**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

Modelos de regressão log-gama generalizado com fração de cura

Fernanda Bühner Rizzato

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba
Dezembro - 2006**

Fernanda Bühner Rizzato
Licenciada em Matemática

Modelos de regressão log-gama generalizado com fração de cura

Orientador:

Prof. Dr. **Edwin Moises Marcos Ortega**

Dissertação apresentada para obtenção do título de Mestre
em Agronomia. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba
Dezembro - 2006

Dados Internacionais de Catalogação na Publicação (CIP)
DI VI SÃO DE BI BLI OTECA E DOCUMENTAÇÃO - ESALQ/ USP

Rizzato, Fernanda Bühler

Modelos de regressão log-gama generalizado com fração de cura / Fernanda Bühler

Rizzato. - - Piracicaba, 2006.

73 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2007.

Bibliografia.

1. Análise de regressão e correlação 2. Análise de sobre vivência 3. Dados censurados
4. Verossimilhança I. Título

CDD 519.536

“Permi tida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

DEDICATÓRIA

*À Deus,
por me dar forças e guiar meus passos.*

*Aos meus pais,
José Orlando Rizzato e Suely Abdalla Bühner Rizzato,
pessoas fundamentais nesta conquista. Por todo o amor, o
carinho, a paciência e o apoio incondicional dedicados a mim.
Amo muito vocês!*

*À minha avó,
Jalile Abdalla Bühner (in memoriam),
o exemplo de vida e de força de vontade.*

Minha eterna gratidão...

AGRADECIMENTOS

Ao Prof. Dr. **Edwin Moisés Marcos Ortega** a orientação, a compreensão e a paciência nos seus ensinamentos para a realização desta conquista.

Ao conselho do programa de Pós-graduação em Estatística e Experimentação Agronômica, os professores Dra. **Clarice Garcia Borges Demétrio**, Dr. **Décio Barbin** e Dra. **Roseli Aparecida Leandro**, pelas valiosas sugestões e confiança.

Aos professores do Departamento de Ciências Exatas da ESALQ/USP, Dr. **Carlos Tadeu dos Santos Dias**, Dr. **César Gonçalves de Lima**, Dr. **Gerson Barreto Mourão**, Dr. **Silvio Sandoval Zocchi** e Dra. **Sônia Maria Stefano Piedade**, pela amizade e formação.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, as secretárias **Solange de Assis Paes Sabadin** e **Luciane Brajão** e aos técnicos em informática **Jorge Alexandre Wiendl** e **Eduardo Bonilha**, pelos auxílios permanentes.

Aos colegas de estudo do mestrado e do doutorado, pela amizade, companheirismo e paciência.

Um agradecimento especial aos amigos, **Alexandre**, **Ana Paula**, **Joseane**, **Juliana** e **Pâmela**, pelo incentivo, motivação, companheirismo, paciência e principalmente por estarem presentes em momentos importantes da minha vida.

Ao CNPq pela concessão de bolsa de estudo.

A todas as pessoas que contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

	Página
RESUMO	7
ABSTRACT	8
1 INTRODUÇÃO	9
2 REVISÃO BIBLIOGRÁFICA	12
2.1 Caracterização da distribuição de tempo de vida	12
2.2 Distribuição gama generalizada	13
2.2.1 Reparametrização	15
2.2.2 Características do modelo log-gama generalizado. Caso 1: considerando tempos de falha	19
2.2.3 Caso 2: considerando tempos de falha e censura	20
2.2.4 Caso 3: Modelo de regressão log-gama generalizado	23
2.3 Modelos de regressão log-gama generalizada com fração de cura	27
2.4 Influência local	29
2.4.1 Metodologia de influência local	29
2.5 Análise de resíduos e diagnósticos	33
2.5.1 Resíduos martingal	33
2.5.2 Resíduos deviance	34
3 METODOLOGIA	35
3.1 Material	35
3.2 Métodos	35
3.2.1 Distribuição log-gama generalizada com fração de cura	35
3.2.2 Modelo de regressão log-gama generalizada com fração de cura	38
3.2.3 Influência local	39

	6
3.2.3.1 Ponderação de casos	40
3.2.3.2 Perturbação da variável resposta	42
3.2.3.3 Perturbação de uma covariável	44
3.2.4 Análise de resíduos	47
3.2.4.1 Resíduos martingal	47
3.2.4.2 Resíduos deviance	48
3.2.5 Impacto das observações influentes	49
4 RESULTADOS E DISCUSSÃO	50
4.1 Resultados dos estimadores de máxima verossimilhança	50
4.2 Análise de influência local	52
4.2.1 Ponderação de casos	52
4.2.2 Perturbação da variável resposta	53
4.2.3 Perturbação de uma covariável	54
4.3 Análise de resíduos	56
4.4 Impacto das observações influentes	58
4.5 Qualidade de ajuste	59
5 CONSIDERAÇÕES FINAIS	61
5.1 Conclusões	61
5.2 Perspectivas para trabalhos futuros	61
REFERÊNCIAS	63
APÊNDICE	66
ANEXOS	71

RESUMO

Modelos de regressão log-gama generalizado com fração de cura

Neste trabalho considera-se uma reparametrização no modelo log-gama generalizado para a inclusão de dados com sobreviventes de longa duração. Os modelos tentam estimar separadamente os efeitos das covariáveis na aceleração ou desaceleração no tempo e na fração de sobreviventes que é a proporção da população para o qual o evento não ocorre. A função logística é usada para o modelo de regressão com fração de cura. Os parâmetros do modelo, serão estimados através do método de máxima verossimilhança. Alguns métodos de influência, como a influência local e a influência local total de um indivíduo, serão introduzidos, calculados, analisados e discutidos. Finalmente, um conjunto de dados médicos será analisado sob o modelo log-gama generalizado com fração de cura. Uma análise de resíduos será executada para verificar a qualidade de ajuste do modelo.

