}}_{\beta\beta} & \ddot{\mathbf{L}}_{\beta\rho} & \ddot{\mathbf{L}}_{\beta\gamma} \\ & \ddot{L}_{\rho\rho} & \ddot{L}_{\rho\gamma} \\ & & \ddot{L}_{\gamma\gamma} \end{pmatrix}.$$

Então, as estatísticas razão de verossimilhança e de Wald para testar H_0 são dadas, respectivamente, por

$$LR = -2\left[l\left(\left(\beta_0, \hat{\boldsymbol{\lambda}}(\beta_0)\right); \mathbf{D}\right) - l\left(\hat{\phi}; \mathbf{D}\right)\right]$$

e

$$W = \left(\hat{\beta} - \beta_0\right)' \ddot{\mathbf{L}}_{\beta\beta}^{-1}\left(\hat{\beta} - \beta_0\right),$$

com $\ddot{\mathbf{L}}_{\beta\beta}$ avaliada em $\phi = \hat{\phi}$. Estas estatísticas têm distribuição assintótica $\chi^2_{(p)}$, sob a hipótese H_0.

Capítulo 3

Modelo com Fração de Cura com Erro nas Variáveis

Neste capítulo estendemos o estudo do modelo anterior considerando que uma das covariáveis do modelo é sujeita a erro de medida. Estimadores consistentes para o modelo são obtidos segundo o método do escore corrigido (Nakamura (1990); Gimenez & Bolfarine (1997)) e um estudo de simulação é apresentado.

3.1 Introdução

Suponhamos agora que no conjunto de covariáveis no modelo com fração de cura, uma das covariáveis não seja precisamente mensurada, ou seja, suponhamos que para cada indivíduo i, $i = 1, \ldots, n$, temos

z_i : covariável medida com erro, possivelmente aleatória;

$\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$: vetor de p covariáveis medidas sem erro.

Dados de tempo de sobrevivência (ou falha) que incluem covariáveis medidas com erro podem ocorrer comumente na prática como, por exemplo, no caso de estudo de pacientes com determinada doença em que alguns dados clínicos (como o valor da pressão sangüínea)

são normalmente sujeitos a erros de medição.

Os chamados modelos com erro nas variáveis possibilitam levar em consideração tais erros de medida em uma análise estatística (ver Fuller (1987) e Carroll et al. (1995) para maiores detalhes). Consideramos aqui um modelo com erro nas variáveis com estrutura aditiva.

Assim, denotando por w_i os valores observados da covariável z_i, $i = 1, \ldots, n$, e, supondo um modelo com erro nas variáveis aditivo, escrevemos

$$w_i = z_i + u_i,$$

com u_i variável aleatória representando o erro de medida. Assumimos aqui que u_i, $i = 1, \ldots, n$, são independentes e identicamente distribuídas, independentes de z_i, y_i e ν_i, $i = 1, \ldots, n$, tal que $u_i \sim N(0, \sigma_u^2)$. No caso em que z_i são consideradas constantes desconhecidas, temos um modelo funcional e, no caso em que supomos z_i como variáveis aleatórias independentes e identicamente distribuídas, temos um modelo estrutural. Na seqüência, consideramos o modelo funcional, que tem maior aplicabilidade na prática.

Considerando as verdadeiras covariáveis $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$, a log-verossimilhança do modelo com fração de cura dada por (2.5), é agora dada por

$$l(\boldsymbol{\phi}; \mathrm{D}) = \sum_{i=1}^{n} \left\{ \nu_i \left(\mathbf{x}_i' \boldsymbol{\beta}_x + z_i \beta_z \right) + \nu_i \ln f(y_i | \boldsymbol{\lambda}) - \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_x + z_i \beta_z \right) [1 - S(y_i | \boldsymbol{\lambda})] \right\}, \tag{3.1}$$

sendo $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{\lambda}')'$ com $\boldsymbol{\beta} = (\boldsymbol{\beta}_x', \beta_z)'$ e $\mathrm{D} = (n, \mathbf{y}, \boldsymbol{\nu}, \mathbf{z}, \mathrm{X})$.

Quando substituímos as verdadeiras covariáveis $\mathbf{z}$ pelas covariáveis observadas $\mathbf{w} = (w_1, w_2, \ldots, w_n)'$, obtemos a log-verossimilhança naive

$$l(\boldsymbol{\phi}; \bar{\mathrm{D}}) = \sum_{i=1}^{n} \left\{ \nu_i \left(\mathbf{x}_i' \boldsymbol{\beta}_x + w_i \beta_z \right) + \nu_i \ln f(y_i | \boldsymbol{\lambda}) - \exp\left(\mathbf{x}_i' \boldsymbol{\beta}_x + w_i \beta_z \right) [1 - S(y_i | \boldsymbol{\lambda})] \right\},$$

com $\bar{\mathbf{D}} = (n, \mathbf{y}, \boldsymbol{\nu}, \mathbf{w}, \mathbf{X})$.

As estimativas obtidas quando simplesmente substituímos as verdadeiras covariáveis $\mathbf{z}$ pelas covariáveis observadas com erro $\mathbf{w}$ (chamadas estimativas "naive"), podem ser assintoticamente viciadas (Stefanski 1985). Em geral, ocorre que o estimador "naive" é inconsistente, ou seja, é atenuado pela presença do erro de medição.

3.2 Estimação dos Parâmetros - Método do Escore Corrigido

O método do escore corrigido é um método para estimação em modelos com erro nas variáveis que possibilita a obtenção de estimadores consistentes e assintoticamente normais. O método supõe o conhecimento da variância do erro de medida (ou uma estimativa desta) e pode ser utilizado tanto em modelos funcionais como estruturais. No que segue, admitimos o modelo funcional com σ_u^2 conhecido.

Introduzido por Nakamura (1990), este método propõe calcular inicialmente uma log-verossimilhança corrigida l^*, construída a partir de uma "correção" na função de log-verossimilhança não observada l (função das covariáveis não observadas).

Na notação utilizada em (3.1), tal log-verossimilhança corrigida l^* é obtida de modo a satisfazer a equação

$$E\left[l^*(\boldsymbol{\phi}; \bar{\mathbf{D}}) \,|\, \mathbf{D}\right] = l(\boldsymbol{\phi}; \mathbf{D}).$$

Quando l^* é diferenciável, definimos então a função escore corrigido $\mathbf{U}^*$ como sendo a derivada da função de log-verossimilhança corrigida, ou seja,

$$\mathbf{U}^*(\boldsymbol{\phi}; \bar{\mathbf{D}}) = \frac{\partial l^*(\boldsymbol{\phi}; \bar{\mathbf{D}})}{\partial \boldsymbol{\phi}}.$$

As estimativas do método do escore corrigido, que denotamos aqui por $\hat{\boldsymbol{\phi}}_{ec}$, são então definidas como a solução da equação $\mathbf{U}^*(\boldsymbol{\phi}; \bar{\mathbf{D}}) = \mathbf{0}$, com $-\frac{\partial \mathbf{U}^*(\boldsymbol{\phi}; \bar{\mathbf{D}})}{\partial \boldsymbol{\phi}}$ matriz positiva definida.

Suponhamos agora que a função escore corrigido satisfaça

$$E\left[\mathbf{U}^*(\boldsymbol{\phi}; \bar{\mathbf{D}})|\mathbf{D}\right] = \mathbf{U}(\boldsymbol{\phi}; \mathbf{D}), \tag{3.2}$$

com $\mathbf{U}(\boldsymbol{\phi}; \mathbf{D}) = \frac{\partial l(\boldsymbol{\phi}; \mathbf{D})}{\partial \boldsymbol{\phi}}$.

Para o caso em que a relação acima é válida, e sob certas condições de regularidade, propriedades assintóticas das estimativas podem ser obtidas usando resultados apresentados em Gimenez & Bolfarine (1997). Suponhamos então válida a relação (3.2) e denotemos

$$\mathbf{U}^*(\boldsymbol{\phi}; \bar{\mathbf{D}}) = \sum_{i=1}^{n} \mathbf{U}_i^*(\boldsymbol{\phi}; y_i, \nu_i, w_i, \mathbf{X}_i) = \sum_{i=1}^{n} \frac{\partial l_i^*(\boldsymbol{\phi}; y_i, \nu_i, w_i, \mathbf{X}_i)}{\partial \boldsymbol{\phi}}$$

e

$$\ddot{\mathbf{L}}^*(\boldsymbol{\phi}; \bar{\mathbf{D}}) = \sum_{i=1}^{n} \ddot{\mathbf{L}}_i^*(\boldsymbol{\phi}; y_i, \nu_i, w_i, \mathbf{X}_i) = \sum_{i=1}^{n} \frac{\partial \mathbf{U}_i^*(\boldsymbol{\phi}; y_i, \nu_i, w_i, \mathbf{X}_i)}{\partial \boldsymbol{\phi}}$$

(para simplificar a notação, no que segue denotamos as funções acima por $\mathbf{U}^*(\boldsymbol{\phi})$, $\mathbf{U}_i^*(\boldsymbol{\phi})$, $\ddot{\mathbf{L}}^*(\boldsymbol{\phi})$ e $\ddot{\mathbf{L}}_i^*(\boldsymbol{\phi})$). Sejam então

$$\overline{\ddot{\mathbf{L}}}_n^*(\boldsymbol{\phi}) = \frac{1}{n} \sum_{i=1}^{n} \ddot{\mathbf{L}}_i^*(\boldsymbol{\phi}),$$

$$\overline{\boldsymbol{\Lambda}}_n(\boldsymbol{\phi}) = \frac{1}{n} \sum_{i=1}^{n} E\left[-\ddot{\mathbf{L}}_i^*(\boldsymbol{\phi})\right],$$

e

$$\overline{\boldsymbol{\Gamma}}_n(\boldsymbol{\phi}) = \frac{1}{n} \sum_{i=1}^{n} E\left[\mathbf{U}_i^*(\boldsymbol{\phi}) \mathbf{U}_i^*(\boldsymbol{\phi})'\right].$$

Denotando por ϕ_0 o verdadeiro valor do parâmetro segue que, sob certas condições de regularidade, $\hat{\phi}_{ec}$ segue uma distribuição assintótica normal com média ϕ_0 e matriz de covariâncias $n^{-1}\Omega_n$, com

$$\Omega_n = \left\{\overline{\Lambda}_n(\phi_0)\right\}^{-1} \overline{\Gamma}_n(\phi_0) \left\{\overline{\Lambda}_n(\phi_0)'\right\}^{-1}. \tag{3.3}$$

Uma estimativa consistente para (3.3) é dada por

$$\hat{\Omega}_n = \left\{-\overline{\ddot{L}}_n^*(\hat{\phi}_{ec})\right\}^{-1} \overline{G}_n(\hat{\phi}_{ec}) \left\{-\overline{\ddot{L}}_n^*(\hat{\phi}_{ec})'\right\}^{-1}, \tag{3.4}$$

com

$$\overline{G}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} U_i^*(\phi) U_i^*(\phi)'.$$

Testes estatísticos baseados nas propriedades assintóticas dos estimadores do escore corrigido também podem ser construídos para testar hipóteses de interesse usando resultados apresentados em Gimenez et al. (2000).

Suponhamos que estamos interessados em testar o subvetor de parâmetros β, através da hipótese $H_0 : \beta = \beta_0$. Suponhamos também que existam matrizes positivas definidas $\Lambda(\phi)$ e $\Gamma(\phi)$ para as quais as matrizes $\overline{\Lambda}_n(\phi)$ e $\overline{\Gamma}_n(\phi)$ convergem quando n cresce para infinito, e definamos as seguintes partições para estas matrizes:

$$\Lambda(\phi) = \begin{pmatrix} \Lambda_{\beta\beta}(\phi) & \Lambda_{\beta\lambda}(\phi) \\ \Lambda_{\lambda\beta}(\phi) & \Lambda_{\lambda\lambda}(\phi) \end{pmatrix}$$

e

$$\Gamma(\phi) = \begin{pmatrix} \Gamma_{\beta\beta}(\phi) & \Gamma_{\beta\lambda}(\phi) \\ \Gamma_{\lambda\beta}(\phi) & \Gamma_{\lambda\lambda}(\phi) \end{pmatrix}.$$

Seja então $\hat{\phi}_{ec} = (\hat{\beta}_{ec}^{'}, \hat{\lambda}_{ec}^{'})'$ e seja $\hat{\phi}_0 = (\beta_0^{'}, \hat{\lambda}_0^{'})'$ a estimativa escore corrigido restrita a H_0 e consideremos a partição

$$\mathbf{U}^*(\boldsymbol{\phi}) = \begin{pmatrix} \mathbf{U}^*_{\boldsymbol{\beta}}(\boldsymbol{\phi}) \\ \mathbf{U}^*_{\boldsymbol{\lambda}}(\boldsymbol{\phi}) \end{pmatrix}.$$

Segue que as estatísticas

$$W = n\left(\hat{\boldsymbol{\beta}}_{ec} - \boldsymbol{\beta}_0\right)' \hat{\boldsymbol{\Omega}}^{-1}_{\boldsymbol{\beta\beta}}(\hat{\boldsymbol{\phi}}_0)\left(\hat{\boldsymbol{\beta}}_{ec} - \boldsymbol{\beta}_0\right)$$

e

$$Q = n^{-1}\mathbf{U}^*_{\boldsymbol{\beta}}(\hat{\boldsymbol{\phi}}_0)'\hat{\boldsymbol{\Lambda}}^{-1}_{\boldsymbol{\beta\beta.\lambda}}(\hat{\boldsymbol{\phi}}_0)\hat{\boldsymbol{\Omega}}^{-1}_{\boldsymbol{\beta\beta}}(\hat{\boldsymbol{\phi}}_0)\hat{\boldsymbol{\Lambda}}^{-1}_{\boldsymbol{\beta\beta.\lambda}}(\hat{\boldsymbol{\phi}}_0)\mathbf{U}^*_{\boldsymbol{\beta}}(\hat{\boldsymbol{\phi}}_0)$$

têm distribuição assintótica $\chi^2_{(p+1)}$, com $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\beta\beta.\lambda}}(\hat{\boldsymbol{\phi}}_0)$ e $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\beta\beta}}(\hat{\boldsymbol{\phi}}_0)$ estimativas consistentes de $\boldsymbol{\Lambda}_{\boldsymbol{\beta\beta.\lambda}}(\boldsymbol{\phi}_0)$ e $\boldsymbol{\Omega}_{\boldsymbol{\beta\beta}}(\boldsymbol{\phi}_0)$, respectivamente, definidas por

$$\boldsymbol{\Lambda}_{\boldsymbol{\beta\beta.\lambda}}(\boldsymbol{\phi}_0) = \boldsymbol{\Lambda}_{\boldsymbol{\beta\beta}}(\boldsymbol{\phi}_0) - \boldsymbol{\Lambda}_{\boldsymbol{\beta\lambda}}(\boldsymbol{\phi}_0)\boldsymbol{\Lambda}^{-1}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\phi}_0)\boldsymbol{\Lambda}_{\boldsymbol{\lambda\beta}}(\boldsymbol{\phi}_0)$$

e

$$\boldsymbol{\Omega}_{\boldsymbol{\beta\beta}}(\boldsymbol{\phi}_0) = (\boldsymbol{\Lambda}^{-1}(\boldsymbol{\phi}_0)\boldsymbol{\Gamma}(\boldsymbol{\phi}_0)\boldsymbol{\Lambda}^{-1}(\boldsymbol{\phi}_0)')_{\boldsymbol{\beta\beta}}.$$

Na seção seguinte, considerando a expressão em (3.1) com distribuição Weibull, calculamos a expressão da log-verossimilhança corrigida para a obtenção de estimadores consistentes. No Capítulo 4, utilizaremos esta log-verossimilhança para o estudo de influência local.

3.3 Modelo Paramétrico com Distribuição Weibull

Assumindo uma distribuição Weibull com vetor de parâmetros $\boldsymbol{\lambda} = (\rho, \gamma)$ e considerando as verdadeiras covariáveis $\mathbf{z}$, temos que a log-verossimilhança, dada por (3.1), é expressa como

$$
\begin{aligned}
l(\boldsymbol{\phi};\mathbf{D}) &= \sum_{i=1}^{n}\left\{\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta}_x + z_i\beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right]\right\} \\
&\quad - \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i'\boldsymbol{\beta}_x + z_i\beta_z)\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\}. \qquad (3.5)
\end{aligned}
$$

A log-verossimilhança naive fica

$$
\begin{aligned}
l(\boldsymbol{\phi};\bar{\mathbf{D}}) &= \sum_{i=1}^{n}\left\{\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right]\right\} \\
&\quad - \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z)\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\}. \qquad (3.6)
\end{aligned}
$$

Vamos agora obter a log-verossimilhança corrigida. Temos que

$$
\begin{aligned}
E\left[l(\boldsymbol{\phi};\bar{\mathbf{D}})|\mathbf{D}\right] &= \sum_{i=1}^{n}\left\{\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta}_x + z_i\beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right]\right\} \\
&\quad - \sum_{i=1}^{n}\left\{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + z_i\beta_z + \frac{\beta_z^2\sigma_u^2}{2}\right)\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\}
\end{aligned}
$$

pois, $E\left[\nu_i w_i\beta_z|\mathbf{D}\right] = \nu_i z_i\beta_z$ e $E\left[\exp(w_i\beta_z)|\mathbf{D}\right] = \exp\left(z_i\beta_z + \frac{\beta_z^2\sigma_u^2}{2}\right)$, $i = 1,\ldots,n$.

Portanto, denotando por $\eta = \frac{\beta_z^2\sigma_u^2}{2}$, temos que a log-verossimilhança corrigida neste caso é expressa como

$$
\begin{aligned}
l^*(\boldsymbol{\phi};\bar{\mathbf{D}}) &= \sum_{i=1}^{n}\left\{\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right]\right\} \\
&\quad - \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta)\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\}. \qquad (3.7)
\end{aligned}
$$

Com a log-verossimilhança acima, calculamos então o vetor escore corrigido $\mathbf{U}^*$, o qual possui dimensão $(p+3)$, com $p = \dim(\boldsymbol{\beta}_x)$, e a forma geral

$$\mathbf{U}^*(\boldsymbol{\phi}) = \left(\mathbf{U}^{\boldsymbol{\beta}_x\prime}, U^{\beta_z}, U^{\rho}, U^{\gamma}\right)' = \left(U_1^{\boldsymbol{\beta}_x}, \ldots, U_p^{\boldsymbol{\beta}_x}, U^{\beta_z}, U^{\rho}, U^{\gamma}\right)',$$

sendo

$$U_k^{\boldsymbol{\beta}_x} = \sum_{i=1}^{n} \left\{ x_{ik} \left[\nu_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta\right) \left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right] \right] \right\}, \; k = 1, \ldots, p,$$

$$U^{\beta_z} = \sum_{i=1}^{n} \left\{ \nu_i w_i - \left(w_i - \beta_z\sigma_u^2\right) \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta\right) \left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right] \right\},$$

$$U^{\rho} = \sum_{i=1}^{n} \left\{ \nu_i \left[\frac{1}{\rho} + \ln(y_i)\right] - y_i^{\rho}e^{\gamma}\ln(y_i) \left[\nu_i + \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\right] \right\},$$

e

$$U^{\gamma} = \sum_{i=1}^{n} \left\{ \nu_i \left[1 - y_i^{\rho}e^{\gamma}\right] - y_i^{\rho}e^{\gamma} \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right) \right\}.$$

É possível mostrar que neste caso $\mathbf{U}^*$ satisfaz a relação (3.2).