Palavras-chave: Modelo logístico; Modelo com fração de cura; Distribuição log-gama generalizada; análise de diagnóstico; dados censurados; dados de sobrevivência

ABSTRACT

The generalized log-gama mixture model with covariates

In this work the generalized log-gama model is modified for possibility that long-term survivors are present in the data . The models attempt to estimate separately the effects of covariates on the acceleration/deceleration of the timing of a given event and surviving fraction; that is, the proportion of the population for which the event never occurs. The logistic function is used for the regression model of the surviving fraction. Inference for the model parameters is considered via maximum likelihood. Some influence methods, such as the local influence, total local influence of an individual are derived, analyzed and discussed. Finally, a data set from the medical area is analyzed under log-gama generalized mixture model. A residual analysis is performed in order to select an appropriate model.

Keywords: Logistic model; Mixture models; log-gama generalized distribution; influence diagnostics; censored data; survival data.

1 INTRODUÇÃO

O modelo log-gama generalizado estendido proposto por Lawless (1980) tem sido largamente aplicado na área de Análise de Sobrevida e Confiabilidade. O modelo permite pesquisar diversos tipos de problemas para estudos analíticos bem como pode ser implementado em situações práticas; além disso, pacotes computacionais tais como SAS introduziram a função gama incompleta e a integral gama incompleta, o que têm contribuído enormemente para a popularidade do modelo no sentido de facilitar sua implementação computacional.

O presente trabalho apresenta o estudo do modelo log-gama generalizado com fração de cura, ou seja, dados em que a maioria dos indivíduos não apresenta o evento de interesse, mesmo sendo observados por um longo período de tempo. Do ponto de vista estatístico, pode haver problemas que podem ser verificados através do estimador produto-limite de Kaplan & Meier (função de sobrevivência empírica). Entretanto, na prática, esses dados são de grande interesse, principalmente na área médica, pois, se se considera um estudo sobre determinado tratamento para a cura do câncer, o interesse é que se tenha um grande número de indivíduos imunes (curados). Assim, será proposto o modelo de regressão, na presença de covariáveis, considerando a distribuição log-gama generalizada para dados que apresentam fração de cura e foi realizada a parte inferencial do modelo proposto. Um trabalho importante, que foi utilizado como motivação, foi o desenvolvido por Yu et al. (2004) onde utilizam a distribuição log-gama generalizada com fração de cura considerando dados agrupados, porém sem utilizar covaráveis.

Entretanto, quando se está ajustando um modelo a um conjunto de dados, é imprescindível que as estimativas obtidas a partir do modelo proposto sejam resistentes a pequenas perturbações nas observações ou no modelo. Se o modelo ajustado não apresentar uma boa descrição dos dados que foram observados, o mesmo pode conduzir a inferências errôneas.

Assim sendo, é importante que se faça um estudo sobre a robustez dos resultados obtidos, em termos dos vários aspectos que envolvem a formulação do modelo e as estimativas

dos seus parâmetros, ou seja, fazer uma análise de diagnóstico, que consiste de métodos para avaliar o grau de sensibilidade das inferências a pequenas perturbações nos dados ou mesmo no modelo proposto.

Dentre as medidas geralmente utilizadas para detectar influência de observações sobre o modelo ajustado, duas se destacam. Uma delas é a distância D_i , sugerida por Cook (1977), e a outra medida de influência é conhecida por $DFFITs_i$ que foi proposta por Belsley, Kuh e Welsch (1980), ambas as medidas utilizam a estratégia da exclusão de casos.

Uma abordagem mais recente é o estudo de influência local, um conceito introduzido por Cook (1986), que se baseia na análise das curvaturas das seções normais de uma determinada superfície numa vizinhança de um ponto especial, denominada gráfico de influência.

O método de influência local será aplicado neste trabalho no modelo de regressão log-gama generalizado com fração de cura. Segundo a proposta de Cook (1986), assume-se o afastamento pela verossimilhança para poder avaliar a influência local e total das observações nas estimativas dos parâmetros.

Também sabe-se que quando modelamos um conjunto de dados um importante passo na validação do modelo é a análise de resíduos, em que se procura detectar pontos em que o modelo não é adequado, bem como verificar se há indícios de afastamentos das suposições feitas para o erro do modelo.

Este trabalho está organizado da seguinte forma: na revisão de literatura é apresentada a distribuição gama generalizada e a seguir será introduzida uma reparametrização da mesma. Serão analisados três casos para o modelo log-gama generalizado: considerando tempos de falha, considerando tempos de falha e censura e considerando tempos de falha e censura na presença de covariáveis. Também será proposta a teoria de influência local bem como a de análise de resíduos. Em metodologia serão desenvolvidos os cálculos necessários para a realização dos procedimentos inferenciais no modelo de regressão log-gama generalizado com fração de cura considerando tempos de falha e censura e no modelo de regressão

log-gama generalizado com fração de cura. Também serão apresentados os cálculos de influência local, para os três casos a serem considerados: ponderação de casos, perturbação da variável resposta e perturbação de uma covariável. Para a análise de resíduos serão calculados os resíduos deviance e martingal, que são os resíduos mais utilizados na análise de sobrevivência. Em resultados e discussão será apresentada a aplicação a conjunto de dados reais de toda a teoria aqui desenvolvida, os parâmetros do modelo proposto serão estimados, as análises de influência local e de resíduos serão realizadas e o impacto das observações influentes e a qualidade de ajuste do modelo serão verificados. Finalmente, em considerações finais serão apresentadas algumas conclusões e alguns trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Nesta seção, apresenta-se um breve histórico acerca da distribuição gama generalizada e suas reparametrizações para facilitar a estimação de seus parâmetros.

2.1 Caracterização da distribuição de tempo de vida

Análise de sobrevivência é constituída por um conjunto de técnicas estatísticas para analisar dados de tempo até a ocorrência de um determinado evento de interesse, que, em geral, é denominado de tempo de sobrevivência. Por exemplo: tempo até a falha de um componente eletrônico, tempo de vida até a morte de um paciente por uma doença previamente diagnosticada; ou tempo até a cura de um paciente após tratamento adequado. É comum, em análise de sobrevivência, a observação parcial de algumas respostas, ou seja, observações que não apresentam o evento de interesse, denominados censurados. Uma censura pode ocorrer por várias razões, sendo as mais comuns:

- (i) O término do experimento antes da ocorrência do evento;
- (ii) O paciente ter se mudado para uma localidade desconhecida;
- (iii) O paciente ter abandonado o tratamento;
- (iv) A não adaptação do indivíduo ao tratamento e a necessidade de retirá-lo do estudo.

Censura de tipo I: os dados apresentam censura tipo I se o experimento for realizado em um período de tempo fixo, de tal maneira que o tempo de vida só será conhecido se ele for menor do que um limite pré-determinado, e então, o número de falhas é uma variável aleatória.

Censura de tipo II: os dados apresentam esse tipo de censura se o número de falhas a ser observado durante o experimento for fixado antes do seu início, e, portanto, o tempo de duração do experimento será uma variável aleatória.

Censura aleatória: esse é o tipo de censura mais comum, em que um indivíduo sai do experimento sem a ocorrência do evento de interesse, e, assim, tanto o número de falhas como o tempo de duração do experimento são variáveis aleatórias.

Agora, considere-se uma variável aleatória T contínua e positiva, que representa o tempo de sobrevivência ou instante de falha e, que associada a ela, exista uma função de densidade de probabilidade definida por:

$$f(x) = \lim_{\Delta t \rightarrow 0} \frac{P\{T \in [t, t + \Delta t)\}}{\Delta t} = \frac{\partial F(t)}{\partial t}$$

em que $F(t) = P(T \leq t)$ é a função distribuição acumulada de T . A função de densidade de probabilidade é estimada pela proporção de indivíduos que morreram em um intervalo de tempo Δt e possui duas propriedades:

$$f(t) \geq 0 \quad \text{e} \quad \int_0^{\infty} f(t) dt = 1.$$

A probabilidade de uma observação não falhar é representada pela função de sobrevivência dada por:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x) dx = 1 - F(t) \quad (1)$$

e que tem como propriedades: $\lim_{t \rightarrow 0} S(t) = 1$ e $\lim_{t \rightarrow \infty} S(t) = 0$.

A função de risco ou taxa de falha é definida como a razão instantânea de morte ou falha no tempo t , dado que o indivíduo sobreviveu até o tempo t , dada por:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P\{T \in [t, t + \Delta t) | T > t\}}{\Delta t} \quad (2)$$

e também pode ser escrita como:

$$\lambda(t) = \frac{f(t)}{S(t)} \geq 0.$$

2.2 Distribuição gama generalizada

A distribuição gama generalizada é de grande interesse na análise de dados de sobrevivência por representar uma família paramétrica. Entretanto, encontram-se dificuldades

na estimação de seus parâmetros, o que não ocorre quando se trabalha com a distribuição log-gama generalizada.

A seguir será apresentada a distribuição gama generalizada, bem como as suas reparametrizações e propriedades até a obtenção da distribuição log-gama generalizada. Serão considerados três esquemas diferentes: considerando apenas a presença de falha, considerando presença de falhas e censuras e, por último, considerando a presença de falhas, censuras e covariáveis.