A matriz $\ddot{\mathbf{L}}^*$ (simétrica) pode ser expressa na forma

$$\ddot{\mathbf{L}}^*(\boldsymbol{\phi}) = \begin{pmatrix} \ddot{\mathbf{L}}^{\boldsymbol{\beta}_x\boldsymbol{\beta}_x} & \ddot{\mathbf{L}}^{\boldsymbol{\beta}_x\beta_z} & \ddot{\mathbf{L}}^{\boldsymbol{\beta}_x\rho} & \ddot{\mathbf{L}}^{\boldsymbol{\beta}_x\gamma} \\ & \ddot{L}^{\beta_z\beta_z} & \ddot{L}^{\beta_z\rho} & \ddot{L}^{\beta_z\gamma} \\ & & \ddot{L}^{\rho\rho} & \ddot{L}^{\rho\gamma} \\ & & & \ddot{L}^{\gamma\gamma} \end{pmatrix}, \tag{3.8}$$

a qual possui dimensão $(p+3)$, sendo suas submatrizes dadas pelas seguintes expressões:

$$\ddot{L}_{k,l}^{\boldsymbol{\beta}_x\boldsymbol{\beta}_x} = -\sum_{i=1}^{n} \left\{ x_{ik}x_{il} \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta\right) \left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right] \right\}, \; k, l = 1, \ldots, p,$$

$$\ddot{L}_k^{\boldsymbol{\beta}_x\beta_z} = -\sum_{i=1}^{n} \left\{ x_{ik} \left(w_i - \beta_z\sigma_u^2\right) \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta\right) \left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right] \right\},$$

$$k = 1, \ldots, p,$$

$$\ddot{L}_k^{\beta_x\rho} = -\sum_{i=1}^{n}\left\{x_{ik}y_i^{\rho}e^{\gamma}\ln(y_i)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\right\},\ k = 1,\ldots,p,$$

$$\ddot{L}_k^{\beta_x\gamma} = -\sum_{i=1}^{n}\left\{x_{ik}y_i^{\rho}e^{\gamma}\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\right\},\ k = 1,\ldots,p,$$

$$\ddot{L}^{\beta_z\beta_z} = -\sum_{i=1}^{n}\left\{\left[\left(w_i - \beta_z\sigma_u^2\right)^2 - \sigma_u^2\right]\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta\right)\left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\},$$

$$\ddot{L}^{\beta_z\rho} = -\sum_{i=1}^{n}\left\{y_i^{\rho}e^{\gamma}\ln(y_i)\left(w_i - \beta_z\sigma_u^2\right)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\right\},$$

$$\ddot{L}^{\beta_z\gamma} = -\sum_{i=1}^{n}\left\{y_i^{\rho}e^{\gamma}\left(w_i - \beta_z\sigma_u^2\right)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\right\},$$

$$\begin{aligned}\ddot{L}^{\rho\rho} &= -\sum_{i=1}^{n}\left\{\nu_i\left[\frac{1}{\rho^2} + y_i^{\rho}e^{\gamma}(\ln(y_i))^2\right]\right\} \\ &\quad -\sum_{i=1}^{n}\left\{y_i^{\rho}e^{\gamma}(\ln(y_i))^2\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\left[1 - y_i^{\rho}e^{\gamma}\right]\right\},\end{aligned}$$

$$\ddot{L}^{\rho\gamma} = -\sum_{i=1}^{n}\left\{\nu_i y_i^{\rho}e^{\gamma}\ln(y_i) + y_i^{\rho}e^{\gamma}\ln(y_i)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\left[1 - y_i^{\rho}e^{\gamma}\right]\right\}$$

e

$$\ddot{L}^{\gamma\gamma} = -\sum_{i=1}^{n}\left\{\nu_i y_i^{\rho}e^{\gamma} + y_i^{\rho}e^{\gamma}\exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta - y_i^{\rho}e^{\gamma}\right)\left[1 - y_i^{\rho}e^{\gamma}\right]\right\}.$$

3.4 Estudo de Simulação

Com o objetivo de analisar o comportamento dos estimadores naive e escore corrigido, apresentamos aqui os resultados de um estudo de simulação do modelo visto na seção anterior, para diversas porcentagens de censura, tamanho de amostra e variância do erro de medida.

No processo de simulação, fixamos $\rho = 1$, ou seja, consideramos os riscos competitivos com função de distribuição Exponencial. Assumimos para cada indivíduo uma única covariável, z, sendo que esta foi considerada fixa, mas teve seu valor gerado a partir de uma distribuição Normal.

Os valores gerados para cada indivíduo i, $i = 1, \ldots, n$, foram:

→ **covariável:**

$w_i = z_i + u_i$, com

$z_i \sim N(0, 1)$,

$u_i \sim N(0, \sigma_u^2)$;

→ **tempos observados e indicador de censura:**

$y_i = \min\{t_i, c_i\}$, $\nu_i = I_{\{t_i \leq c_i\}}$, com

$T_i = \min\{R_{i0}, R_{i1}, \ldots, R_{i,N_i}\}$,

$R_{il} \sim Exponencial(e^{\gamma})$, $l = 1, \ldots, N_i$, $\gamma = -1,5$,

$N_i \sim Poisson(\theta_i)$, $\theta_i = \exp(\beta_z w_i)$, $\beta_z = 0,3$,

$C_i \sim Exponencial(e^{\mu})$, sendo $\mu = \mu(p_c)$ calculado de acordo com a proporção de censura fixada para a população de não curados p_c, através da relação $p_c = P(T > C)$.

Foram geradas amostras com proporções de censura $p_c = 0\%, p_c = 15\%, p_c = 25\%$ e

$p_c = 50\%$, tamanhos de amostra $n = 50, n = 100$ e $n = 300$ e variâncias do erro de medida $\sigma_u^2 = 0,1, \sigma_u^2 = 0,5$ e $\sigma_u^2 = 1$. Para cada combinação destes valores foram geradas 1.500 réplicas e tomadas a média das estimativas dos parâmetros, o erro padrão e a raiz quadrada do erro quadrático médio. Todo o processo foi implementado com a linguagem de programação Ox (http://www.nuff.ox.ac.uk/Users/Doornik/index.html). As estimativas de máxima verossimilhança foram obtidas com o uso da função de maximização MaxBFGS, o qual utiliza o método quasi-Newton de Broyden, Fletcher, Goldfarb e Shanno (BFGS) (ver Doornik (1996), p.240).

Apresentamos os resultados destas simulações nas tabelas subseqüentes. Denotamos por

- $\hat{\beta}_z$ e $\hat{\gamma}$: estimativas de máxima verossimilhança obtidas com amostras geradas sem erro de medida ($w_i = z_i$);
- $\hat{\beta}_{ec}$ e $\hat{\gamma}_{ec}$: estimativas do método do escore corrigido, obtidas através da maximização de 3.7;
- $\hat{\beta}_{naive}$ e $\hat{\gamma}_{naive}$: estimativas naive, obtidas maximizando (3.6).

Estimador	n = 50			n = 100			n = 300		
	média	EP	REQM	média	EP	REQM	média	EP	REQM
$\hat{\beta}_z$	0,3312	0,2020	0,2044	0,3198	0,1329	0,1344	0,3060	0,0755	0,0757
$\hat{\gamma}$	-1,6370	0,2507	0,2857	-1,5876	0,1637	0,1857	-1,5341	0,0854	0,0920
$\sigma_u^2 = 0,1$									
$\hat{\beta}_{ec}$	0,3356	0,2219	0,2247	0,3208	0,1430	0,1445	0,3074	0,0799	0,0802
$\hat{\beta}_{naive}$	0,2991	0,1924	0,1924	0,2885	0,1263	0,1268	0,2780	0,0715	0,0748
$\hat{\gamma}_{ec}$	-1,6389	0,2531	0,2887	-1,5886	0,1653	0,1875	-1,5345	0,0856	0,0923
$\hat{\gamma}_{naive}$	-1,6335	0,2512	0,2845	-1,5844	0,1641	0,1846	-1,5314	0,0852	0,0908
$\sigma_u^2 = 0,5$									
$\hat{\beta}_{ec}$	0,3574	0,3455	0,3503	0,3354	0,2246	0,2273	0,3125	0,0994	0,1002
$\hat{\beta}_{naive}$	0,2096	0,1562	0,1805	0,2082	0,1070	0,1410	0,2032	0,0612	0,1145
$\hat{\gamma}_{ec}$	-1,6442	0,2609	0,2981	-1,5936	0,1702	0,1943	-1,5361	0,0864	0,0936
$\hat{\gamma}_{naive}$	-1,6199	0,2488	0,2762	-1,5761	0,1633	0,1801	-1,5242	0,0847	0,0881
$\sigma_u^2 = 1$									
$\hat{\beta}_{ec}$	0,3680	0,6641	0,6676	0,3493	0,3805	0,3837	0,3223	0,1293	0,1312
$\hat{\beta}_{naive}$	0,1381	0,1238	0,2038	0,1486	0,0893	0,1758	0,1520	0,0528	0,1571
$\hat{\gamma}_{ec}$	-1,6338	0,2646	0,2965	-1,5974	0,1797	0,2044	-1,5390	0,0882	0,0964
$\hat{\gamma}_{naive}$	-1,6094	0,2448	0,2681	-1,5685	0,1620	0,1758	-1,5194	0,0844	0,0866

Tabela 3.1: *Estimativas para dados simulados com* $\beta_z = 0,3$ *e* $\gamma = -1,5$ *- sem censura.*

Estimador	n = 50			n = 100			n = 300		
	média	EP	REQM	média	EP	REQM	média	EP	REQM
$\hat{\beta}_z$	0,3219	0,2157	0,2169	0,3062	0,1372	0,1374	0,2887	0,0760	0,0768
$\hat{\gamma}$	-1,6077	0,3108	0,3289	-1,5314	0,2061	0,2085	-1,4433	0,1019	0,1166
$\sigma_u^2 = 0,1$									
$\hat{\beta}_{ec}$	0,3272	0,2360	0,2376	0,3070	0,1474	0,1475	0,2899	0,0802	0,0808
$\hat{\beta}_{naive}$	0,2909	0,2043	0,2045	0,2757	0,1299	0,1322	0,2621	0,0717	0,0811
$\hat{\gamma}_{ec}$	-1,6100	0,3142	0,3329	-1,5323	0,2076	0,2101	-1,4438	0,1020	0,1165
$\hat{\gamma}_{naive}$	-1,6036	0,3118	0,3286	-1,5275	0,2062	0,2080	-1,4402	0,1015	0,1178
$\sigma_u^2 = 0,5$									
$\hat{\beta}_{ec}$	0,3468	0,3551	0,3581	0,3194	0,1994	0,2003	0,2947	0,099	0,1001
$\hat{\beta}_{naive}$	0,2012	0,1625	0,1902	0,1983	0,1092	0,1492	0,1912	0,0613	0,1249
$\hat{\gamma}_{ec}$	-1,6156	0,3242	0,3442	-1,5387	0,2145	0,2179	-1,4457	0,1030	0,1165
$\hat{\gamma}_{naive}$	-1,5877	0,3088	0,3210	-1,5180	0,2053	0,2061	-1,4321	0,1008	0,1215
$\sigma_u^2 = 1$									
$\hat{\beta}_{ec}$	0,3840	0,6884	0,6935	0,3377	0,4265	0,4281	0,3047	0,1323	0,1324
$\hat{\beta}_{naive}$	0,1319	0,1273	0,2108	0,1406	0,0904	0,1833	0,1430	0,0530	0,1657
$\hat{\gamma}_{ec}$	-1,6026	0,3271	0,3428	-1,5394	0,2216	0,2251	-1,4491	0,1055	0,1171
$\hat{\gamma}_{naive}$	-1,5782	0,3026	0,3126	-1,5087	0,2027	0,2029	-1,4266	0,1003	0,1243

Tabela 3.2: *Estimativas para dados simulados com* $\beta_z = 0,3$ *e* $\gamma = -1,5$ - *15% de censura.*

Estimador	n = 50			n = 100			n = 300		
	média	EP	REQM	média	EP	REQM	média	EP	REQM
$\hat{\beta}_z$	0,3202	0,2272	0,2281	0,3014	0,1415	0,1415	0,2803	0,0785	0,0810
$\hat{\gamma}$	-1,6156	0,3766	0,3939	-1,5173	0,2542	0,2548	-1,3971	0,1240	0,1612
$\sigma_u^2 = 0,1$									
$\hat{\beta}_{ec}$	0,3264	0,2496	0,2510	0,3025	0,1516	0,1516	0,2815	0,0831	0,0851
$\hat{\beta}_{naive}$	0,2894	0,2151	0,2154	0,2713	0,1332	0,1362	0,2543	0,0743	0,0872
$\hat{\gamma}_{ec}$	-1,6187	0,3800	0,3981	-1,5184	0,2557	0,2564	-1,3976	0,1242	0,1610
$\hat{\gamma}_{naive}$	-1,6111	0,3767	0,3928	-1,5129	0,2541	0,2544	-1,3935	0,1233	0,1630
$\sigma_u^2 = 0,5$									
$\hat{\beta}_{ec}$	0,3408	0,3611	0,3634	0,3184	0,2394	0,2401	0,2864	0,1038	0,1046
$\hat{\beta}_{naive}$	0,1978	0,1692	0,1977	0,1948	0,1114	0,1533	0,1853	0,0634	0,1311
$\hat{\gamma}_{ec}$	-1,6247	0,3871	0,4067	-1,5255	0,2649	0,2662	-1,3998	0,1258	0,1608
$\hat{\gamma}_{naive}$	-1,5941	0,3720	0,3837	-1,5021	0,2534	0,2534	-1,3840	0,1221	0,1684
$\sigma_u^2 = 1$									
$\hat{\beta}_{ec}$	0,3539	0,7192	0,7212	0,3348	0,3990	0,4005	0,2962	0,1366	0,1366
$\hat{\beta}_{naive}$	0,1287	0,1322	0,2164	0,1379	0,0922	0,1866	0,1383	0,0547	0,1707
$\hat{\gamma}_{ec}$	-1,6071	0,3988	0,4130	-1,5282	0,2730	0,2745	-1,4038	0,1297	0,1615
$\hat{\gamma}_{naive}$	-1,5840	0,3681	0,3776	-1,4920	0,2498	0,2499	-1,3777	0,1213	0,1722

Tabela 3.3: *Estimativas para dados simulados com* $\beta_z = 0,3$ *e* $\gamma = -1,5$ - *25% de censura.*

Estimador	n = 50			n = 100			n = 300		
	média	EP	REQM	média	EP	REQM	média	EP	REQM
$\hat{\beta}_z$	0,3418	0,2917	0,2947	0,3213	0,1855	0,1867	0,2874	0,0968	0,0976
$\hat{\gamma}$	-1,8590	0,5690	0,6728	-1,7576	0,4808	0,5455	-1,5518	0,3610	0,3647
$\sigma_u^2 = 0,1$									
$\hat{\beta}_{ec}$	0,3538	0,3262	0,3306	0,3239	0,1994	0,2008	0,2897	0,1018	0,1023
$\hat{\beta}_{naive}$	0,3120	0,2791	0,2794	0,2895	0,1740	0,1744	0,2610	0,0907	0,0987
$\hat{\gamma}_{ec}$	-1,8655	0,5742	0,6806	-1,7589	0,4834	0,5484	-1,5527	0,3606	0,3644
$\hat{\gamma}_{naive}$	-1,8565	0,5699	0,6722	-1,7529	0,4832	0,5454	-1,5463	0,3610	0,3639
$\sigma_u^2 = 0,5$									
$\hat{\beta}_{ec}$	0,3644	0,4877	0,4919	0,3370	0,2568	0,2595	0,2973	0,1266	0,1267
$\hat{\beta}_{naive}$	0,2049	0,2111	0,2315	0,2055	0,1397	0,1687	0,1902	0,0760	0,1336
$\hat{\gamma}_{ec}$	-1,8625	0,5822	0,6858	-1,7673	0,4925	0,5603	-1,5573	0,3627	0,3672
$\hat{\gamma}_{naive}$	-1,8289	0,5599	0,6494	-1,7399	0,4865	0,5424	-1,5317	0,3624	0,3638
$\sigma_u^2 = 1$									
$\hat{\beta}_{ec}$	0,3627	0,7991	0,8016	0,3385	0,4707	0,4722	0,3072	0,1582	0,1584
$\hat{\beta}_{naive}$	0,1315	0,1541	0,2283	0,1403	0,1115	0,1948	0,1412	0,0641	0,1712
$\hat{\gamma}_{ec}$	-1,8094	0,5987	0,6739	-1,7507	0,5057	0,5644	-1,5609	0,3662	0,3712
$\hat{\gamma}_{naive}$	-1,8071	0,5546	0,6340	-1,7197	0,4879	0,5351	-1,5193	0,3639	0,3644

Tabela 3.4: *Estimativas para dados simulados com* $\beta_z = 0,3$ *e* $\gamma = -1,5$ - *50% de censura.*

Das Tabelas 3.1 a 3.4, podemos então extrair alguns resultados da simulação.

Observamos pelas amostras geradas sem erro de medida que o erro quadrático médio dos estimadores aumenta com o crescimento da proporção de censura e diminui com o aumento do tamanho da amostra, como esperado.

Os resultados também mostram claramente a atenuação do estimador de máxima verossimilhança naive $\hat{\beta}_{naive}$ de acordo com o aumento da variância do erro de medida. Notamos também que com o aumento da variância do erro de medida, a variabilidade de $\hat{\beta}_{naive}$ diminui e de $\hat{\beta}_{ec}$ aumenta. Ou seja, o estimador naive não capta o aumento na variabilidade do erro de medida, ao contrário do estimador do escore corrigido. Com este aumento do vício e diminuição da variância de $\hat{\beta}_{naive}$, observamos que para amostras maiores ($n = 300$) e variância do erro de medida grande ($\sigma_u^2 = 1$), o verdadeiro valor do parâmetro β_z não pertence ao intervalo de confiança calculado com o estimador naive.

O vício dos estimadores de γ apresenta pouca variação com o aumento da variância do erro de medida.

Os estimadores naive têm menor variabilidade do que os estimadores do escore corrigido em todos os casos.

O erro quadrático médio dos estimadores naive é menor do que dos estimadores do escore corrigido em quase todos os casos (devido a menor variabilidade). Esta situação não ocorre para os estimadores de β_z quando as variâncias do erro de medida são maiores ($\sigma_u^2 = 0,5$ e $\sigma_u^2 = 1$) e o tamanho da amostra é grande. Observamos que, como os estimadores do método do escore corrigido são consistentes, o erro quadrático médio de $\hat{\beta}_{ec}$ tende a diminuir com o aumento do tamanho da amostra.

Em geral, na presença de erro de medida, ocorre um aumento no erro quadrático médio dos estimadores quando a proporção de censura aumenta, como esperado.

Capítulo 4

Influência Local

Dentro do contexto de metodologias para análise de diagnóstico, uma das técnicas atualmente mais utilizadas consiste na técnica de influência local, desenvolvida por Cook (1986). Neste método, pequenas perturbações são introduzidas no modelo ajustado ou no conjunto de dados, a fim de se investigar a influência que estas podem produzir nas inferências. Apresentamos aqui uma introdução a esta técnica e, em seguida, a sua aplicação nos modelos com fração de cura vistos anteriormente.

4.1 Metodologia

Seja LD a função afastamento pela verossimilhança, definida por

$$LD(\omega) = 2\left[l(\hat{\phi}) - l(\hat{\phi}_\omega)\right],$$

com

- ω: vetor de perturbações tal que $\omega \in \Omega \subseteq \mathbb{R}^q, \Omega$ aberto;
- $\hat{\phi}$: E.M.V. sob $l(\phi)$, com $l(\phi)$ denotando a log-verossimilhança do modelo ajustado, sendo ϕ vetor $p \times 1$ de parâmetros;

- $\hat{\phi}_\omega$: E.M.V. sob $l(\phi|\omega)$, com $l(\phi|\omega)$ função duas vezes continuamente diferenciável representando a log-verossimilhança construída considerando modelo ou dados perturbados pelo vetor ω, sob a suposição que existe $\omega_0 \in \Omega$ tal que $l(\phi|\omega_0) = l(\phi)$.