Seja T uma variável aleatória com distribuição gama generalizada, cuja função de densidade é dada por:

$$f(t; \alpha, \beta, k) = \frac{\beta}{\alpha \Gamma(k)} \left(\frac{t}{\alpha}\right)^{\beta k - 1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\beta\right\}; \quad t, \alpha, \beta, k > 0 \quad (3)$$

com média e variância

$$E(T) = \frac{\alpha \Gamma(\frac{\beta k - 1}{\beta})}{\Gamma(k)} \quad \text{e} \quad V(T) = \frac{\alpha^2}{\Gamma(k)} \left[\Gamma\left(\frac{\beta k + 2}{\beta}\right) - \frac{(\Gamma(\frac{\beta k + 1}{\beta}))^2}{\Gamma(k)} \right].$$

Como casos particulares dessa família paramétrica podem ser citadas as distribuições: exponencial, $f(t; \alpha, 1, 1)$, gama, $f(t; \alpha, 1, k)$, Weibull, $f(t; \alpha, \beta, 1)$, qui-quadrado, $f(t; 2, n/2, 2)$, log-normal, quando $k \rightarrow \infty$

As funções de distribuição acumulada, de sobrevivência e de taxa de risco são dadas, respectivamente, por:

$$F(t) = P[T \leq t] = \int_0^t f(t; \alpha, \beta, k) dt = \frac{1}{\Gamma(k)} \int_0^{(\frac{t}{\alpha})^\beta} w^{k-1} e^{-w} dw = I\left[k, \left(\frac{t}{\alpha}\right)^\beta\right],$$

$$S(t) = 1 - F(t) = 1 - I\left[k, \left(\frac{t}{\alpha}\right)^\beta\right]$$

e

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{t^{\beta k - 1} e^{-(\frac{t}{\alpha})^\beta}}{\int_t^\infty x^{\beta k - 1} e^{-(\frac{x}{\alpha})^\beta} dx}.$$

em que

$$I(k, x) = \frac{1}{\Gamma(k)} \int_0^x u^{(k-1)} e^{-u} du$$

é a função gama incompleta, que é facilmente implementada na maioria dos pacotes estatísticos.

Dada $t_1, t_2, \dots, t_n$ uma amostra aleatória de uma distribuição gama generalizada com vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \beta, k)^T$; a estimação dos parâmetros pode ser feita pelo método de máxima verossimilhança, sendo o logaritmo da função de verossimilhança dado por:

$$l(\boldsymbol{\theta}) = n\{\log(\beta) - \log(\alpha) - \log[\Gamma(k)]\} + (\beta k - 1) \sum_{i=1}^n \log\left(\frac{t_i}{\alpha}\right) - \sum_{i=1}^n \left(\frac{t_i}{\alpha}\right)^\beta \quad (4)$$

Derivando-se $l(\boldsymbol{\theta})$ em relação a α , β e k e fazendo-se $\frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} = 0$, $\frac{\partial l(\boldsymbol{\theta})}{\partial \beta} = 0$ e $\frac{\partial l(\boldsymbol{\theta})}{\partial k} = 0$, têm-se

$$\begin{cases} -n\hat{k} + \sum_{i=1}^n \left(\frac{t_i}{\hat{\alpha}}\right)^{\hat{\beta}} = 0, \\ \frac{n}{\hat{\beta}} + \hat{k} \sum_{i=1}^n \log\left(\frac{t_i}{\hat{\alpha}}\right) - \sum_{i=1}^n \left(\frac{t_i}{\hat{\alpha}}\right)^{\hat{\beta}} \log\left(\frac{t_i}{\hat{\alpha}}\right) = 0 \text{ e} \\ -n\Psi(\hat{k}) + \hat{\beta} \sum_{i=1}^n \log\left(\frac{t_i}{\hat{\alpha}}\right) = 0 \end{cases}$$

em que $\Psi(\hat{k})$ é a função digama que é definida como $\Psi(k) = \frac{\partial \log \Gamma(k)}{\partial k} = \frac{\Gamma'(k)}{\Gamma(k)}$, em que $k > 0$.

Hager e Brain (1970) concluíram que o método de Newton-Raphson não leva a resultados satisfatórios e, para amostras de tamanho maior do que 200, os estimadores de máxima verossimilhança deixam de convergir. Assim, Prentice (1974) propôs uma reparametrização.

2.2.1 Reparametrização

Seja T uma variável aleatória com distribuição gama generalizada dada pela expressão (3). Prentice (1974) mostra que a variável T pode ser escrita da seguinte forma:

$$T = \alpha X^{\frac{1}{\beta}}$$

em que $X \sim \text{Gama}(1, k)$. Logo, tem-se que:

$$Y = \log T = \log(\alpha) + \frac{1}{\beta} \log X \sim \text{log-gama generalizada}(\alpha, \beta, k) \quad (5)$$

pois se $X \sim \text{Gama}(1, k)$ então $Z = \log X \sim \text{log-gama}(1, k)$ com

$$E(Z) = \mu^* = \Psi(k) \cong \log(k) - \frac{1}{2k} - \frac{1}{12k^2} + \frac{1}{120k^4} - \frac{1}{252k^6} + \dots \quad (6)$$

e

$$V(Z) = \sigma^{*2} = \Psi'(k) \cong \frac{1}{k} - \frac{1}{2k^2} + \frac{1}{6k^3} - \frac{1}{30k^5} + \frac{1}{42k^7} + \dots \quad (7)$$

Prentice (1974), ainda, considera um parâmetro de transformação $q = k^{-c}$, $c > 0$, pois quando $k \rightarrow \infty$ a distribuição da variável aleatória T tende para a distribuição log-normal. Define também, um ponto $q = 0$ onde a distribuição de Y aproxima-se da distribuição normal. Mostra que $c = \frac{1}{2}$ é o valor para o qual a distribuição de Y mais se aproxima da distribuição normal.