Cook sugere mensurar a influência das perturbações introduzidas no modelo ou nos dados através do vetor ω, fazendo um estudo da variação da função LD. Para isto, consideramos a superfície em $\mathbb{R}^{q+1}$ dada por

$$\boldsymbol{\alpha}(\omega) = \begin{pmatrix} \omega \\ LD(\omega) \end{pmatrix} \tag{4.1}$$

e analisamos o seu comportamento numa vizinhança do ponto de mínimo global ω_0 (vetor de não perturbação), usando o conceito em geometria diferencial de curvaturas normais.

Mostra-se que a curvatura normal desta superfície $\boldsymbol{\alpha}$ no ponto ω_0 na direção de um dado vetor unitário $\mathbf{v}$ pode ser expressa como

$$C_{\mathbf{v}} = 2\left|\mathbf{v}'\,\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}\,\mathbf{v}\right|, \tag{4.2}$$

sendo

$$\boldsymbol{\Delta} = \{\Delta_{ij}\}_{\substack{i=1,\ldots,p \\ j=1,\ldots,q}},$$

com

$$\Delta_{ij} = \left.\frac{\partial^2 l(\phi|\omega)}{\partial\phi_i\partial\omega_j}\right|_{\substack{\phi=\hat{\phi} \\ \omega=\omega_0}}$$

e $-\ddot{\mathbf{L}}$ a matriz de informação observada.

Através da expressão acima, podemos então caracterizar a superfície $\boldsymbol{\alpha}$ e, conseqüentemente, analisar o comportamento da função LD. Um procedimento bastante utilizado consiste em calcular a máxima curvatura $C_{\mathbf{v}_{\max}} = \max_{\mathbf{v}} C_{\mathbf{v}}$ e a direção correspondente $\mathbf{v}_{\max}$. Denotando por $\ddot{\mathbf{F}} = \boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}$, temos que $C_{\mathbf{v}_{\max}}$ corresponde ao maior autovalor

absoluto da matriz $\ddot{\mathbf{F}}$ e $\mathbf{v_{max}}$ corresponde ao autovetor normalizado associado. Os elementos de índice i de $\mathbf{v_{max}}$ que são relativamente grandes (em valor absoluto), indicam que perturbações nos correspondentes elementos ω_i de $\boldsymbol{\omega}$ podem produzir alterações consideráveis nas inferências. Assim, pontos que se destacam no gráfico de $\mathbf{v_{max}}$ contra a ordem das observações, indicam possíveis elementos influentes.

4.1.1 Gráfico de influência correspondente a um subvetor de parâmetros

Cook propõe também o uso da metodologia acima no caso específico em que há o interesse somente em parte do conjunto de parâmetros. Neste caso, se o vetor de parâmetros do modelo ajustado é da forma $\boldsymbol{\phi}' = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2')$, com $\boldsymbol{\phi}_1$ e $\boldsymbol{\phi}_2$ respectivamente vetores de dimensões $p_1 \times 1$ e $p_2 \times 1$, com $p_1 + p_2 = p$, e assumimos interesse somente em $\boldsymbol{\phi}_1$, então analogamente à expressão (4.1), consideramos agora a superfície

$$\boldsymbol{\alpha}_s(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ LD_s(\boldsymbol{\omega}) \end{pmatrix},$$

com LD_s a função afastamento pela verossimilhança definida como

$$LD_s(\boldsymbol{\omega}) = 2\left[l\left(\hat{\boldsymbol{\phi}}\right) - l\left(\hat{\boldsymbol{\phi}}_{1\boldsymbol{\omega}}, g\left(\hat{\boldsymbol{\phi}}_{1\boldsymbol{\omega}}\right)\right)\right],$$

sendo $\hat{\boldsymbol{\phi}}_{1\boldsymbol{\omega}}$ o subvetor obtido de $\hat{\boldsymbol{\phi}}_{\boldsymbol{\omega}}' = \left(\hat{\boldsymbol{\phi}}_{1\boldsymbol{\omega}}', \hat{\boldsymbol{\phi}}_{2\boldsymbol{\omega}}'\right)$ e $g(\boldsymbol{\phi}_1)$ a função que, para cada $\boldsymbol{\phi}_1$ fixado, maximiza $l(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$.

É possível mostrar que a curvatura correspondente neste caso pode ser obtida de maneira simplificada pela expressão

$$C_{\mathbf{v}}(\boldsymbol{\phi}_1) = 2\left|\mathbf{v}'\,\boldsymbol{\Delta}'(\ddot{\mathbf{L}}^{-1} - \mathbf{B_{22}})\boldsymbol{\Delta}\,\mathbf{v}\right|, \tag{4.3}$$

sendo

$$\mathbf{B_{22}} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{L}_{22}^{-1} \end{pmatrix}, \tag{4.4}$$

com $\mathbf{L}_{22}$ submatriz de $\ddot{\mathbf{L}}$ obtida segundo a partição

$$\ddot{\mathbf{L}} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix}.$$

4.2 Esquemas de Perturbação

De acordo com o modelo em análise e os interesses do pesquisador, escolhemos os possíveis tipos de perturbação a serem aplicados. Os esquemas mais comumente utilizados são:

Ponderação de casos:

Perturbamos o modelo ajustado através da introdução de pesos em cada elemento da log-verossimilhança, definindo

$$l(\boldsymbol{\phi}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \, l_i(\boldsymbol{\phi}).$$

Observe que este esquema pode ser visto como uma generalização da técnica de retirada de casos. O vetor de não-perturbação é $\boldsymbol{\omega}_0 = (1, \ldots, 1)'$.

Perturbação das respostas:

Adicionamos ao vetor de dados da variável resposta pequenas perturbações fazendo

$$y_i(\omega) = y_i + \omega_i,$$

Aqui, o vetor de não-perturbação é dado por $\boldsymbol{\omega}_0 = (0, \ldots, 0)'$.

Perturbação de uma covariável:

Perturbamos o vetor de dados da k-ésima covariável, fazendo

$$x_{ik}(\omega) = x_{ik} + s\omega_i, \quad i = 1, \ldots, n,$$

sendo s a norma do vetor $\mathbf{x_k} = (x_{1k}, \ldots, x_{nk})'$. Neste caso, temos $\boldsymbol{\omega}_0 = (0, \ldots, 0)'$.

Quando o modelo postulado é um modelo com erro nas variáveis, outro esquema de interesse é a perturbação da variância do erro de medida.

Perturbação da variância do erro de medida:

Introduzimos heteroscedasticidade no modelo, através de pesos multiplicativos na variância do erro de medida, definindo

$$Var(u_i|\omega) = \frac{\sigma_u^2}{\omega_i}, \quad \omega_i > 0, \quad i = 1, \ldots, n$$

Observe que aqui $\boldsymbol{\omega}_0 = (1, \ldots, 1)'$.

4.3 Influência Local no Modelo com Fração de Cura

Descrevemos nesta seção as matrizes necessárias para a aplicação da técnica de influência local no modelo com distribuição Weibull visto no Capítulo 2, considerando os esquemas de perturbação citados na seção anterior.

Consideremos o modelo visto na Seção 2.4 e seja $\hat{\boldsymbol{\phi}} = \left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\lambda}}'\right)'$ o vetor das estimativas dos parâmetros da log-verossimilhança (2.8).

Para a obtenção da curvatura

$$C_{\mathbf{v}} = 2\left|\mathbf{v}'\,\boldsymbol{\Delta}'(\ddot{\mathbf{L}})^{-1}\boldsymbol{\Delta}\,\mathbf{v}\right|,$$

calculamos inicialmente a matriz $\ddot{\mathbf{L}}$ que neste caso tem dimensão $(p+2)$ e é dada pela fórmula em (2.9), avaliada em $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$.

O cálculo da matriz $\mathbf{\Delta}$ depende do esquema de perturbação a ser considerado. Seja então $l(\boldsymbol{\phi}|\boldsymbol{\omega})$ a log-verossimilhança construída com base num determinado esquema de perturbação escolhido. A matriz $\mathbf{\Delta}$ possui então dimensão $(p+2) \times n$ e a forma geral

$$\mathbf{\Delta} = (\mathbf{\Delta}_1, \mathbf{\Delta}_2, \ldots, \mathbf{\Delta}_n), \tag{4.5}$$

sendo $\mathbf{\Delta}_j$, $j = 1, \ldots, n$, vetores coluna de tamanho $(p+2)$ dados por

$$\begin{aligned} \mathbf{\Delta}_j &\equiv \left(\Delta_{j_1}, \ldots, \Delta_{j_p}, \Delta_{j_{(p+1)}}, \Delta_{j_{(p+2)}}\right)' \\ &= \left(\frac{\partial^2 l(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\beta_1\partial\omega_j}, \ldots, \frac{\partial^2 l(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\beta_p\partial\omega_j}, \frac{\partial^2 l(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\rho\partial\omega_j}, \frac{\partial^2 l(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\gamma\partial\omega_j}\right)' \Bigg|_{\substack{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}} \\ \boldsymbol{\omega}=\boldsymbol{\omega}_0}}. \end{aligned}$$

A seguir, descrevemos a log-verossimilhança $l(\boldsymbol{\phi}|\boldsymbol{\omega})$ e os elementos da matriz $\mathbf{\Delta}$ para cada um dos esquemas considerados.

4.3.1 Ponderação de casos

Para este esquema, temos que a função de log-verossimilhança do modelo perturbado é escrita como

$$l(\boldsymbol{\phi}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \left\{ \nu_i \left[\mathbf{x}_i'\boldsymbol{\beta} + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho} e^{\gamma} \right] - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \left[1 - \exp\left(-y_i^{\rho} e^{\gamma}\right)\right] \right\}.$$

Segue então que os elementos dos vetores $\mathbf{\Delta}_j = \left(\Delta_{j_1}, \ldots, \Delta_{j_p}, \Delta_{j_{(p+1)}}, \Delta_{j_{(p+2)}}\right)'$, $j = 1, \ldots, n$, em (4.5) tomam as seguintes expressões:

$$\Delta_{j_l} = \nu_j x_{jl} - x_{jl} \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}}\right) \left[1 - \exp\left(-y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)\right], \; l = 1, \ldots, p,$$

$$\Delta_{j(p+1)} = \frac{\nu_j}{\hat{\rho}} + \nu_j \left(1 - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right) \ln(y_j) - y_j^{\hat{\rho}} e^{\hat{\gamma}} \ln(y_j) \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}} - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)$$

e

$$\Delta_{j(p+2)} = \nu_j \left(1 - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right) - y_j^{\hat{\rho}} \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}} + \hat{\gamma} - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right).$$

4.3.2 Perturbação das respostas

Introduzindo perturbações na variável resposta $\mathbf{y}$, construímos a seguinte função de log-verossimilhança:

$$\begin{aligned} l(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{i=1}^{n} \left\{ \nu_i \left[\mathbf{x}_i'\boldsymbol{\beta} + \gamma + \ln\left(\rho\left(y_i + \hat{\sigma}\omega_i\right)^{\rho-1}\right) - e^{\gamma}\left(y_i + \hat{\sigma}\omega_i\right)^{\rho} \right] \right\} \\ &\quad - \sum_{i=1}^{n} \left\{ \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \left[1 - \exp\left(-e^{\gamma}\left(y_i + \hat{\sigma}\omega_i\right)^{\rho}\right)\right] \right\}. \end{aligned}$$

Assim, para este esquema de perturbação, os elementos dos vetores coluna $\boldsymbol{\Delta}_j$, $j = 1, \ldots, n$, em (4.5) são escritos como

$$\Delta_{jl} = -x_{jl}\hat{\sigma}\hat{\rho}e^{\hat{\gamma}} y_j^{\hat{\rho}-1} \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}} - e^{\hat{\gamma}} y_j^{\hat{\rho}}\right),\ l = 1, \ldots, p,$$

$$\begin{aligned} \Delta_{j(p+1)} = &\frac{\nu_j \hat{\sigma}}{y_j} \left[1 - e^{\hat{\gamma}} y_j^{\hat{\rho}} \left(1 + \hat{\rho}\ln(y_j)\right)\right] \\ &+ \hat{\sigma}\hat{\rho}e^{\hat{\gamma}} y_j^{\hat{\rho}-1} \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}} - e^{\hat{\gamma}} y_j^{\hat{\rho}}\right) \ln(y_j) \left(1 + \frac{1}{\hat{\rho}\ln(y_j)} + e^{\hat{\gamma}} y_j^{\hat{\rho}}\right) \end{aligned}$$

e

$$\Delta_{j(p+2)} = -\hat{\sigma}\hat{\rho}e^{\hat{\gamma}} y_j^{\hat{\rho}-1} \left[\nu_j + \exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}} - e^{\hat{\gamma}} y_j^{\hat{\rho}}\right) \left(1 - e^{\hat{\gamma}} y_j^{\hat{\rho}}\right)\right].$$

4.3.3 Perturbação de uma covariável

Neste esquema, adicionamos perturbações na k-ésima coluna da matriz de covariáveis $\mathbf{X}$. Definindo o vetor

$$\widetilde{\mathbf{x}}_i' = (x_{i1}, x_{i2}, \ldots, x_{ik} + s\omega_i, \ldots, x_{ip}),\ i = 1, \ldots, n,$$

segue que a função de log-verossimilhança do modelo construído com base na perturbação acima é dada por:

$$\begin{aligned} l(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{i=1}^{n} \left\{ \nu_i \left[\widetilde{\mathbf{x}}_i' \boldsymbol{\beta} + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho} e^{\gamma} \right] \right\} \\ &\quad - \sum_{i=1}^{n} \left\{ \exp\left(\widetilde{\mathbf{x}}_i' \boldsymbol{\beta}\right) \left[1 - \exp\left(-y_i^{\rho} e^{\gamma}\right)\right] \right\}. \end{aligned}$$

Portanto, segue que os elementos dos vetores $\boldsymbol{\Delta}_j$, $j = 1, \ldots, n$ possuem neste caso as seguintes expressões:

$$\Delta_{jl} = \begin{cases} -s\hat{\beta}_k \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}\right) x_{jl} \left[1 - \exp\left(-y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)\right], & l = 1, \ldots, p,\ l \neq k, \\ s\left[\nu_j - \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}\right)\left(1 - \exp\left(-y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)\right)\left(1 + \hat{\beta}_k x_{jk}\right)\right], & l = k, \end{cases}$$

$$\Delta_{j_{(p+1)}} = -s\hat{\beta}_k y_j^{\hat{\rho}} e^{\hat{\gamma}} \ln(y_j) \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}} - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)$$

e

$$\Delta_{j_{(p+2)}} = -s\hat{\beta}_k y_j^{\hat{\rho}} e^{\hat{\gamma}} \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}} - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right).$$

4.4 Influência Local no Modelo com Fração de Cura com Erro nas Variáveis

Mostraremos a seguir as expressões obtidas para as matrizes constantes da fórmula da curvatura (4.2), quando consideramos o modelo com fração de cura com erro nas variáveis com distribuição Weibull visto no Capítulo 3.

Vimos na Seção 3.3 que a log-verossimilhança corrigida é dada por

$$\begin{aligned} l^*(\boldsymbol{\phi};\bar{\mathbf{D}}) &= \sum_{i=1}^{n}\left\{\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right]\right\} \\ &\quad - \sum_{i=1}^{n}\left\{\exp(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta)\left[1 - \exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\right\}, \end{aligned} \tag{4.6}$$

em que $\eta = \frac{\beta_z^2\sigma_u^2}{2}$.

Seja $\hat{\boldsymbol{\phi}} = \left(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\lambda}}'\right)'$, com $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_x', \hat{\beta}_z\right)'$ e $\hat{\boldsymbol{\lambda}} = (\hat{\rho}, \hat{\gamma})'$ o vetor das estimativas dos parâmetros da log-verossimilhança acima e definamos $\hat{\eta} = \frac{\hat{\beta}_z^2\sigma_u^2}{2}$.

A matriz $\ddot{\mathbf{L}}$ para o cálculo da curvatura (4.2) neste caso possui dimensão $(p+3)$, com $p = \dim(\boldsymbol{\beta}_x)$, e é obtida através da expressão (3.8) avaliada em $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$.

Vamos agora determinar a forma geral da matriz $\boldsymbol{\Delta}$ e nas subseções seguintes as suas expressões específicas para cada esquema de perturbação a ser considerado.

Seja então $l^*(\boldsymbol{\phi}|\boldsymbol{\omega})$ a log-verossimilhança construída de acordo com um determinado esquema de perturbação. Segue que a matriz $\boldsymbol{\Delta}$ neste caso tem dimensão $(p+3)\times n$ e possui a forma geral

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \ldots, \boldsymbol{\Delta}_n), \tag{4.7}$$

sendo $\boldsymbol{\Delta}_j$, $j = 1, \ldots, n$, vetores coluna de tamanho $(p+3)$ calculados por

$$\begin{aligned}\boldsymbol{\Delta}_j &\equiv \left(\Delta_{j_1},\ldots,\Delta_{j_p},\Delta_{j_{(p+1)}},\Delta_{j_{(p+2)}},\Delta_{j_{(p+3)}}\right)' \\ &= \left(\frac{\partial^2 l^*(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\beta_{x_1}\partial\omega_j},\ldots,\frac{\partial^2 l^*(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\beta_{x_p}\partial\omega_j},\frac{\partial^2 l^*(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\beta_z\partial\omega_j},\frac{\partial^2 l^*(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\rho\partial\omega_j},\frac{\partial^2 l^*(\boldsymbol{\phi}|\boldsymbol{\omega})}{\partial\gamma\partial\omega_j}\right)'\Bigg|_{\substack{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}\\ \boldsymbol{\omega}=\boldsymbol{\omega}_0}}.\end{aligned}$$

Para cada um dos esquemas de perturbação considerados, descrevemos a seguir, a log-verossimilhança perturbada $l^*(\boldsymbol{\phi}|\boldsymbol{\omega})$ e as expressões obtidas para cada um dos elementos da matriz $\boldsymbol{\Delta}$.

4.4.1 Ponderação de casos

Quando consideramos como esquema de perturbação a ponderação de casos, temos que a função de log-verossimilhança do modelo perturbado é escrita como

$$\begin{aligned}l^*(\boldsymbol{\phi}|\boldsymbol{\omega}) = \sum_{i=1}^{n}\omega_i\Big\{&\nu_i\left[\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho}e^{\gamma}\right] \\ &- \exp\left(\mathbf{x}_i'\boldsymbol{\beta}_x + w_i\beta_z - \eta\right)\left[1-\exp\left(-y_i^{\rho}e^{\gamma}\right)\right]\Big\}.\end{aligned}$$

Segue daí que os elementos dos vetores $\boldsymbol{\Delta}_j$, $j=1,\ldots,n$, em (4.7) tomam as seguintes expressões:

$$\Delta_{j_l} = \nu_j x_{jl} - x_{jl}\exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}}_x + w_j\hat{\beta}_z - \hat{\eta}\right)\left[1-\exp\left(-y_j^{\hat{\rho}}e^{\hat{\gamma}}\right)\right],\ l=1,\ldots,p,$$

$$\Delta_{j_{(p+1)}} = \nu_j w_j - \left(w_j - \hat{\beta}_z\sigma_u^2\right)\exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}}_x + w_j\hat{\beta}_z - \hat{\eta}\right)\left[1-\exp\left(-y_j^{\hat{\rho}}e^{\hat{\gamma}}\right)\right],$$

$$\Delta_{j_{(p+2)}} = \frac{\nu_j}{\hat{\rho}} + \nu_j\ln(y_j)\left(1-y_j^{\hat{\rho}}e^{\hat{\gamma}}\right) - y_j^{\hat{\rho}}e^{\hat{\gamma}}\ln(y_j)\exp\left(\mathbf{x}_j'\hat{\boldsymbol{\beta}}_x + w_j\hat{\beta}_z - \hat{\eta} - y_j^{\hat{\rho}}e^{\hat{\gamma}}\right)$$

e

$$\Delta_{j_{(p+3)}} = \nu_j \left(1 - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right) - y_j^{\hat{\rho}} e^{\hat{\gamma}} \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_x + w_j \hat{\beta}_z - \hat{\eta} - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right).$$

4.4.2 Perturbação da variância do erro de medida

Inserindo pesos na variância da variável u, que representa o erro de medida da covariável z, construímos a seguinte log-verossimilhança perturbada:

$$\begin{aligned} l^*(\boldsymbol{\phi}|\boldsymbol{\omega}) &= \sum_{i=1}^{n} \left\{ \nu_i \left[\mathbf{x}_i' \boldsymbol{\beta}_x + w_i \beta_z + \gamma + \ln\left(\rho y_i^{\rho-1}\right) - y_i^{\rho} e^{\gamma} \right] \right\} \\ &\quad - \sum_{i=1}^{n} \left\{ \exp(\mathbf{x}_i' \boldsymbol{\beta}_x + w_i \beta_z - \eta_i) \left[1 - \exp\left(-y_i^{\rho} e^{\gamma}\right)\right] \right\} \end{aligned}$$

em que $\eta_i = \frac{\beta_z^2 \sigma_u^2}{2\omega_i}$.

Obtemos assim, as seguintes expressões para os elementos dos vetores $\boldsymbol{\Delta}_j$, $j = 1, \ldots, n$:

$$\Delta_{jl} = -x_{jl} \frac{\hat{\beta}_z^2 \sigma_u^2}{2\omega_j^2} \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_x + w_j \hat{\beta}_z - \hat{\eta}_j\right) \left[1 - \exp\left(-y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)\right], \; l = 1, \ldots, p,$$

$$\Delta_{j_{(p+1)}} = -\frac{\sigma_u^2}{\omega_j^2} \left[\left(w_j - \frac{\hat{\beta}_z \sigma_u^2}{\omega_j} \right) \frac{\hat{\beta}_z^2}{2} + \hat{\beta}_z \right] \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_x + w_j \hat{\beta}_z - \hat{\eta}_j\right) \left[1 - \exp\left(-y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)\right],$$

$$\Delta_{j_{(p+2)}} = -y_j^{\hat{\rho}} e^{\hat{\gamma}} \ln(y_j) \frac{\hat{\beta}_z^2 \sigma_u^2}{2\omega_j^2} \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_x + w_j \hat{\beta}_z - \hat{\eta}_j - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right)$$

e

$$\Delta_{j_{(p+3)}} = -y_j^{\hat{\rho}} e^{\hat{\gamma}} \frac{\hat{\beta}_z^2 \sigma_u^2}{2\omega_j^2} \exp\left(\mathbf{x}_j' \hat{\boldsymbol{\beta}}_x + w_j \hat{\beta}_z - \hat{\eta}_j - y_j^{\hat{\rho}} e^{\hat{\gamma}}\right).$$

Capítulo 5

Aplicação

Como ilustração, apresentamos aqui a aplicação dos resultados desenvolvidos no Capítulo 4, mediante um conjunto de dados reais e conjuntos de dados simulados. A implementação de todos os programas computacionais foi realizada em linguagem de programação Ox (Doornik 1996).

5.1 Dados de Melanoma

Discutimos nesta seção a aplicação da teoria de influência local em um conjunto de dados reais sobre reincidência de câncer.

Os dados fazem parte de um ensaio clínico sobre melanoma cutâneo (tipo de câncer de pele maligno), para a avaliação do desempenho de tratamento pós-cirurgia, com alta dose de determinada droga (interferon alfa-2b) para prevenir a reincidência do câncer. O período de entrada dos pacientes no estudo foi de 1991 até 1995 e o seguimento foi realizado até 1998. Os dados foram obtidos de Ibrahim et al. (2001b) (para maiores informações a respeito do ensaio, ver Kirkwood et al. (2000)) e contêm dois tempos observados: o tempo de morte do paciente e o tempo de reincidência da doença. Analisaremos os dois casos, admitindo a variável T_1 representando o tempo, a partir da aleatorização, até a morte do paciente (a qual denominaremos sobrevida global) e a variável T_2 representando o tempo

a partir da aleatorização até a reincidência (denominada sobrevida livre de reincidência).

As variáveis $R_1, R_2, \ldots, R_N$ representam os tempos de promoção de cada uma das N células carcinogênicas do paciente.

O tamanho original da amostra é de $n = 427$ pacientes, sendo que 10 pacientes não apresentam valor para a covariável Breslow. Retirando estes casos, consideramos uma amostra de tamanho $n = 417$ pacientes . A porcentagem de observações censuradas na amostra, quando consideramos os tempos T_1, é de aproximadamente 56%. Para o caso em que consideramos os tempos T_2, 9 pacientes apresentam o tempo observado igual a zero, e trabalhamos então com uma amostra de tamanho $n = 408$ pacientes, com uma porcentagem de censura de aproximadamente 43%.

Para cada paciente i, $i = 1, \ldots, n$, temos associados os seguintes dados:

- y_i: tempo efetivamente observado (em anos);
- ν_i: indicador de censura (0 = censura, 1 = falha);
- x_{i1}: tratamento (0 = observação, 1 = interferon);
- x_{i2}: idade (em anos);
- x_{i3}: nódulo (categoria do nódulo: 1 a 4);
- x_{i4}: sexo (0 = masculino, 1 = feminino);
- x_{i5}: p.s. ("performance status" - escala de capacidade funcional do paciente quanto às suas atividades diárias: 0 = completamente ativo, 1 = outro);
- x_{i6}: Breslow (espessura do tumor, em mm).

Na Subseção 5.1.1, realizamos um estudo para este conjunto de dados, assumindo que as covariáveis foram observadas sem erro e, na Subseção 5.1.2, assumimos que a covariável

contínua Breslow tenha sido observada com erro e analisamos vários cenários para diversos valores da variância do erro de medida σ_u^2.

5.1.1 Estudo assumindo covariáveis sem erro de medida

I. Variável resposta: sobrevida global

O gráfico da função de sobrevivência, estimada de Kaplan-Meier, foi apresentado no Capítulo 1, Figura 1.1. Ajustando o modelo paramétrico Weibull, visto na Seção 2.4, obtivemos as seguintes estimativas (utilizando a função de maximização MaxBFGS do Ox no passo M do algoritmo EM):

Tabela 5.1: *Estimativas de máxima verossimilhança para dados de melanoma.*

Parâmetro	Estimativa	EP	p-Valor
$\hat{\beta}_{intercep.}$	-1,831	0,380	<0,0001
$\hat{\beta}_{tratam.}$	0,078	0,150	0,6001
$\hat{\beta}_{idade}$	0,009	0,006	0,1059
$\hat{\beta}_{nodulo}$	0,378	0,069	<0,0001
$\hat{\beta}_{sexo}$	-0,172	0,156	0,2700
$\hat{\beta}_{p.s.}$	0,159	0,211	0,4505
$\hat{\beta}_{Breslow}$	0,021	0,022	0,3316
$\hat{\rho}$	1,740	0,113	<0,0001
$\hat{\gamma}$	-1,647	0,133	<0,0001

A fração de cura média estimada foi 0,510.

Aplicando a teoria de influência local desenvolvida na Seção 4.3 a esse caso, apresentamos os resultado obtidos considerando-se os esquemas de ponderação de casos, perturbação das respostas e perturbação das covariáveis idade e Breslow.

Ponderação de casos:

Considerando como parâmetro de interesse o vetor $\boldsymbol{\beta}$, calculamos o vetor $\mathbf{v_{max}}$ correspondente à direção da maior curvatura dada por (4.3). Obtivemos como curvatura máxima o valor $C_{\mathbf{v_{max}}} = 2,50$. Na Tabela 5.2 a seguir, apresentamos os dados referentes aos indivíduos que mais se destacaram no gráfico de $|v_{max_i}|$ contra o índice das observações (ver Figura 5.1).

Quando retiramos da amostra este conjunto de observações, não há mudanças significativas nas estimativas. Observamos porém que, com a retirada somente das observações #47, #68 e #263, as quais correspondem às observações associadas aos maiores elementos positivos de $\mathbf{v_{max}}$ e são todas observações censuradas, ocorre uma mudança importante nos níveis de significância das covariáveis idade e Breslow (ver Tabela 5.3). A fração de cura média estimada, apresenta uma ligeira diminuição, passando de 0,510 para 0,506.

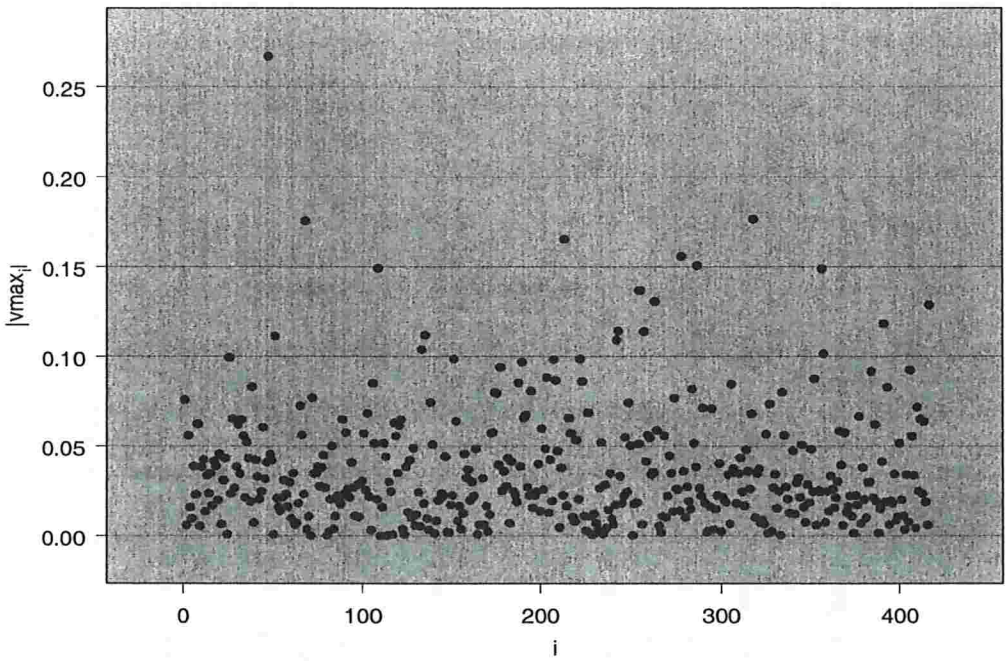

Figura 5.1: *Gráfico de influência - ponderação de casos com interesse em* β.

Tabela 5.2: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *- ponderação de casos.*

Indiv.	$\lvert v_{max_i}\rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
263	0,131	3,504	0	0	57,755	4	1	0	11,00
255	0,137	0,460	1	0	73,374	2	1	0	6,50
356	0,149	0,977	1	0	53,788	1	1	1	4,00
109	0,149	0,767	1	1	40,838	1	1	1	6,70
287	0,151	1,763	1	1	76,816	1	1	1	5,10
278	0,156	2,021	1	0	55,718	1	1	0	12,00
214	0,165	1,580	1	0	70,609	2	1	1	10,00
68	0,176	5,043	0	0	38,508	3	1	1	13,00
318	0,177	2,174	1	1	27,321	1	1	1	15,00
47	0,267	5,076	0	0	60,474	4	1	0	16,00

Tabela 5.3: *Estimativas dos coeficientes de regressão para dados de melanoma - ponderação de casos.*

	$\hat{\beta}_{interc}$	$\hat{\beta}_{tratam}$	$\hat{\beta}_{idade}$	$\hat{\beta}_{nodulo}$	$\hat{\beta}_{sexo}$	$\hat{\beta}_{p.s.}$	$\hat{\beta}_{Breslow}$
dados completos							
estimativa	-1,831	0,078	0,009	0,378	-0,172	0,159	0,021
erro padrão	0,380	0,150	0,006	0,069	0,156	0,211	0,022
p-valor	<0,0001	0,6001	0,1059	<0,0001	0,2700	0,4505	0,3316
dados s/ obs. 47							
estimativa	-1,927	0,061	0,010	0,393	-0,140	0,140	0,031
erro padrão	0,387	0,150	0,006	0,069	0,157	0,211	0,023
p-valor	<0,0001	0,6831	0,0858	<0,0001	0,3733	0,5079	0,1633
dados s/ obs. 47,68							
estimativa	-1,937	0,050	0,010	0,396	-0,121	0,177	0,037
erro padrão	0,386	0,150	0,006	0,069	0,157	0,211	0,023
p-valor	<0,0001	0,7393	0,0981	<0,0001	0,4420	0,4024	0,1066
dados s/ obs. 47,68,263							
estimativa	-2,004	0,032	0,010	0,407	-0,092	0,162	0,043
erro padrão	0,390	0,150	0,006	0,069	0,158	0,211	0,023
p-valor	<0,0001	0,8318	0,0841	<0,0001	0,5632	0,4442	0,0612

Perturbação das respostas

A seguir, analisamos a influência de perturbações nos tempos de sobrevivência observados.

O valor da curvatura máxima calculada foi $C_{\mathbf{v_{max}}} = 10.62$. A Figura 5.2, com o gráfico de $|v_{max_i}|$ contra o índice das observações, mostra claramente que alguns pontos se destacam dos demais. Na tabela 5.4 apresentamos os dados associados à estas observações. Analisando a amostra inicial, verifica-se que as observações em destaque se referem a pacientes com os menores tempos de sobrevivência, todos não censurados. As estimativas calculadas sem estas observações não apresentaram mudanças significativas.

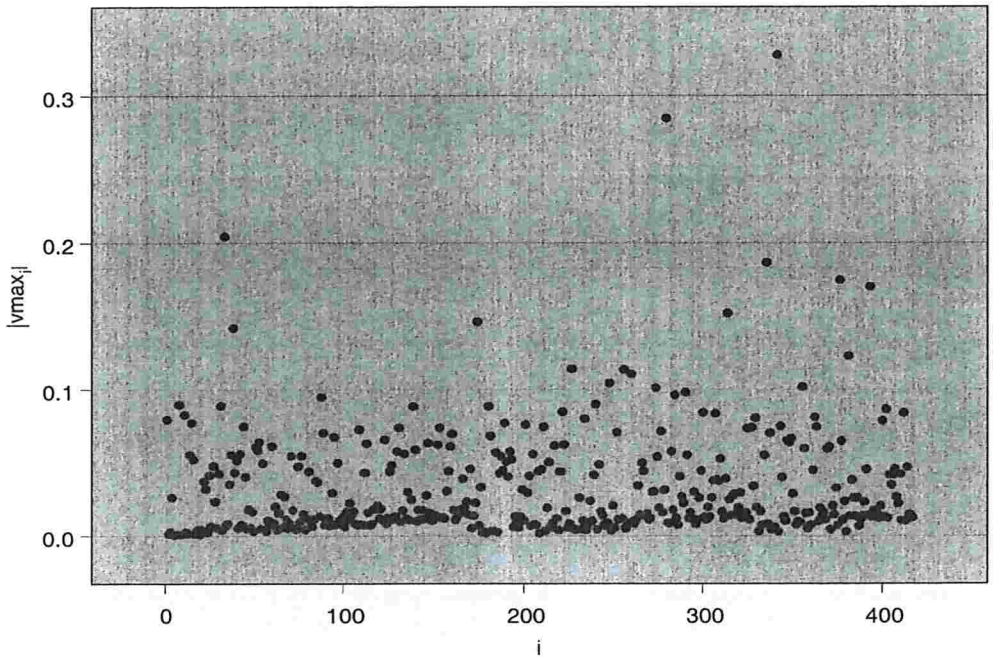

Figura 5.2: *Gráfico de influência - perturbação da resposta.*

Tabela 5.4: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *- perturbação da resposta.*

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
38	0,146	0,356	1	0	43,118	2	1	0	5,50
174	0,151	0,348	1	0	37,372	4	1	1	3,00
314	0,157	0,331	1	0	51,926	3	1	0	3,10
393	0,175	0,293	1	0	55,926	2	1	0	3,50
376	0,180	0,285	1	0	30,957	2	1	0	1,02
335	0,192	0,266	1	0	44,958	3	0	0	1,35
33	0,210	0,241	1	1	69,821	4	0	0	1,05
279	0,291	0,170	1	0	68,479	4	0	1	3,00
341	0,335	0,148	1	0	39,655	3	1	0	1,00

Perturbação das covariáveis idade e Breslow:

Investigamos aqui a perturbação dos vetores de covariáveis idade e Breslow.

Para a perturbação da covariável idade, obtivemos como curvatura máxima o valor $C_{\mathbf{v_{max}}} = 82,70$ e para a perturbação da covariável Breslow o valor $C_{\mathbf{v_{max}}} = 73,96$. Os respectivos gráficos de $|v_{max_i}|$ contra o índice das observações são apresentados na Figura 5.3 e Figura 5.4. Em ambos os gráficos algumas observações se destacam das demais. Os dados associados às dez observações com os maiores valores de $|v_{max_i}|$ em cada caso são listados na Tabela 5.5 e na Tabela 5.6. Para o caso da perturbação da covariável idade, destacam-se os pacientes com tempos de sobrevivência censurados, com categoria do nódulo 3 ou 4 e idade acima da média ($\bar{\mathbf{x}}_{idade} \cong 48$ anos). Para o caso da perturbação da covariável Breslow, observamos também que são todos pacientes com tempos censurados e categoria do nódulo 4.

As mudanças nas estimativas dos coeficientes de regressão com a retirada destes conjuntos de observações da amostra inicial, são apresentadas na Tabela 5.7. Nos resultados obtidos para os dois casos, verificamos um aumento na estimativa $\hat{\beta}_{idade}$, sendo que esta covariável passou a ser altamente significativa (p-valor $< 0,01$). Para o caso da perturbação da covariável Breslow, também observamos um aumento na correspondente estimativa de seu coeficiente e a diminuição do p-valor de $0,3316$ para $0,0257$. Houve também uma redução na fração de cura estimada, de $0,510$ para $0,493$ e $0,487$, respectivamente.