Considerando-se uma padronização da variável aleatória Z , isto é;

$$\varepsilon = \frac{Z - \mu^*}{q}.$$

em que $\sigma^{*2} = \Psi(k) \cong \frac{1}{k} = q^2$ é uma aproximação de primeira ordem para a variância, tem-se

$$Z = q\varepsilon + \mu^*. \quad (8)$$

Por outro lado, substituindo a expressão (8) na equação (5) tem-se que:

$$Y = \gamma + \sigma\varepsilon,$$

em que $\gamma = \log(\alpha) + \frac{\mu^*}{\beta}$ e $\sigma = \frac{q}{\beta}$. Assim, pode-se reescrever a variável aleatória Z como:

$$Z = q\left(\frac{Y - \gamma}{\sigma}\right) + \mu^*. \quad (9)$$

Portanto, a função de densidade de Y pode ser obtida como

$$f(y) = f_Z\left(\frac{y - \gamma}{\sigma}q + \mu^*\right)|J|,$$

em que $|J| = \frac{\partial z}{\partial y} = \left| \frac{q}{\sigma} \right|$ é o jacobiano da transformação. Logo,

$$f(y; \gamma, \sigma, q) = \frac{q}{\sigma \Gamma(q^{-2})} \exp \left\{ \left[\frac{(y - \gamma)}{\sigma} q + \mu^* \right] q^{-2} - \exp \left[\frac{(y - \gamma)}{\sigma} q + \mu^* \right] \right\},$$

em que: $y \in \mathbb{R}$, $\gamma > 0$, $\sigma > 0$ e $q > 0$.

Prentice (1974) estendeu esse modelo para valores de $q \in \mathbb{R}$.

$$f(y; \gamma, \sigma, q) = \begin{cases} \frac{|q|}{\sigma \Gamma(q^{-2})} \exp \left\{ \left[\frac{(y - \gamma)}{\sigma} q + \mu^* \right] q^{-2} + \exp \left[\frac{(y - \gamma)}{\sigma} q + \mu^* \right] \right\}, & \text{se } q \neq 0 \\ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (y - \mu^*)^2 \right], & \text{se } q = 0 \end{cases} \quad (10)$$

em que $\mu^* = \Psi(q^{-2})$.

Uma reparametrização alternativa foi proposta por Farewell e Prentice (1977), e por Lawless (1980), sendo que a média e a variância são aproximadas pelo primeiro termo das expressões (6) e (7) da seguinte maneira:

$$\mu^* = \Psi(k) \cong \log(k) = \log(q^{-2}) \quad \text{e} \quad \sigma^{*2} = \Psi'(k) \cong \frac{1}{k} = q^2. \quad (11)$$

Assim

$$W = \frac{Z - \log(q^{-2})}{q} \quad (12)$$

tem função de densidade

$$f(w; q) = \frac{1}{\Gamma(q^{-2})} (q^{-2})^{q^{-2} - \frac{1}{2}} \exp \left[q^{-1} w - q^{-2} \exp(qw) \right], \quad w \in \mathbb{R} \quad \text{e} \quad q > 0. \quad (13)$$

Por outro lado, substituindo a expressão (12) na equação (5), a variável Y pode ser escrita como

$$\begin{aligned} Y &= \log(\alpha) + \frac{1}{\beta} Z \\ &= \log(\alpha) + \frac{1}{\beta} \left[qW + \log(q^{-2}) \right] \\ &= \log(\alpha) + \frac{1}{\beta} \log(q^{-2}) + \frac{1}{\beta} qW \end{aligned}$$

ou, ainda,

$$Y = \mu + \sigma W, \quad (14)$$

em que $\mu = \log(\alpha) + \frac{1}{\beta} \log(q^{-2})$ e $\sigma = \frac{q}{\beta}$ e com função de densidade

$$f(y; \mu, \sigma, q) = \begin{cases} \frac{|q|}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[q \left(\frac{y-\mu}{\sigma} \right) - \exp \left\{ q \left(\frac{y-\mu}{\sigma} \right) \right\} \right] \right\}, & \text{se } q \neq 0 \\ \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \right\}, & \text{se } q = 0 \end{cases} \quad (15)$$

conhecido como modelo log-gama generalizado estendido.

A média e a variância de Y são dadas por:

(i) Para $q \neq 0$

$$E(Y) = \mu + \sigma \left[\frac{\Psi(q^{-2}) - \log(q^{-2})}{|q|} \right] \quad e \quad V(Y) = \frac{\sigma^2 \Psi'(q^{-2})}{q^2}$$

(ii) Para $q = 0$

$$E(Y) = \mu \quad e \quad V(Y) = \sigma^2$$

Pela Figura 1, pode-se ver que essa distribuição é simétrica para $q=0$ e assimétrica para $q=2$ e $q=-2$.