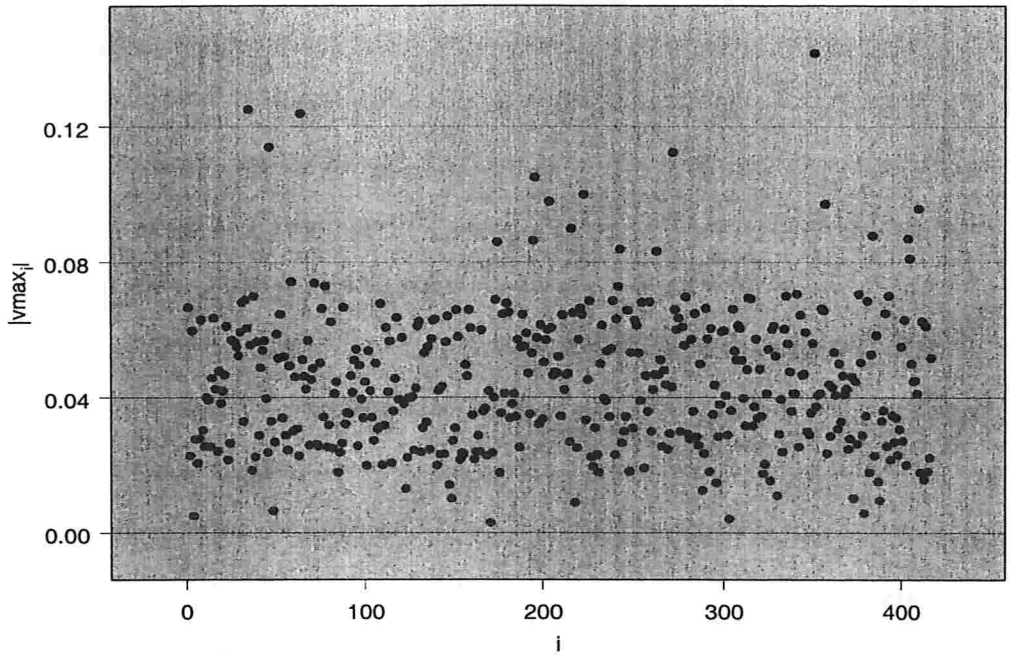

Figura 5.3: *Gráfico de influência - perturbação da covariável idade.*

Tabela 5.5: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *- perturbação da covariável idade.*

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
410	0,096	3,192	0	1	66,387	4	0	0	1,51
357	0,097	5,076	0	0	59,458	4	1	1	1,60
204	0,098	6,040	0	0	67,934	3	0	1	4,45
223	0,100	5,043	0	1	54,713	4	0	0	2,50
196	0,105	5,955	0	1	49,703	4	0	0	8,00
272	0,112	4,129	0	0	70,954	4	0	0	2,65
47	0,114	5,076	0	0	60,474	4	1	0	16,00
64	0,124	5,311	0	0	70,650	4	0	0	3,70
35	0,125	6,045	0	0	73,366	4	0	0	1,08
351	0,142	5,410	0	1	64,077	4	0	1	3,00

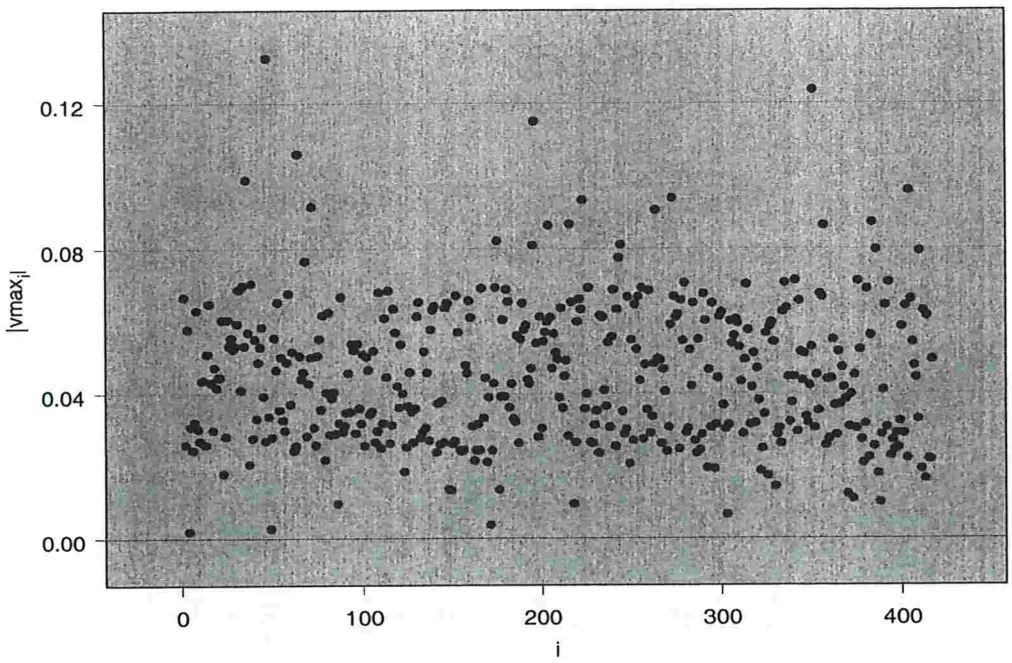

Figura 5.4: *Gráfico de influência - perturbação da covariável Breslow.*

Tabela 5.6: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *- perturbação da covariável Breslow.*

Indiv.	$\lvert v_{max_i}\rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
263	0,091	3,504	0	0	57,755	4	1	0	11,00
72	0,092	4,720	0	1	39,020	4	1	0	10,00
223	0,093	5,043	0	1	54,713	4	0	0	2,50
272	0,094	4,129	0	0	70,954	4	0	0	2,65
404	0,096	3,269	0	0	55,880	4	0	0	10,00
35	0,099	6,045	0	0	73,366	4	0	0	1,08
64	0,106	5,311	0	0	70,650	4	0	0	3,70
196	0,115	5,955	0	1	49,703	4	0	0	8,00
351	0,124	5,410	0	1	64,077	4	0	1	3,00
47	0,132	5,076	0	0	60,474	4	1	0	16,00

Tabela 5.7: *Estimativas dos coeficientes de regressão para dados completos e dados parciais de melanoma - perturbação nas covariáveis idade e Breslow.*

	$\hat{\beta}_{interc}$	$\hat{\beta}_{tratam}$	$\hat{\beta}_{idade}$	$\hat{\beta}_{nodulo}$	$\hat{\beta}_{sexo}$	$\hat{\beta}_{p.s.}$	$\hat{\beta}_{Breslow}$
dados completos							
estimativa	-1,831	0,078	0,009	0,378	-0,172	0,159	0,021
erro padrão	0,380	0,150	0,006	0,069	0,156	0,211	0,022
p-valor	<0,0001	0,6001	0,1059	<0,0001	0,2700	0,4505	0,3316
dados parciais - perturbação na cov. idade							
estimativa	-2,436	0,078	0,018	0,488	-0,244	0,256	0,029
erro padrão	0,404	0,150	0,006	0,071	0,157	0,212	0,022
p-valor	<0,0001	0,6035	0,0030	<0,0001	0,1207	0,2265	0,1839
dados parciais - perturbação na cov. Breslow							
estimativa	-2,451	0,051	0,016	0,508	-0,188	0,092	0,050
erro padrão	0,403	0,150	0,006	0,072	0,158	0,214	0,022
p-valor	<0,0001	0,7327	0,0080	<0,0001	0,2344	0,6691	0,0257

II. Variável resposta: sobrevida livre de reincidência

Na Figura 5.5, temos o gráfico da função de sobrevivência estimada (estimador Kaplan-Meier).

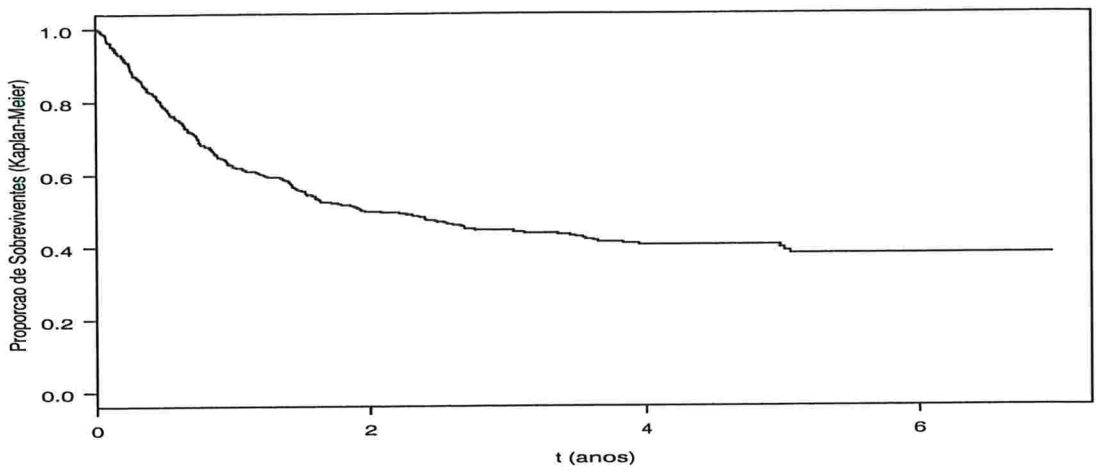

Figura 5.5: *Estimativas Kaplan-Meier para os dados de melanoma.*

Obtivemos neste caso as seguintes estimativas:

Tabela 5.8: *Estimativas de máxima verossimilhança para dados de melanoma (reincidência).*

Parâmetro	Estimativa	EP	p-Valor
$\hat{\beta}_{intercep.}$	-1,270	0,335	0,0002
$\hat{\beta}_{tratam.}$	-0,198	0,132	0,1348
$\hat{\beta}_{idade}$	0,010	0,005	0,0576
$\hat{\beta}_{nodulo}$	0,340	0,061	<0,0001
$\hat{\beta}_{sexo}$	-0,220	0,140	0,1161
$\hat{\beta}_{p.s.}$	0,071	0,193	0,7120
$\hat{\beta}_{Breslow}$	0,034	0,019	0,0672
$\hat{\rho}$	1,129	0,065	<0,0001
$\hat{\gamma}$	-0,561	0,102	<0,0001

O valor da fração de cura média estimada foi $0,384$.

Na seqüência, aplicamos a teoria de influência local ao conjunto de dados considerando os esquemas de ponderação de casos e perturbação das covariáveis idade e Breslow. Os resultados para o esquema de perturbação das respostas foram análogos aos obtidos na subseção anterior e não são apresentados aqui.

Ponderação de casos:

Tomando como parâmetro de interesse o subvetor $\boldsymbol{\beta}$, calculamos o vetor $\mathbf{v_{max}}$ correspondente à direção da maior curvatura, calculada como $C_{\mathbf{v_{max}}} = 2.80$. Na Figura 5.6 a seguir, apresentamos o gráfico de $|v_{max_i}|$ contra o índice das observações e na Tabela 5.9, apresentamos os dados referentes aos indivíduos associados aos maiores elementos de $|v_{max_i}|$.

Tabela 5.9: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *(reincidência) - ponderação de casos.*

Indiv.	$\lvert v_{max_i}\rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
132	0,095	0,342	1	0	61,780	2	0	0	15,00
162	0,096	3,124	0	0	48,827	2	1	0	8,10
31	0,108	5,676	0	0	47,554	1	0	0	14,00
274	0,113	0,750	1	0	55,718	1	1	0	12,00
408	0,116	0,392	1	0	46,872	1	1	0	10,50
251	0,117	0,033	1	0	73,374	2	1	0	6,50
344	0,119	5,410	0	1	64,077	4	0	1	3,00
70	0,156	4,720	0	1	39,020	4	1	0	10,00
259	0,335	3,504	0	0	57,755	4	1	0	11,00
46	0,595	5,076	0	0	60,474	4	1	0	16,00

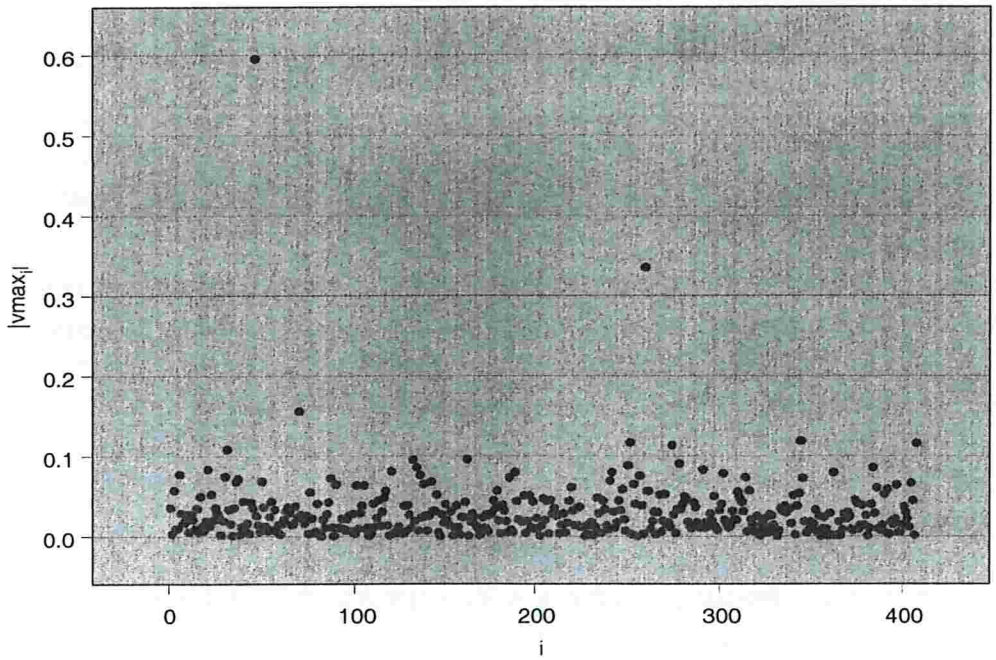

Figura 5.6: *Gráfico de influência - ponderação de casos com interesse em* β.

Observamos claramente no gráfico um destaque maior das observações #46 e #259, as quais são observações censuradas e que possuem valores grandes para as covariáveis nódulo e Breslow. Retiramos então da amostra somente estas duas observações e recalculamos as estimativas (ver Tabela 5.10). Dentro das mudanças observadas, destacamos a redução do nível de significância da covariável tratamento de $0,1348$ para $0,0612$. Também, a covariável idade passou a ser significativa ao nível de 5% e a covariável Breslow passou a ser altamente significativa.

Tabela 5.10: *Estimativas dos coeficientes de regressão para dados completos e dados parciais de melanoma (reincidência) - ponderação de casos.*

	$\hat{\beta}_{interc}$	$\hat{\beta}_{tratam}$	$\hat{\beta}_{idade}$	$\hat{\beta}_{nodulo}$	$\hat{\beta}_{sexo}$	$\hat{\beta}_{p.s.}$	$\hat{\beta}_{Breslow}$
dados completos							
estimativa	-1,270	-0,198	0,010	0,340	-0,220	0,071	0,034
erro padrão	0,335	0,132	0,005	0,061	0,140	0,193	0,019
p-valor	0,0002	0,1348	0,0576	<0,0001	0,1161	0,7120	0,0672
dados parciais							
estimativa	-1,517	-0,248	0,012	0,379	-0,131	0,013	0,056
erro padrão	0,349	0,133	0,005	0,061	0,142	0,194	0,019
p-valor	<0,0001	0,0612	0,0253	<0,0001	0,3568	0,9477	0,0039

Perturbação das covariáveis idade e Breslow:

Para as duas covariáveis, identificamos no gráfico de $|v_{max_i}|$ contra o índice das observações algumas possíveis observações influentes (Figuras 5.7 e 5.8). Os dados associados às dez observações com os maiores valores de $|v_{max_i}|$ em cada caso são listados nas Tabelas 5.11 e 5.12. Notamos, em ambos os casos, que destacam-se os pacientes com tempos de reincidência censurados e categoria do nódulo 3 ou 4.

Na Tabela 5.13 são apresentadas as alterações nas estimativas dos coeficientes de regressão com a retirada de algumas destas observações da amostra inicial. Para o caso da perturbação da covariável idade retiramos sete observações e no caso da covariável Breslow retiramos nove observações. Nos resultados obtidos para os dois casos, verificamos que a covariável tratamento passou a ser significativa ao nível de 5%. Também, as covariáveis idade e Breslow passaram a ser altamente significativas. Houve uma ligeira redução na fração de cura média estimada, de $0,384$ para 0.370 e $0,366$, respectivamente.

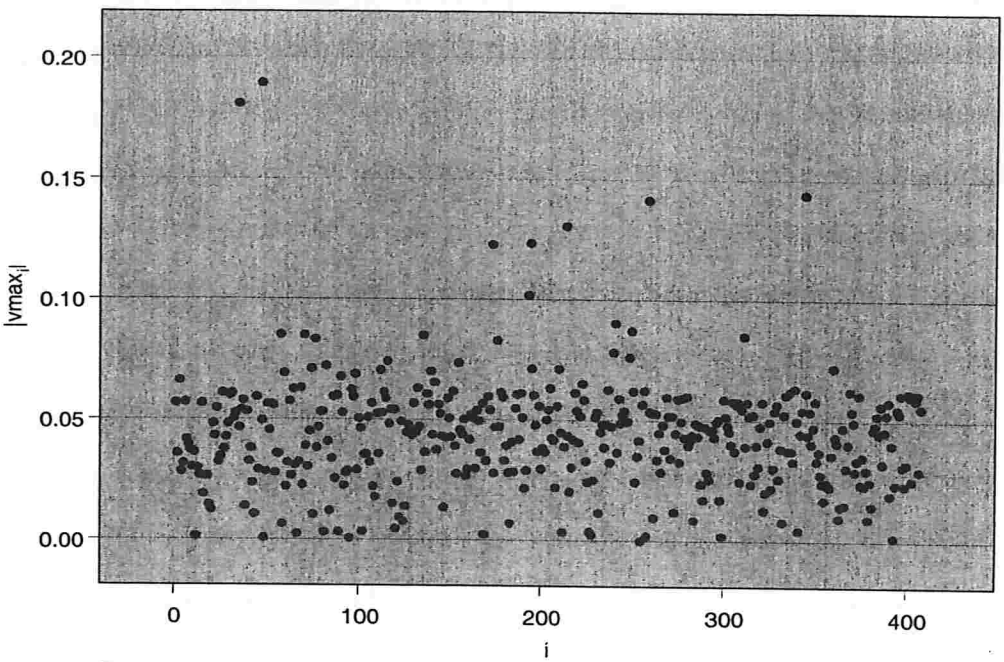

Figura 5.7: *Gráfico de influência - perturbação da covariável idade.*

Tabela 5.11: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *(reincidência) - perturbação da covariável idade.*

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	**tratam.**	**idade**	**nódulo**	**sexo**	**p.s.**	**Breslow**
250	0,087	4,923	0	0	40,099	3	0	0	9,16
241	0,090	5,150	0	1	46,787	4	0	0	0,82
193	0,102	3,907	0	0	52,953	4	1	1	1,37
173	0,123	5,999	0	0	51,989	4	0	0	1,36
194	0,123	5,955	0	1	49,703	4	0	0	8,00
214	0,131	5,771	0	0	53,531	4	0	0	2,40
259	0,141	3,504	0	0	57,755	4	1	0	11,00
344	0,144	5,410	0	1	64,077	4	0	1	3,00
34	0,181	6,045	0	0	73,366	4	0	0	1,08
46	0,189	5,076	0	0	60,474	4	1	0	16,00

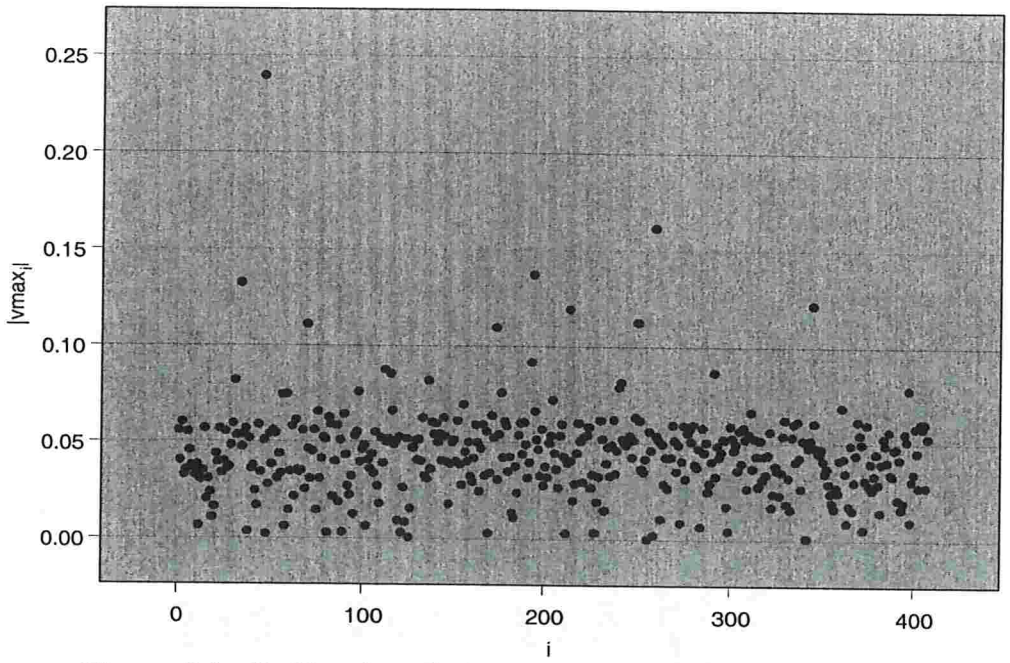

Figura 5.8: *Gráfico de influência - perturbação da covariável Breslow.*

Tabela 5.12: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ *(reincidência) - perturbação da covariável Breslow.*

Indiv.	$\lvert v_{max_i}\rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
193	0,092	3,907	0	0	52,953	4	1	1	1,37
173	0,110	5,999	0	0	51,989	4	0	0	1,36
70	0,111	4,720	0	1	39,020	4	1	0	10,00
250	0,113	4,923	0	0	40,099	3	0	0	9,16
214	0,119	5,771	0	0	53,531	4	0	0	2,40
344	0,122	5,410	0	1	64,077	4	0	1	3,00
34	0,132	6,045	0	0	73,366	4	0	0	1,08
194	0,137	5,955	0	1	49,703	4	0	0	8,00
259	0,162	3,504	0	0	57,755	4	1	0	11,00
46	0,240	5,076	0	0	60,474	4	1	0	16,00

Tabela 5.13: *Estimativas dos coeficientes de regressão para dados completos e dados parciais de melanoma (reincidência) - perturbação nas covariáveis idade e Breslow.*

	$\hat{\beta}_{interc}$	$\hat{\beta}_{tratam}$	$\hat{\beta}_{idade}$	$\hat{\beta}_{nodulo}$	$\hat{\beta}_{sexo}$	$\hat{\beta}_{p.s.}$	$\hat{\beta}_{Breslow}$
dados completos							
estimativa	-1,270	-0,198	0,010	0,340	-0,220	0,071	0,034
erro padrão	0,335	0,132	0,005	0,061	0,140	0,193	0,019
p-valor	0,0002	0,1348	0,0576	<0,0001	0,1161	0,7120	0,0672
dados parciais - perturbação na cov. idade							
estimativa	-1,745	-0,268	0,015	0,452	-0,229	0,020	0,054
erro padrão	0,354	0,133	0,005	0,063	0,143	0,196	0,019
p-valor	<0,0001	0,0444	0,0042	<0,0001	0,1098	0,9200	0,0051
dados parciais - perturbação na cov. Breslow							
estimativa	-1,770	-0,275	0,014	0,471	-0,209	-0,010	0,063
erro padrão	0,352	0,133	0,005	0,063	0,143	0,196	0,019
p-valor	<0,0001	0,0393	0,0063	<0,0001	0,1444	0,9603	0,0010

5.1.2 Estudo assumindo covariável com erro de medida

I. Variável resposta: sobrevida global

Sob a suposição de que os valores da covariável Breslow tenham sido observados com erro de medida, ajustamos ao conjunto de dados de melanoma o modelo visto na Seção 3.3, assumindo vários valores para a variância do erro de medida σ_u^2. Para cada valor assumido, aplicamos então a teoria de influência local considerando os esquemas de perturbação de ponderação de casos e de perturbação da variância do erro de medida. Nos dois casos, assumimos para σ_u^2 os valores: $\sigma_u^2 = 1$, $\sigma_u^2 = 2$ e $\sigma_u^2 = 5$ (observamos que a variância amostral da covariável Breslow é aproximadamente igual a 10,239).

Ponderação de casos:

Considerando como interesse o subvetor de parâmetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^{'}, \beta_z)^{'}$, aplicamos aqui o esquema de ponderação de casos.

Para $\sigma_u^2 = 1$, $\sigma_u^2 = 2$ e $\sigma_u^2 = 5$ obtivemos como curvatura máxima os valores $C_{\mathbf{v_{max}}} = 2,60$, $C_{\mathbf{v_{max}}} = 2,77$ e $C_{\mathbf{v_{max}}} = 3,73$, respectivamente. Nas Figuras 5.9 a 5.11, apresentamos os gráficos de $|v_{max_i}|$ contra o índice das observações e nas Tabelas 5.14 a 5.16 os respectivos dados associados às dez observações com maior destaque nestes gráficos. Observamos em todos os casos que os pacientes com maior destaque nos gráficos possuem os valores da covariável Breslow acima da média ($\bar{\mathrm{x}}_{Breslow} \cong 3,94$). Como ilustração, retiramos em cada caso as cinco observações com os maiores valores de $|v_{max_i}|$ e observamos as mudanças nas estimativas (ver Tabela 5.17).

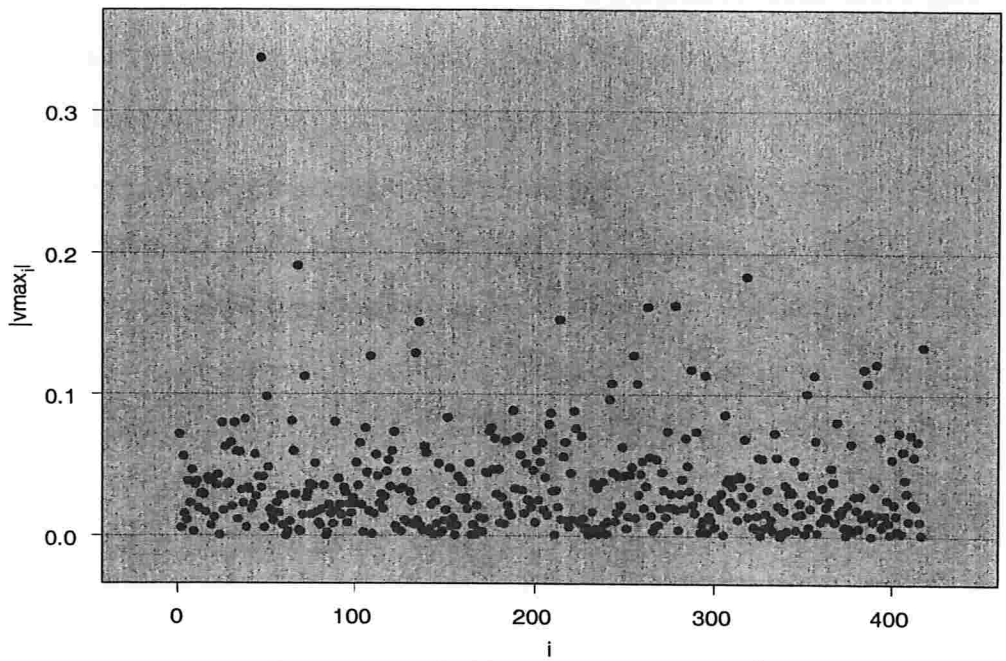

Figura 5.9: *Gráfico de influência* - $\sigma_u^2 = 1$.

Tabela 5.14: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ - *ponderação de casos com* $\sigma_u^2 = 1$.

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
255	0,128	0,460	1	0	73,374	2	1	0	6,50
134	0,129	1,342	1	0	61,780	2	0	0	15,00
417	0,134	2,979	1	0	46,872	1	1	0	10,50
136	0,151	2,220	1	1	29,547	1	0	0	20,00
214	0,153	1,580	1	0	70,609	2	1	1	10,00
263	0,162	3,504	0	0	57,755	4	1	0	11,00
278	0,163	2,021	1	0	55,718	1	1	0	12,00
318	0,184	2,174	1	1	27,321	1	1	1	15,00
68	0,191	5,043	0	0	38,508	3	1	1	13,00
47	0,338	5,076	0	0	60,474	4	1	0	16,00

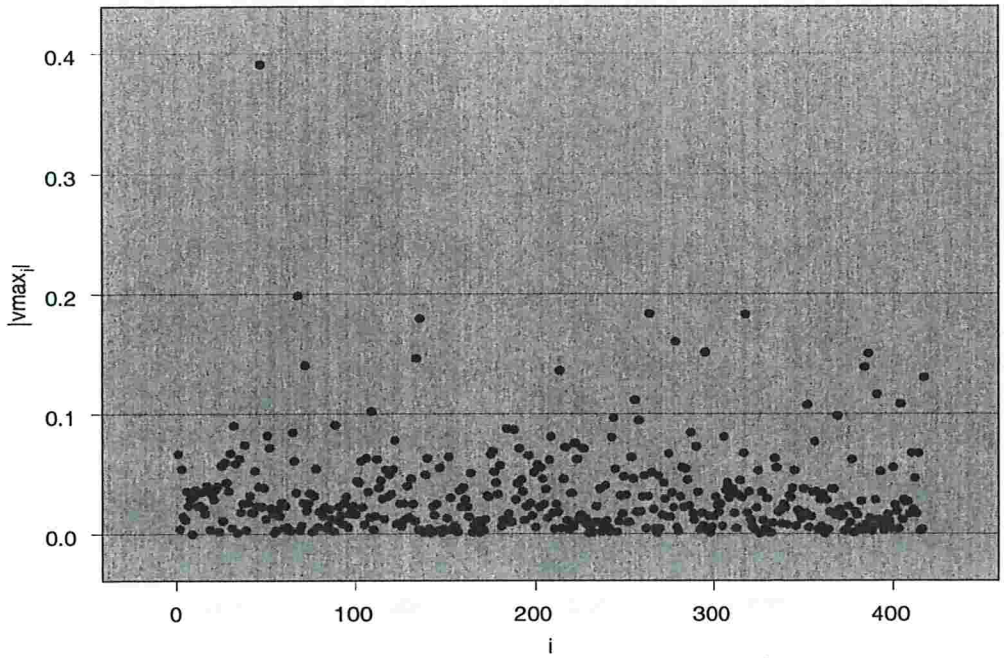

Figura 5.10: *Gráfico de influência* - $\sigma_u^2 = 2$.

Tabela 5.15: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ - *ponderação de casos com* $\sigma_u^2 = 2$.

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
72	0,141	4,720	0	1	39,020	4	1	0	10,00
134	0,147	1,342	1	0	61,780	2	0	0	15,00
386	0,150	3,652	0	1	39,647	3	0	0	15,00
295	0,152	2,867	0	0	37,769	3	0	0	17,00
278	0,160	2,021	1	0	55,718	1	1	0	12,00
136	0,180	2,220	1	1	29,547	1	0	0	20,00
318	0,183	2,174	1	1	27,321	1	1	1	15,00
263	0,184	3,504	0	0	57,755	4	1	0	11,00
68	0,199	5,043	0	0	38,508	3	1	1	13,00
47	0,391	5,076	0	0	60,474	4	1	0	16,00

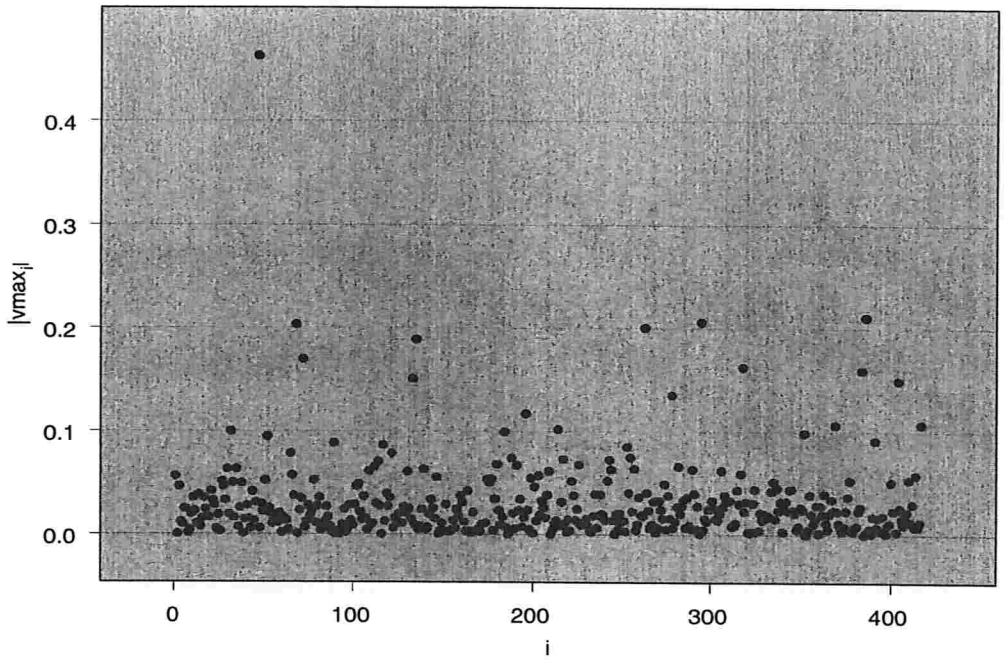

Figura 5.11: *Gráfico de influência* - $\sigma_u^2 = 5$.

Tabela 5.16: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ - *ponderação de casos com* $\sigma_u^2 = 5$.

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	**tratam.**	**idade**	**nódulo**	**sexo**	**p.s.**	**Breslow**
134	0,151	1,342	1	0	61,780	2	0	0	15,00
384	0,160	4,003	0	0	69,380	3	0	0	11,00
318	0,162	2,174	1	1	27,321	1	1	1	15,00
72	0,170	4,720	0	1	39,020	4	1	0	10,00
136	0,189	2,220	1	1	29,547	1	0	0	20,00
263	0,200	3,504	0	0	57,755	4	1	0	11,00
68	0,203	5,043	0	0	38,508	3	1	1	13,00
295	0,206	2,867	0	0	37,769	3	0	0	17,00
386	0,211	3,652	0	1	39,647	3	0	0	15,00
47	0,463	5,076	0	0	60,474	4	1	0	16,00

Tabela 5.17: *Estimativas dos coeficientes de regressão para dados completos e dados parciais de melanoma assumindo erro em uma covariável - ponderação de casos.*

	$\hat{\beta}_{interc}$	$\hat{\beta}_{tratam}$	$\hat{\beta}_{idade}$	$\hat{\beta}_{nodulo}$	$\hat{\beta}_{sexo}$	$\hat{\beta}_{p.s.}$	$\hat{\beta}_{Breslow}$
$\sigma_u^2 = 1$							
dados completos							
estimativa	-1,841	0,079	0,009	0,379	-0,171	0,159	0,023
erro padrão	0,391	0,154	0,006	0,070	0,163	0,222	0,025
p-valor	<0,0001	0,6083	0,1260	<0,0001	0,2926	0,4730	0,3456
dados parciais							
estimativa	-2,024	0,034	0,010	0,417	-0,119	0,135	0,039
erro padrão	0,401	0,155	0,006	0,070	0,167	0,225	0,026
p-valor	<0,0001	0,8248	0,0960	<0,0001	0,4744	0,5480	0,1288
$\sigma_u^2 = 2$							
dados completos							
estimativa	-1,853	0,079	0,009	0,380	-0,170	0,159	0,026
erro padrão	0,396	0,154	0,006	0,070	0,163	0,222	0,027
p-valor	<0,0001	0,6065	0,1266	<0,0001	0,2967	0,4729	0,3413
dados parciais							
estimativa	-2,062	0,018	0,011	0,417	-0,095	0,131	0,042
erro padrão	0,408	0,155	0,006	0,070	0,167	0,225	0,029
p-valor	<0,0001	0,9055	0,0780	<0,0001	0,5680	0,5611	0,1565
$\sigma_u^2 = 5$							
dados completos							
estimativa	-1,907	0,081	0,009	0,385	-0,165	0,159	0,036
erro padrão	0,416	0,154	0,006	0,070	0,165	0,221	0,036
p-valor	<0,0001	0,5975	0,1296	<0,0001	0,3159	0,4715	0,3146
dados parciais							
estimativa	-2,311	0,004	0,009	0,455	-0,040	0,149	0,109
erro padrão	0,428	0,159	0,006	0,071	0,169	0,218	0,034
p-valor	<0,0001	0,9822	0,1469	<0,0001	0,8124	0,4935	0,0014

Notamos que para os casos $\sigma_u^2 = 1$ e $\sigma_u^2 = 2$, com a retirada das cinco observações houve um aumento na estimativa $\hat{\beta}_{Breslow}$ e uma diminuição no p-valor associado. Destas cinco observações retiramos posteriormente nos dois casos somente as observações #47, #68 e #263, que são observações censuradas, e observamos que esta covariável passou a ser significativa ao nível de 10%, resultado análogo ao observado na Seção 5.1.1 para o esquema de ponderação de casos.

Para o caso $\sigma_u^2 = 5$, a retirada das cinco observações provocou um aumento na estimativa $\hat{\beta}_{Breslow}$ sendo que a covariável associada passou a ser altamente significativa.

Perturbação da variância do erro de medida:

Para este esquema de perturbação, apresentamos nas Figuras 5.12 a 5.14 os gráficos de $|v_{max_i}|$ contra o índice das observações e nas Tabelas 5.18 a 5.20 os respectivos dados associados às observações com os maiores valores de $|v_{max_i}|$.

Verificamos em todos os casos que a maioria dos dados são de pacientes com tempo de sobrevivência censurado e que todos têm categoria do nódulo com valor 4.

A fim de avaliar a sua influência nas estimativas, retiramos algumas destas observações da amostra inicial. Para os casos $\sigma_u^2 = 1$ e $\sigma_u^2 = 2$, retiramos as quatro primeiras observações e, para o caso $\sigma_u^2 = 5$, retiramos as três primeiras observações. Comparamos então as estimativas dos parâmetros obtidas para o conjunto de dados completo e o conjunto de dados parciais, para cada um dos valores assumidos para σ_u^2 (ver Tabela 5.21).

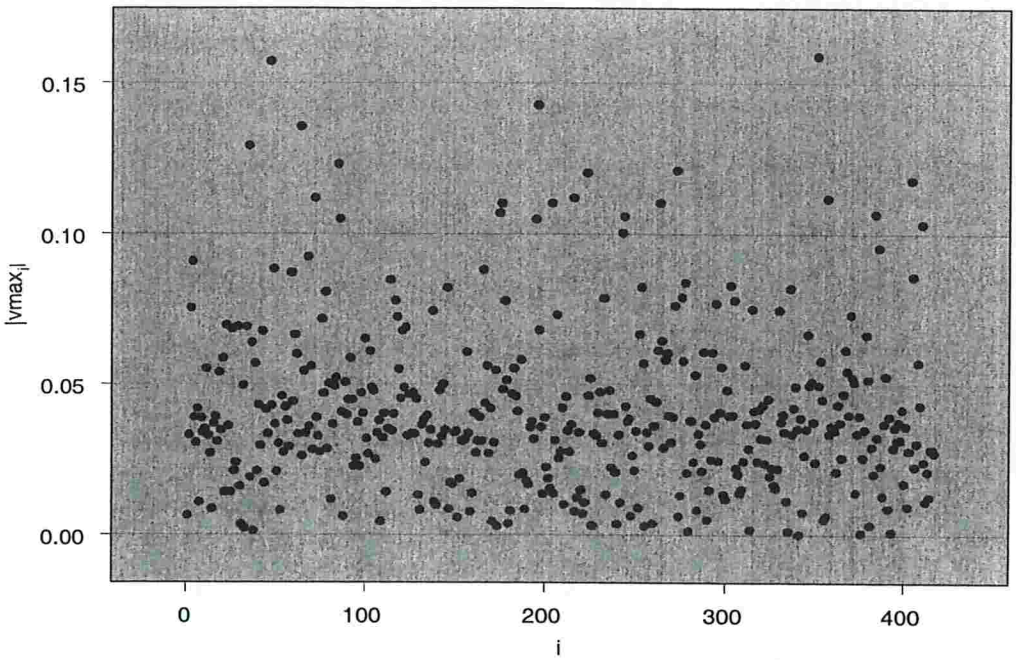

Figura 5.12: *Gráfico de influência* - $\sigma_u^2 = 1$.

Tabela 5.18: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ - *perturbação de* σ_u^2 *com* $\sigma_u^2 = 1$.