Figura 1 - Gráfico da função de densidade log-gama generalizada estendida para os valores do parâmetro $q = -2, 0, 2$

2.2.2 Características do modelo log-gama generalizado. Caso 1: considerando tempos de falha

Seja $y_1, y_2, \dots, y_n$ uma amostra aleatória com distribuição log-gama generalizada e vetor de parâmetros desconhecidos $\boldsymbol{\theta} = (\mu, \sigma, q)^T$. O logaritmo da função de verossimilhança para $q \neq 0$ é dado por:

$$l(\boldsymbol{\theta}) = n \left[\log|q| + \log(q^{-2})^{q^{-2}} - \log(\sigma \Gamma(q^{-2})) \right] + \sum_{i=1}^n \left\{ q^{-1} \left(\frac{y_i - \mu}{\sigma} \right) - q^{-2} \exp \left[q \left(\frac{y_i - \mu}{\sigma} \right) \right] \right\}.$$

Para se obterem os estimadores de máxima verossimilhança, basta resolver, simultaneamente, as equações:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \mu} = 0, \quad \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = 0 \quad \text{e} \quad \frac{\partial l(\boldsymbol{\theta})}{\partial q} = 0.$$

Entretanto, o método da máxima verossimilhança pode ter problemas de convergência e tem sido sugerido (Lawless, 2003) em seu lugar, o uso do método da máxima verossimilhança perfilada. Esse método consiste em atribuir uma grade de valores para q dentro de uma amplitude fixada $(-2, 2)$. Para cada valor fixado de q , obtêm-se os valores $\tilde{\mu}(q)$

e $\tilde{\sigma}(q)$ que maximizam $l(\mu(q), \sigma(q), q)$, fazendo-se $\frac{\partial l(\mu(q), \sigma(q), q)}{\partial \mu} = \frac{\partial l(\mu(q), \sigma(q), q)}{\partial \sigma} = 0$ sendo que

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \mu} = -\frac{n}{q\sigma} + \frac{1}{q\sigma} \sum_{i=1}^n \exp\left\{q\left(\frac{y_i - \mu}{\sigma}\right)\right\}$$

e

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = -\frac{n}{\sigma} - \sum_{i=1}^n \left(\frac{y_i - \mu}{q\sigma^2}\right) + \sum_{i=1}^n \left(\frac{y_i - \mu}{q\sigma^2}\right) \exp\left[q\left(\frac{y_i - \mu}{\sigma}\right)\right].$$

A seguir faz-se o gráfico de $l(\tilde{\mu}(q), \tilde{\sigma}(q), q)$ e as estimativas de μ , σ e q serão dadas pelos valores em que $l(\hat{\mu}(q), \hat{\sigma}(q), q)$ é máximo.

No caso de $q = 0$ têm-se os estimadores de máxima verossimilhança da distribuição normal:

$$\tilde{\mu}(0) = \bar{y} \quad \text{e} \quad \tilde{\sigma}(0) = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}.$$

2.2.3 Caso 2: considerando tempos de falha e censura

Seja $y_1, y_2, \dots, y_n$ uma amostra aleatória com distribuição log-gama generalizada com vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, q)^T$, em que $y_i = \log(t_i)$, $i = 1, 2, \dots, n$. A cada observação associa-se um indicador de censura δ_i , isto é,

$$\delta_i = \begin{cases} 1, & \text{se for um tempo de falha} \\ 0, & \text{se for um tempo de censura.} \end{cases}$$

Então, a função de verossimilhança considerando censura aleatória é dada por:

$$L(\boldsymbol{\theta}) = \prod_{i:\delta_i=1} f(y_i; \boldsymbol{\theta}) \prod_{i:\delta_i=0} S(y_i; \boldsymbol{\theta})$$

e seu logaritmo por:

$$l(\boldsymbol{\theta}) = \log[L(\boldsymbol{\theta})] = \sum_{i:\delta_i=1} \log[f(y_i; \boldsymbol{\theta})] + \sum_{i:\delta_i=0} \log[S(y_i; \boldsymbol{\theta})]. \quad (16)$$

A função de sobrevivência para a distribuição log-gama generalizada estendida é dada por:

(i) para $q \neq 0$

$$\begin{aligned}
 S(y; \boldsymbol{\theta}) &= 1 - F(y) = P(Y > y) \\
 &= P(\mu + \sigma W > y) \\
 &= P\left(W > \frac{y - \mu}{\sigma}\right) = P(W > w) \\
 &= \int_w^\infty \frac{|q|}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}u - q^{-2}\exp(qu)\right\} du.
 \end{aligned}$$

(ii) para $q = 0$

$$\begin{aligned}
 S(y; \boldsymbol{\theta}) &= 1 - F(y) = P(Y > y) \\
 &= P(\mu + \sigma W > y) \\
 &= P\left(W > \frac{y - \mu}{\sigma}\right) = P(W > w) \\
 &= 1 - \Phi(w)
 \end{aligned}$$

e, portanto,

$$S(y; \boldsymbol{\theta}) = \begin{cases} \int_w^\infty \frac{q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}u - q^{-2}\exp(qu)\right\} du, & \text{se } q > 0; \\ \int_w^\infty \frac{-q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}u - q^{-2}\exp(qu)\right\} du, & \text{se } q < 0; \\ 1 - \Phi(w), & \text{se } q = 0, \end{cases} \quad (17)$$

em que $w = \frac{y - \mu}{\sigma}$ e $\Phi(w)$ é a função de distribuição da normal padrão.