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
216	0,112	5,771	0	0	53,531	4	0	0	2,40
404	0,118	3,269	0	0	55,880	4	0	0	10,00
223	0,120	5,043	0	1	54,713	4	0	0	2,50
272	0,121	4,129	0	0	70,954	4	0	0	2,65
85	0,123	3,335	1	1	50,420	4	0	0	10,00
35	0,129	6,045	0	0	73,366	4	0	0	1,08
64	0,136	5,311	0	0	70,650	4	0	0	3,70
196	0,143	5,955	0	1	49,703	4	0	0	8,00
47	0,157	5,076	0	0	60,474	4	1	0	16,00
351	0,159	5,410	0	1	64,077	4	0	1	3,00

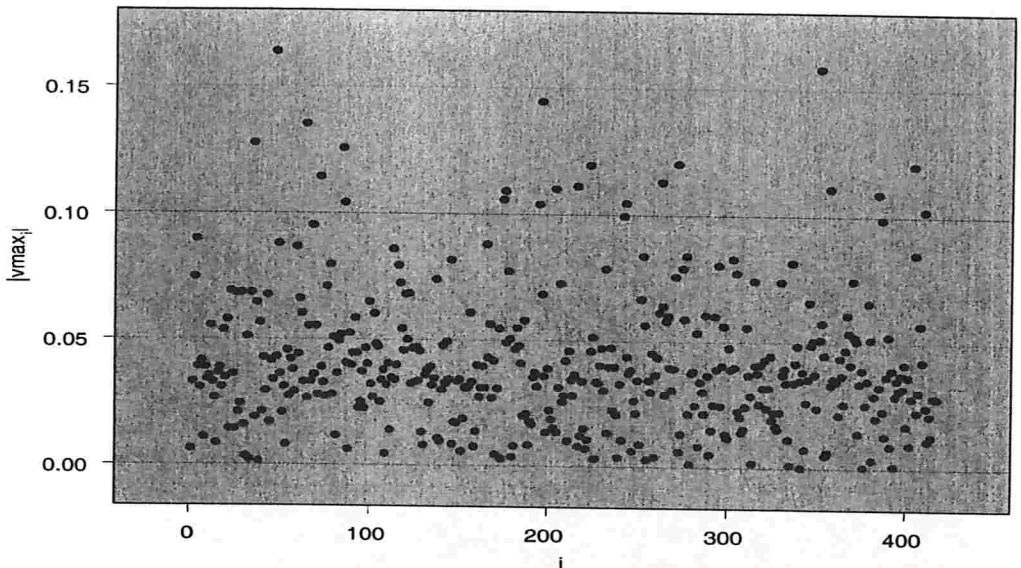

Figura 5.13: *Gráfico de influência* - $\sigma_u^2 = 2$.

Tabela 5.19: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ - *perturbação de* σ_u^2 *com* $\sigma_u^2 = 2$.

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
72	0,114	4,720	0	1	39,020	4	1	0	10,00
223	0,120	5,043	0	1	54,713	4	0	0	2,50
404	0,120	3,269	0	0	55,880	4	0	0	10,00
272	0,120	4,129	0	0	70,954	4	0	0	2,65
85	0,126	3,335	1	1	50,420	4	0	0	10,00
35	0,128	6,045	0	0	73,366	4	0	0	1,08
64	0,135	5,311	0	0	70,650	4	0	0	3,70
196	0,145	5,955	0	1	49,703	4	0	0	8,00
351	0,158	5,410	0	1	64,077	4	0	1	3,00
47	0,164	5,076	0	0	60,474	4	1	0	16,00

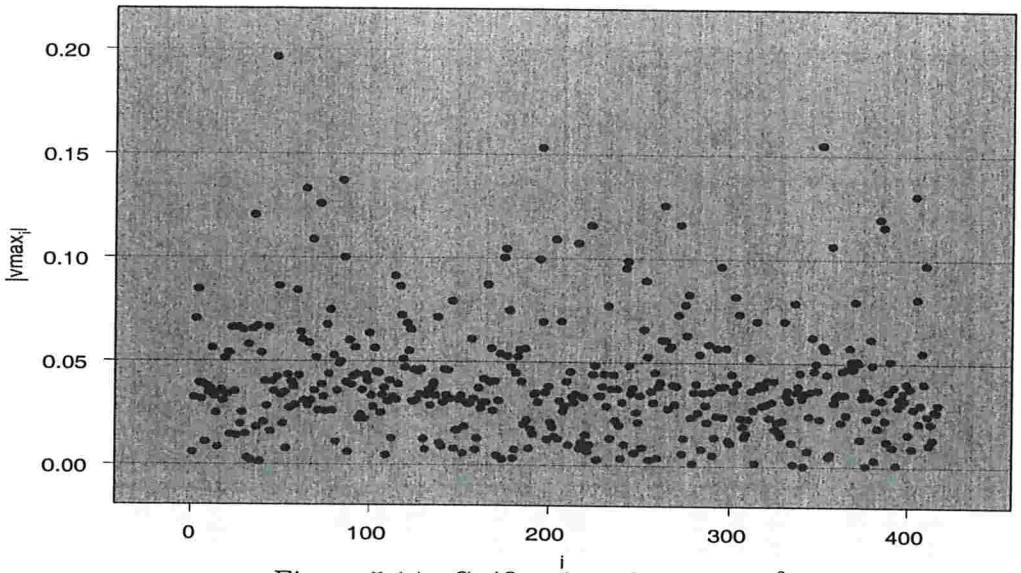

Figura 5.14: *Gráfico de influência* - $\sigma_u^2 = 5$.

Tabela 5.20: *Dados dos indivíduos associados às maiores componentes de* $\mathbf{v_{max}}$ - *perturbação de* σ_u^2 *com* $\sigma_u^2 = 5$.

Indiv.	$\lvert v_{max_i} \rvert$	y_i	ν_i	tratam.	idade	nódulo	sexo	p.s.	Breslow
35	0,120	6,045	0	0	73,366	4	0	0	1,08
263	0,126	3,504	0	0	57,755	4	1	0	11,00
72	0,126	4,720	0	1	39,020	4	1	0	10,00
404	0,131	3,269	0	0	55,880	4	0	0	10,00
64	0,133	5,311	0	0	70,650	4	0	0	3,70
85	0,137	3,335	1	1	50,420	4	0	0	10,00
196	0,153	5,955	0	1	49,703	4	0	0	8,00
351	0,155	5,410	0	1	64,077	4	0	1	3,00
47	0,196	5,076	0	0	60,474	4	1	0	16,00

Tabela 5.21: *Estimativas dos coeficientes de regressão para dados completos e dados parciais de melanoma - perturbação de* σ_u^2.

	$\hat{\beta}_{interc}$	$\hat{\beta}_{tratam}$	$\hat{\beta}_{idade}$	$\hat{\beta}_{nodulo}$	$\hat{\beta}_{sexo}$	$\hat{\beta}_{p.s.}$	$\hat{\beta}_{Breslow}$
$\sigma_u^2 = 1$							
dados completos							
estimativa	-1,841	0,079	0,009	0,379	-0,171	0,159	0,023
erro padrão	0,391	0,154	0,006	0,070	0,163	0,222	0,025
p-valor	<0,0001	0,6083	0,1260	<0,0001	0,2926	0,4730	0,3456
dados parciais							
estimativa	-2,129	0,095	0,012	0,430	-0,184	0,189	0,038
erro padrão	0,400	0,154	0,006	0,069	0,164	0,217	0,024
p-valor	<0,0001	0,5346	0,0494	<0,0001	0,2625	0,3842	0,1137
$\sigma_u^2 = 2$							
dados completos							
estimativa	-1,853	0,079	0,009	0,380	-0,170	0,159	0,026
erro padrão	0,396	0,154	0,006	0,070	0,163	0,222	0,027
p-valor	<0,0001	0,6065	0,1266	<0,0001	0,2967	0,4729	0,3413
dados parciais							
estimativa	-2,153	0,096	0,012	0,433	-0,180	0,188	0,043
erro padrão	0,405	0,154	0,006	0,069	0,164	0,217	0,027
p-valor	<0,0001	0,5339	0,0491	<0,0001	0,2724	0,3868	0,1083
$\sigma_u^2 = 5$							
dados completos							
estimativa	-1,907	0,081	0,009	0,385	-0,165	0,159	0,036
erro padrão	0,416	0,154	0,006	0,070	0,165	0,221	0,036
p-valor	<0,0001	0,5975	0,1296	<0,0001	0,3159	0,4715	0,3146
dados parciais							
estimativa	-2,202	0,112	0,011	0,431	-0,153	0,199	0,061
erro padrão	0,428	0,155	0,006	0,070	0,167	0,217	0,035
p-valor	<0,0001	0,4686	0,0698	<0,0001	0,3582	0,3593	0,0823

Notamos que com o aumento da variância do erro de medida, ocorre um aumento no valor estimado do coeficiente $\beta_{Breslow}$, tanto para os dados completos quanto para os dados parciais. Observamos também para este coeficiente uma diminuição no p-valor quando analisamos os dados parciais, sendo que para o caso $\sigma_u^2 = 5$ a covariável passa a ser significativa ao nível de 10%. Nas estimativas para a covariável idade, verificamos que esta passa a ser significativa com a retirada das observações.

5.2 Dados Simulados

Apresentamos nesta seção um estudo de influência local com dados simulados segundo o modelo com fração de cura visto na Seção 2.3, com o objetivo de avaliar perturbações sobre uma covariável (Seção 4.3.3).

Simulamos uma amostra de tamanho $n = 100$, gerando os seguintes valores para cada indivíduo i, $i = 1, \ldots, n$:

→ **covariável:**

$x_i \sim N(1, 1)$;

→ **tempos observados e indicador de censura:**

$y_i = \min\{t_i, c_i\}$, $\nu_i = I_{\{t_i \leq c_i\}}$, com

$T_i = \min\{R_{i0}, R_{i1}, \ldots, R_{i,N_i}\}$,

$R_{il} \sim Exponencial(e^{\gamma})$, $l = 1, \ldots, N_i$, $\gamma = -1,3$,

$N_i \sim Poisson(\theta_i)$, $\theta_i = \exp(\beta_0 + \beta_1 x_i)$, $\beta_0 = -0,1$, $\beta_1 = 0,3$,

$C_i \sim Exponencial(e^{\mu})$, sendo $\mu = \mu(p_c)$ calculado de acordo com a proporção de censura fixada para a população de não curados p_c, através da relação $p_c = P(T > C)$.

A covariável $\mathbf{x}$ teve seu valor máximo $\mathbf{x}_{max}$ perturbado, assumindo o novo valor $\mathbf{x}_{max} + \|\mathbf{x}\|$, com $\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}}$. Foram geradas amostras com diferentes proporções de censura ($p_c = 0\%$, $p_c = 15\%$, $p_c = 25\%$ e $p_c = 50\%$). Para cada uma destas amostras, obtivemos como curvatura máxima os valores $C_{\mathbf{v_{max}}} = 53,36$, $C_{\mathbf{v_{max}}} = 52,45$, $C_{\mathbf{v_{max}}} = 51,22$ e $C_{\mathbf{v_{max}}} = 50,34$, respectivamente. As Figuras 5.15 a 5.18 mostram os respectivos gráficos de $|v_{max_i}|$ contra o índice das observações. Observamos claramente nos gráficos, o maior destaque da observação perturbada (#31) em todos os casos.

As mudanças nas estimativas do coeficiente de regressão β_1 com a retirada desta ob-

servação das amostras, podem ser vistas na Tabela 5.22. Os resultados indicam que o valor de β_1 é subestimado na presença da observação extrema.

Tabela 5.22: *Estimativas do coeficiente de regressão β_1 para dados completos e dados parciais da simulação*

	dados completos	dados parciais
sem censura		
estimativa	0,122	0,364
erro padrão	0,085	0,130
p-valor	0,1500	0,0050
15% de censura		
estimativa	0,089	0,292
erro padrão	0,096	0,139
p-valor	0,3522	0,0360
25% de censura		
estimativa	0,143	0,436
erro padrão	0,093	0,150
p-valor	0,1227	0,0037
50% de censura		
estimativa	0,093	0,336
erro padrão	0,125	0,191
p-valor	0,4558	0,0790

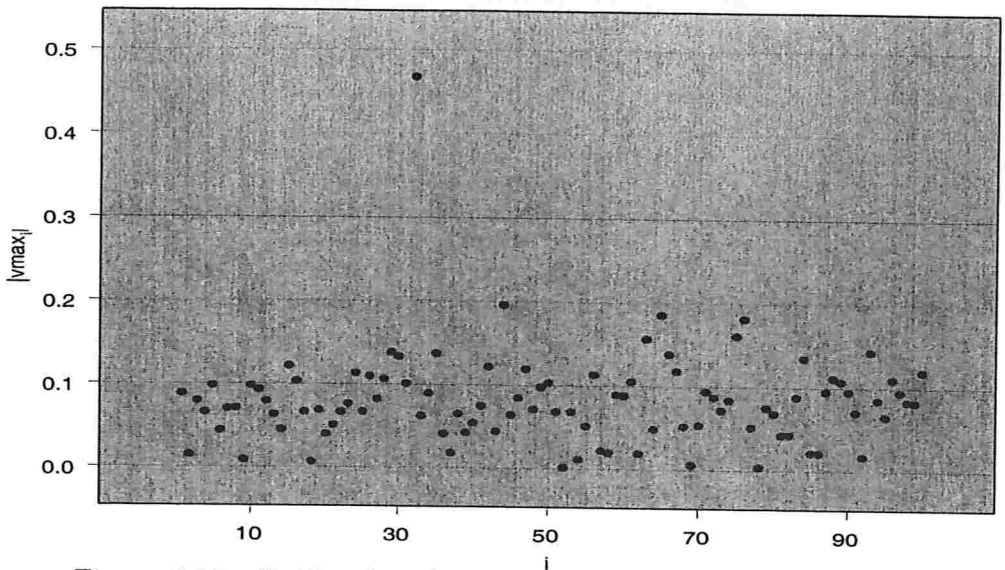

Figura 5.15: *Gráfico de influência para dados simulados - sem censura.*

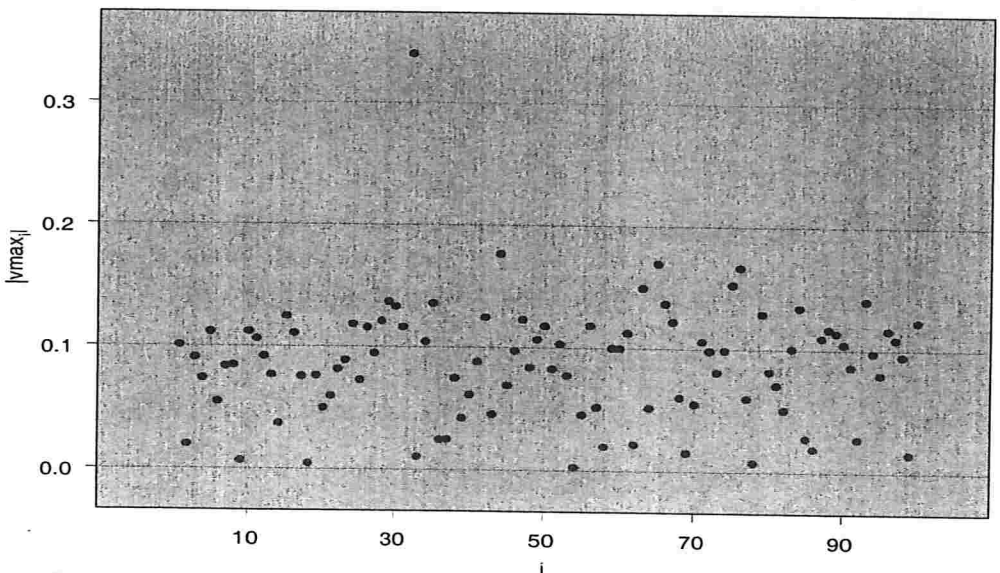

Figura 5.16: *Gráfico de influência para dados simulados* - 15% *de censura.*

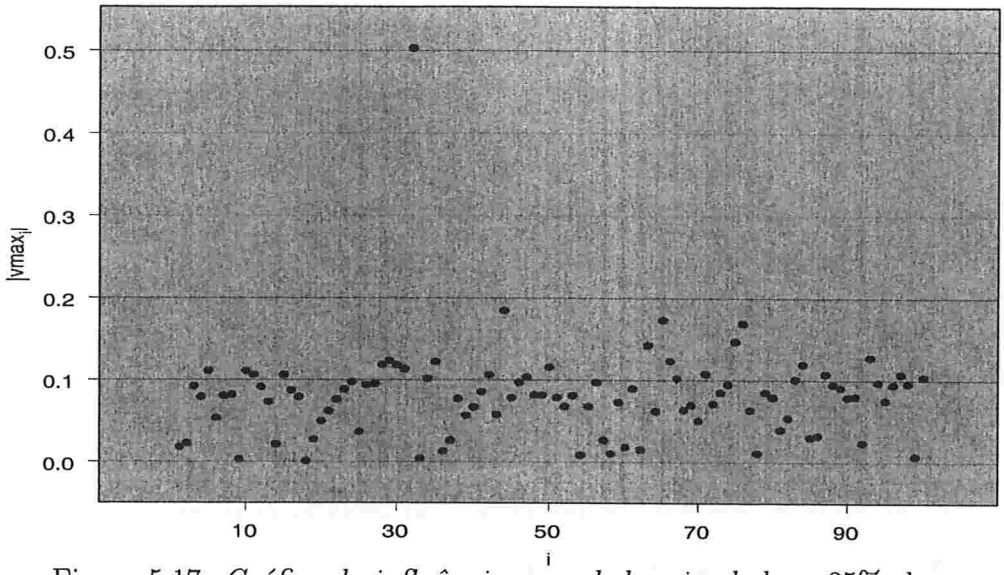

Figura 5.17: *Gráfico de influência para dados simulados - 25% de censura.*

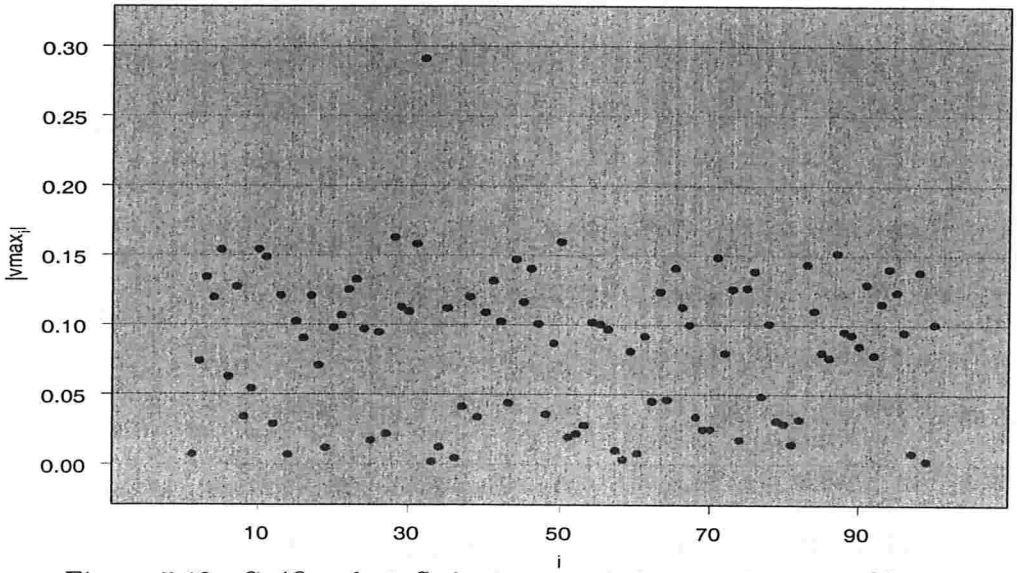

Figura 5.18: *Gráfico de influência para dados simulados - 50% de censura.*

Capítulo 6

Considerações Finais

6.1 Conclusões

Discutimos neste trabalho a aplicação da teoria de influência local (Cook 1986) em um modelo de sobrevivência com fração de cura, admitindo covariáveis observadas com e sem erro de medida.

Obtivemos as matrizes necessárias para a aplicação da técnica, considerando vários tipos de perturbação nos elementos dos dados e do modelo. Aplicando-se estes resultados em um conjunto de dados, podemos obter indicações de quais observações ou conjunto de observações influenciam de maneira sensível os resultados da análise. Ilustramos este fato no Capítulo 5, através de um conjunto de dados reais e de dados simulados. Para o conjunto de dados reais analisados, observamos que, para alguns esquemas de perturbação, a presença de algumas observações pode mudar consideravelmente os níveis de significância de certas covariáveis. Em alguns casos, a retirada de um pequeno número de observações censuradas acarretou em uma mudança expressiva nos resultados. Através de dados simulados com diferentes proporções de censura, pudemos observar a sensibilidade da metodologia de influência local, quando perturbamos o valor de uma covariável. Os resultados das aplicações indicam que o uso da técnica de influência local no modelo com

fração de cura pode ser bastante útil na detecção de possíveis pontos influentes.

Analisamos também a utilização do método do escore corrigido (Nakamura (1990); Gimenez & Bolfarine (1997)) para o processo de estimação no modelo com erro nas variáveis, através de dados simulados para vários tamanhos de amostra, proporção de censura e variância do erro de medida. Vimos que o estimador escore corrigido possibilita a obtenção de estimadores consistentes, ao mesmo tempo em que observamos a atenuação do estimador naive, o qual pode ser seriamente viciado.

6.2 Pesquisas Futuras

Propomos aqui possíveis pesquisas complementares que podem ser desenvolvidas com base neste trabalho.

1) Uma versão semi-paramétrica do modelo aqui analisado pode ser construída assumindo uma distribuição exponencial particionada para as variáveis que representam os riscos competitivos (Chen & Ibrahim 2001). Para tanto, consideramos uma partição finita do eixo temporal em M intervalos e assumimos que, para cada intervalo, temos associada uma função de taxa de falha constante e, portanto, uma distribuição exponencial para cada intervalo. Um estudo de influência local pode ser conduzido neste modelo, considerando-se diversos valores para M, que neste caso controla o grau de não-parametricidade do modelo.

2) Através da inclusão de um vetor de covariáveis $\mathbf{x}$ no modelo com fração de cura, estudamos aqui um modelo com função de sobrevivência

$$S_p(t|\mathbf{x}) = \exp\left(-\exp(\mathbf{x}'\boldsymbol{\beta})F(t)\right), \quad t \geq 0.$$

Vimos que desta forma, as covariáveis $\mathbf{x}$ ficam relacionadas à fração de curados e aos tempos de sobrevivência da população de não-curados. A inclusão de covariáveis pode ser feita de uma forma mais geral, admitindo que o vetor de covariáveis $\mathbf{x}$ de dimensão p possa ser particionado como $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, tal que $dim(\mathbf{x}_1) = r$ e $dim(\mathbf{x}_2) = s$, $p = r + s$, de modo que o subvetor $\mathbf{x}_1$ tenha relação somente com a fração de cura e que o subvetor $\mathbf{x}_2$ tenha relação tanto com a fração de curados quanto com os tempos de sobrevivência dos não-curados. Neste sentido, Asselain et al. (1996) propõem uma extensão da função de sobrevivência anterior, fazendo

$$S_p(t|\mathbf{x}) = \exp\left(-\theta \exp\left(\mathbf{x}_1'\boldsymbol{\alpha} + \mathbf{x}_2'\boldsymbol{\beta}^{(1)}\right) F\left(t\exp(-\mathbf{x}_2'\boldsymbol{\beta}^{(2)})\right)\right), \quad t \geq 0,$$

com $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)'$ vetor com os coeficientes de regressão associados a $\mathbf{x}_1$ e, $\boldsymbol{\beta}^{(1)} = \left(\beta_1^{(1)}, \ldots, \beta_s^{(1)}\right)'$ e $\boldsymbol{\beta}^{(2)} = \left(\beta_1^{(2)}, \ldots, \beta_s^{(2)}\right)'$ vetores com os coeficientes de regressão associados a $\mathbf{x}_2$ para a fração de curados e não-curados, respectivamente.

Podemos então neste modelo fazer um estudo de influência local considerando um esquema de perturbação com interesse somente nos parâmetros associados à fração de cura, o que pode ser de grande interesse ao pesquisador.

3) No modelo com erro nas variáveis descrito na Seção 3.3, observamos mediante dados simulados que o erro de medição atenua o estimador de máxima verossimilhança naive $\hat{\beta}_{naive}$. Propomos obter a expressão para tal fator de atenuação, ou seja, obter o fator K tal que $\hat{\beta}_{naive} \xrightarrow{P} K\beta$, com β o verdadeiro valor do parâmetro.

Apêndice A

Função de Verossimilhança

Vamos mostrar aqui a obtenção da função de verossimilhança (2.3).

Temos que $y_i = \min\{T_i, C_i\}$, com $T_i = \min\{R_{i0}, R_{i1}, \ldots, R_{i,N_i}\}$, sendo $N_i \sim Poisson(\theta)$.

Sejam f_T e g as funções densidade de probabilidade de T_i e C_i, respectivamente, e S_T e G as funções de sobrevivência de T_i e C_i, respectivamente, para $i = 1, \ldots, n$.

Então,

$$\begin{aligned} S_T(t|n_i) = P(T_i > t|N_i = n_i) &= P\left(\min\{R_{i0}, R_{i1}, \ldots, R_{i,n_i}\} > t\right) \\ &= P\left(R_{i0} > t\right) P\left(R_{i1} > t\right) \ldots P\left(R_{i,n_i} > t\right) \\ &= 1\ S(t|\boldsymbol{\lambda}) \ldots S(t|\boldsymbol{\lambda}) \\ &= S(t|\boldsymbol{\lambda})^{n_i} \end{aligned}$$

e, portanto,

$$f_T(t|n_i) = -\frac{dS_T(t|n_i)}{dt} = n_i f(t|\boldsymbol{\lambda}) S(t|\boldsymbol{\lambda})^{n_i - 1}.$$

Segue então que

$$\begin{aligned} P(y_i = t, \nu_i = 0 | N_i = n_i) &= P(C_i = t, T_i > C_i | N_i = n_i) \\ &= P(T_i > C_i | C_i = t, N_i = n_i) P(C_i = t) \\ &= S_T(t|n_i) g(t) \\ &= S(t|\boldsymbol{\lambda})^{n_i} g(t) \end{aligned}$$

e

$$\begin{aligned} P(y_i = t, \nu_i = 1 | N_i = n_i) &= P(T_i = t, T_i \leq C_i | N_i = n_i) \\ &= P(T_i \leq C_i | T_i = t, N_i = n_i) P(T_i = t | N_i = n_i) \\ &= G(t) f_T(t|n_i) \\ &= G(t) n_i f(t|\lambda) S(t|\lambda)^{n_i - 1}. \end{aligned}$$

Portanto, a distribuição de (y_i, ν_i) dado $N_i = n_i$, $i = 1, \ldots, n$, sob a suposição de censura não-informativa é

$$\begin{aligned} f(y_i, \nu_i | n_i) &= \left[S(y_i|\boldsymbol{\lambda})^{n_i}\right]^{1-\nu_i} \left[n_i f(y_i|\boldsymbol{\lambda}) S(y_i|\boldsymbol{\lambda})^{n_i - 1}\right]^{\nu_i} \\ &= S(y_i|\boldsymbol{\lambda})^{n_i - \nu_i} \left[n_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i}. \end{aligned} \tag{A.1}$$

Assim, podemos escrever a densidade conjunta como

$$
\begin{aligned}
f(\mathbf{y}, \boldsymbol{\nu}, \mathbf{n}) &= \prod_{i=1}^{n} f(y_i, \nu_i, n_i) \\
&= \prod_{i=1}^{n} f(y_i, \nu_i | n_i) f(n_i) \\
&= \prod_{i=1}^{n} S(y_i|\boldsymbol{\lambda})^{n_i - \nu_i} \left[n_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i} \frac{\theta^{n_i}}{n_i!} e^{-\theta} \\
&= \prod_{i=1}^{n} S(y_i|\boldsymbol{\lambda})^{n_i - \nu_i} \left[n_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i} \exp\left\{\sum_{i=1}^{n} \left[n_i \ln\theta - \ln(n_i!) - \theta\right]\right\}.
\end{aligned}
$$

Logo, a função de verossimilhança dos dados completos é dada por

$$
L(\theta, \boldsymbol{\lambda}; \mathbf{D_c}) = \prod_{i=1}^{n} S(y_i|\boldsymbol{\lambda})^{N_i - \nu_i} \left[N_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i} \exp\left\{\sum_{i=1}^{n} \left[N_i \ln\theta - \ln(N_i!) - \theta\right]\right\}.
$$

Introduzindo covariáveis no modelo através do parâmetro θ através da relação $\theta_i \equiv \theta(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, segue que

$$
L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \mathbf{D_c}) = \left\{\prod_{i=1}^{n} S(y_i|\boldsymbol{\lambda})^{N_i - \nu_i} \left[N_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i}\right\} \exp\left\{\sum_{i=1}^{n} \left[N_i x_i'\boldsymbol{\beta} - \ln(N_i!) - \exp(\mathbf{x_i'}\boldsymbol{\beta})\right]\right\}.
$$

Apêndice B

Distribuição Condicional de N

Descrevemos a seguir a obtenção da distribuição de N_i dado $\mathbf{D}$, $i = 1, \ldots, n$, (Seção 2.3).

Vimos anteriormente na equação (A.1) que

$$f(y_i, \nu_i | n_i) = S(y_i|\boldsymbol{\lambda})^{n_i - \nu_i} \left[n_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i}.$$

Então,

$$\begin{aligned}
f(y_i, \nu_i) &= \sum_{n_i=0}^{+\infty} f(y_i, \nu_i, n_i) \\
&= \sum_{n_i=0}^{+\infty} f(y_i, \nu_i | n_i) f(n_i) \\
&= \sum_{n_i=0}^{+\infty} S(y_i|\boldsymbol{\lambda})^{n_i - \nu_i} \left[n_i f(y_i|\boldsymbol{\lambda})\right]^{\nu_i} \frac{\theta^{n_i}}{n_i!} e^{-\theta} \\
&= \left\{ \sum_{n_i=0}^{+\infty} \frac{n_i^{\nu_i} \left[\theta S(y_i|\boldsymbol{\lambda})\right]^{n_i} e^{-\theta S(y_i|\boldsymbol{\lambda})}}{n_i!} \right\} S(y_i|\boldsymbol{\lambda})^{-\nu_i} e^{-\theta} e^{\theta S(y_i|\boldsymbol{\lambda})} f(y_i|\boldsymbol{\lambda})^{\nu_i} \\
&= \theta^{\nu_i} e^{-\theta} e^{\theta S(y_i|\boldsymbol{\lambda})} f(y_i|\boldsymbol{\lambda})^{\nu_i}.
\end{aligned}$$

Portanto, podemos escrever

$$\begin{aligned} f(n_i|y_i,\nu_i) &= \frac{f(y_i,\nu_i,n_i)}{f(y_i,\nu_i)} \\ &= \frac{n_i^{\nu_i} S(y_i|\boldsymbol{\lambda})^{n_i-\nu_i} f(y_i|\boldsymbol{\lambda})^{\nu_i} \theta^{n_i} e^{-\theta}}{n_i!\theta^{\nu_i} e^{-\theta} e^{\theta S(y_i|\lambda)} f(y_i|\boldsymbol{\lambda})^{\nu_i}} \\ &= \frac{\theta^{n_i-\nu_i} e^{-\theta S(y_i|\lambda)} S(y_i|\boldsymbol{\lambda})^{n_i-\nu_i}}{n_i! n_i^{-\nu_i}}. \end{aligned}$$

Mas,

$$\frac{n_i!}{n_i^{\nu_i}} = \begin{cases} n_i! & , \text{ se } \nu_i = 0 \\ (n_i-1)! & , \text{ se } \nu_i = 1 \end{cases}$$

e, portanto,

$$\frac{n_i!}{n_i^{\nu_i}} = (n_i - \nu_i)!.$$

Logo,

$$\begin{aligned} f(n_i|y_i,\nu_i) &= \frac{e^{-\theta S(y_i|\lambda)} \left[\theta S(y_i|\boldsymbol{\lambda})\right]^{n_i-\nu_i}}{(n_i-\nu_i)!} \\ &= \frac{e^{-S(y_i|\lambda)e^{\mathbf{x_i'}\boldsymbol{\beta}}} \left[S(y_i|\boldsymbol{\lambda})e^{\mathbf{x_i'}\boldsymbol{\beta}}\right]^{n_i-\nu_i}}{(n_i-\nu_i)!}. \end{aligned}$$

Da expressão acima, concluímos que $N_i \,|\, \mathbf{D}$ possui a mesma distribuição de $V_i + \nu_i$ com V_i variável Poisson tal que $E(V_i) = S(y_i|\boldsymbol{\lambda}) \exp(\mathbf{x_i'}\beta)$.

Referências Bibliográficas

Asselain, B., Fourque, A., Hoang, T., Tsodikov, A. & Yakovlev, A. (1996). A parametric regression model of tumor recurrence: An application to the analysis of clinical data on breast cancer, *Statistics and Probability Letters* **29**: 271–278.

Berkson, J. & Gage, R. (1952). Survival curve for cancer patients following treatment, *Journal of American Statistical Association* **47**: 501–515.

Carroll, R., Ruppert, D. & Stefanski, L. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall, London.

Chen, M.-H., Harrington, D. & Ibrahim, J. (2002). Bayesian cure rate models for malignant melanoma: A case study of ECOG Trial E1690, *Applied Statistics* **51**: 135–150.

Chen, M.-H. & Ibrahim, J. (2001). Maximum likelihood methods for cure rate models with missing covariates, *Biometrics* **57**: 43–52.

Chen, M.-H., Ibrahim, J. & Lipsitz, S. (2002). Bayesian methods for missing covariates in cure rate models, *Lifetime Data Analysis* **8**: 117–146.

Chen, M.-H., Ibrahim, J. & Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction, *Journal of American Statistical Association* **94**: 909–919.

Chen, M.-H., Ibrahim, J. & Sinha, D. (2002). Bayesian inference for multivariate survival data with a surviving fraction, *Journal of Multivariate Analysis* **80**: 101–126.

Cook, R. (1977). Detection of influential observation in linear regression, *Technometrics* **19**(1): 15–18.

Cook, R. (1986). Assessment of local influence (with discussion), *Journal of Royal Statistical Society, Ser. B* **48**(2): 133–169.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, Ser. B* **39**: 1–38.

Doornik, J. (1996). *Ox: An Object-Oriented Matrix Language*, International Thomson Business Press, London.

Escobar, L. & Meeker, W. (1992). Assessing local influence in regression analysis with censored data, *Biometrics* **48**: 507–528.

Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**.

Fuller, W. (1987). *Measurement Error Models*, John Wiley and Sons, New York.

Gimenez, P. (1997). *Inferência em Modelos com Erro nas Variáveis Através do Método do Escore Corrigido*, PhD thesis, Universidade de São Paulo.

Gimenez, P. & Bolfarine, H. (1997). Corrected score functions in classical error-in-variables and incidental parameter models, *Australian Journal of Statistics* **39**: 325–344.

Gimenez, P., Bolfarine, H. & Colosimo, E. (2000). Hypotheses testing for error-in-variables models, *Annals of the Institute of Statistical Mathematics* **52**: 698–711.

Goldman, A. (1984). Survivorship analysis when cure is a possibility: a Monte Carlo study, *Statistics in Medicine* **3**: 153–163.

Goldman, A. (1991). The cure model and time confounded risk in the analysis of survival and other timed events, *Journal of Clinical Epidemiology* **44**: 1327–1340.

Greenhouse, J. & Wolfe, R. (1984). A competing risks derivation of a mixture model for the analysis of survival data, *Communications in Statistics - Theory Meth.* **13**: 3133–3154.

Halpern, J. & Brown, B. (1987). Cure rate models: Power of the log-rank and gerneralized Wilcoxon tests, *Statistics in Medicine* **6**: 483–489.

Ibrahim, J., Chen, M.-H. & Sinha, D. (2001a). Bayesian semi-parametric models for survival data with a cure fraction, *Biometrics* **57**: 383–388.

Ibrahim, J., Chen, M.-H. & Sinha, D. (2001b). *Bayesian Survival Analysis*, Springer-Verlag, New York.

Kirkwood, J., Ibrahim, J., Sondak, V., Richards, J., Flaherty, L., Ernstoff, M., Smith, T., Rao, U., Steele, M. & Blum, R. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S9111/C9190, *Journal of Clinical Oncology* **18**: 2444–2458.

Kirkwood, J., Strawderman, M., Ernstoff, M., Smith, T., Borden, E. & Blum, R. (1996). Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: The Eastern Cooperative Oncology Group Trial EST 1684, *Journal of Clinical Oncology* **14**: 7–17.

Klebanov, L., Rachev, S. & Yakovlev, A. (1993). On the parametric estimation of survival functions, *Statistics and Decisions* **3**: 83–102.

Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York.

Maller, R. & Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*, John Wiley and Sons, New York.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models, *Biometrika* **77**: 127–137.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error, *Biometrics* **48**: 829–838.

Ortega, E. M. M., Bolfarine, H. & Paula, G. A. (2003). Influence diagnostics in generalized log-gamma regression models, *Computational Statistics and Data Analysis* **42**: 165–186.

Pettitt, A. & Bin Daud, I. (1989). Case-weighted measures of influence for proportional hazards regression, *Applied Statistics* **38**: 51–67.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**: 331–42.

Sposto, R., Sather, H. & Baker, S. (1992). A comparison of tests of the difference in the proportion of patients who are cured, *Biometrics* **48**: 87–99.

Stefanski, L. (1985). The effects of measurement error on parameter estimation, *Biometrika* **72**: 583–92.

Stefanski, L. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models, *Communications in Statistics - Theory Meth.* **18**: 4335–4358.

Stefanski, L. & Carroll, R. (1987). Conditional scores and optimal scores for generalized linear measurement- error models, *Biometrika* **74**: 703–716.

Tsodikov, A. (1998a). Asymptotic efficiency of a proportional hazards model with cure, *Statistics and Probability Letters* **39**: 237–244.

Tsodikov, A. (1998b). A proportional hazards model taking account of long-term survivors, *Biometrics* **54**: 138–146.

Tsodikov, A., Asselain, B., Fourque, A., Hoang, T. & Yakovlev, A. (1995). Discrete strategies of cancer posttreatment surveillance - estimation and optimization problems, *Biometrics* **51**: 437–447.

Tsodikov, A., Loeffler, M. & Yakovlev, A. (1998). A cure model with time-changing risk factor: an application to the analysis of secondary leukaemia. a report from the international database on hodgkin's disease, *Statistics in Medicine* **17**: 27–40.

Weissfeld, L. (1990). Influence diagnostics for the proportional hazard model, *Statistics and Probability Letters* **10**: 411–417.

Weissfeld, L. & Schneider, H. (1990). Influence diagnostics for the Weibull model fit to censored data, *Statistics and Probability Letters* **9**: 67–73.

Yakovlev, A., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefediere, A. & Tsodikov, A. (1993). A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer, *in* B. Asselain, M. Boniface, C. Duby, C. Lopez, J. Masson & J. Tranchefort (eds), *Biometrie et Analyse de Donnees Spatio-Temporelles*, number 12, pp. 66–82.

Yakovlev, A. & Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*, Singapore: World Scientific Publications.