Se se considerar a mudança de variável $x = q^{-2}\exp(qu)$ na expressão (18) tem-se que:

(i) para $q > 0$

$$\begin{aligned}
 S(y; \boldsymbol{\theta}) &= \int_w^\infty \frac{q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}u - q^{-2}\exp(qu)\right\} du \\
 &= \frac{q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \int_w^\infty (e^{qu})^{q^{-2}} e^{-(q^{-2}\exp(qu))} du \\
 &= \frac{1}{\Gamma(q^{-2})} \int_{q^{-2}\exp(qw)}^\infty x^{q^{-2}-1} e^{-x} dx.
 \end{aligned}$$

(ii) para $q < 0$

$$\begin{aligned}
S(y; \boldsymbol{\theta}) &= \int_w^\infty \frac{-q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\{q^{-1}u - q^{-2}\exp(qu)\} du \\
&= \frac{-q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \int_w^\infty (e^{qu})^{q^{-2}} e^{-(q^{-2}\exp(qu))} du \\
&= \frac{-1}{\Gamma(q^{-2})} \left(- \int_0^{q^{-2}\exp(qw)} x^{q^{-2}-1} e^{-x} dx \right) \\
&= \frac{1}{\Gamma(q^{-2})} \int_0^{q^{-2}\exp(qw)} x^{q^{-2}-1} e^{-x} dx.
\end{aligned}$$

(iii) para $q = 0$

$$S(y; \boldsymbol{\theta}) = 1 - \Phi(w).$$

e, portanto,

$$S(y; \boldsymbol{\theta}) = \begin{cases} \frac{1}{\Gamma(q^{-2})} \int_{q^{-2}\exp(qw)}^\infty x^{q^{-2}-1} e^{-x} dx, & \text{se } q > 0; \\ \frac{1}{\Gamma(q^{-2})} \int_0^{q^{-2}\exp(qw)} x^{q^{-2}-1} e^{-x} dx, & \text{se } q < 0; \\ 1 - \Phi(w), & \text{se } q \neq 0. \end{cases} \quad (18)$$

Por outro lado, sabe-se que a função integral gama incompleta pode ser escrita como:

$$Q(k, a) = \int_a^\infty \frac{x^{k-1} e^{-x}}{\Gamma(k)} dx. \quad (19)$$

Agora, substituindo a expressão (19) em (17) e fazendo $w = \frac{y-\mu}{\sigma}$, tem-se:

$$S(y; \boldsymbol{\theta}) = \begin{cases} Q\left(q^{-2}, q^{-2}\exp\left[q\left(\frac{y-\mu}{\sigma}\right)\right]\right), & \text{se } q > 0; \\ 1 - Q\left(q^{-2}, q^{-2}\exp\left[q\left(\frac{y-\mu}{\sigma}\right)\right]\right), & \text{se } q < 0; \\ 1 - \Phi\left(\frac{y-\mu}{\sigma}\right), & \text{se } q \neq 0. \end{cases} \quad (20)$$

Desta maneira, substituindo a função de densidade (15) e a de sobrevivência

(20) na expressão (16), tem-se que o logaritmo da função de verossimilhança é dado por:

$$l(\boldsymbol{\theta}) = \begin{cases} \sum_{i:\delta_i=1} \log \left\{ \frac{q}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[\left(\frac{y_i - \mu}{\sigma} \right) q \right] - \exp \left(\frac{y_i - \mu}{\sigma} q \right) \right] \right\} \right\} \\ + \sum_{i:\delta_i=0} \log \left\{ Q \left(q^{-2}, q^{-2} \exp \left[q \left(\frac{y_i - \mu}{\sigma} \right) \right] \right) \right\}, & \text{se } q > 0; \\ \\ \sum_{i:\delta_i=1} \log \left\{ \frac{-q}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[\left(\frac{y_i - \mu}{\sigma} \right) q \right] - \exp \left(\frac{y_i - \mu}{\sigma} q \right) \right] \right\} \right\} \\ + \sum_{i:\delta_i=0} \log \left\{ 1 - Q \left(q^{-2}, q^{-2} \exp \left[q \left(\frac{y_i - \mu}{\sigma} \right) \right] \right) \right\}, & \text{se } q < 0; \\ \\ \sum_{i:\delta_i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (y_i - \mu)^2 \right] \\ + \sum_{i:\delta_i=0} \log \left\{ 1 - \Phi \left(\frac{y_i - \mu}{\sigma} \right) \right\}, & \text{se } q \neq 0. \end{cases} \quad (21)$$

Assim, como no modelo em que se consideram apenas falhas, também será utilizado o método da máxima verossimilhança perfilada para a obtenção das estimativas dos parâmetros.

2.2.4 Caso 3: Modelo de regressão log-gama generalizado

Sabe-se que o tempo de vida pode ser influenciado por variáveis regressoras ou covariáveis, que podem ser: idade, sexo, estágio da doença no diagnóstico etc, sendo de grande interesse prático o modelo de regressão log-gama generalizado estendido.

A classe de modelos de locação e escala se caracteriza pelo fato de $Y = \log(T)$ ter uma distribuição com parâmetros de locação $\mu(\mathbf{x})$ dependendo das variáveis regressoras, e um parâmetro de escala σ constante. Pode-se então escrever

$$Y = \mu(\mathbf{x}) + \sigma W, \quad (22)$$

em que $\sigma > 0$ e W tem uma distribuição que não depende de $\mathbf{x}$. Agora, por outro lado, considere-se um modelo de regressão em que se assume uma relação linear entre o logaritmo do tempo de vida, denotado por Y e um vetor de covariáveis $\mathbf{x} = (x_1, x_2, \dots, x_p)$. Além disso, suponha que Y dado $\mathbf{x}$ tem distribuição log-gama generalizada e pode ser representada na

forma:

$$Y = \mathbf{x}^T \boldsymbol{\beta} + \sigma Z, \quad (23)$$

em que $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$ e $\sigma > 0$ são os parâmetros desconhecidos.

Sejam $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)$ n observações provenientes de uma distribuição log-gama generalizada, dada em (15), em que y_i representa o logaritmo do tempo de vida ou tempo até a censura e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})^T$ o vetor de covariáveis associadas.

Utilizando a expressão (23) o modelo pode ser escrito como

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma Z_i, \quad i = 1, 2, \dots, n, \quad (24)$$

em que o vetor de parâmetros é $\boldsymbol{\theta} = (\sigma, q, \boldsymbol{\beta}^T)^T$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$.

A distribuição da variável aleatória Z pode ser obtida utilizando a transformação de variáveis que neste caso é dada por

$$f(z; \boldsymbol{\theta}) = \begin{cases} \frac{|q|}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\{q^{-1}z - q^{-2} \exp(qz)\}, & \text{se } q \neq 0 \\ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), & \text{se } q = 0. \end{cases} \quad (25)$$

E a função de densidade da variável Y pode ser escrita substituindo $\mu = \mathbf{x}_i^T \boldsymbol{\beta}$ na expressão (15).

Assim, a função de sobrevivência pode ser escrita como

(i) para $q \neq 0$

$$\begin{aligned} S(y; \boldsymbol{\theta}) &= 1 - F_{Y|x}(y) = P(Y > y) \\ &= P(\mathbf{x}^T \boldsymbol{\beta} + \sigma Z > y) \\ &= P\left(Z > \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right) \\ &= \int_{\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}}^{\infty} \frac{|q|}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\{q^{-1}z - q^{-2} \exp(qz)\} dz \end{aligned}$$

(ii) para $q = 0$

$$S(y; \boldsymbol{\theta}) = 1 - F_{Y|x}(y) = P(Y > y)$$

$$\begin{aligned}
&= P(\mathbf{x}^T \boldsymbol{\beta} + \sigma Z > y) \\
&= P\left(Z > \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right) \\
&= \int_{\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \\
&= 1 - \Phi\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right),
\end{aligned}$$

em que $z = \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}$ e $\Phi\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right)$ é a função de distribuição da normal padrão.

Utilizando a mudança de variável $t = q^{-2} \exp(qz)$, tem-se que:

(i) para $q > 0$

$$\begin{aligned}
S(y; \boldsymbol{\theta}) &= \int_{\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}}^{\infty} \frac{q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}z - q^{-2} \exp(qz)\right\} dz \\
&= \frac{q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \int_{\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}}^{\infty} (e^{qz})^{q^{-2}} e^{-(q^{-2} \exp(qz))} dz \\
&= \frac{1}{\Gamma(q^{-2})} \int_{q^{-2} \exp(q \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma})}^{\infty} t^{q^{-2}-1} e^{-t} dt
\end{aligned}$$

(ii) para $q < 0$

$$\begin{aligned}
S(y; \boldsymbol{\theta}) &= \int_{\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}}^{\infty} \frac{-q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}z - q^{-2} \exp(qz)\right\} dz \\
&= \frac{-q}{\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \int_{\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}}^{\infty} (e^{qz})^{q^{-2}} e^{-(q^{-2} \exp(qz))} dz \\
&= \frac{-1}{\Gamma(q^{-2})} \left(- \int_0^{q^{-2} \exp(q \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma})} t^{q^{-2}-1} e^{-t} dt \right) \\
&= \frac{1}{\Gamma(q^{-2})} \int_0^{q^{-2} \exp(q \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma})} t^{q^{-2}-1} e^{-t} dt
\end{aligned}$$

(iii) para $q = 0$

$$S(y; \boldsymbol{\theta}) = 1 - \Phi\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right).$$

Dessa maneira a função de sobrevivência pode ser expressa como

$$S(y; \boldsymbol{\theta}) = \begin{cases} \frac{1}{\Gamma(q^{-2})} \int_{q^{-2} \exp(q \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma})}^{\infty} t^{q^{-2}-1} e^{-t} dt, & \text{se } q > 0; \\ \frac{1}{\Gamma(q^{-2})} \int_0^{q^{-2} \exp(q \frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma})} t^{q^{-2}-1} e^{-t} dt, & \text{se } q < 0; \\ 1 - \Phi\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right), & \text{se } q = 0. \end{cases} \quad (26)$$

Finalmente, utilizando a função integral gama incompleta dada na expressão (20) e a expressão (28), a função de sobrevivência, considerando um modelo de regressão, pode ser escrita como

$$S(y; \boldsymbol{\theta}) = \begin{cases} Q\left(q^{-2}, q^{-2} \exp\left[q\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right)\right]\right), & \text{se } q > 0; \\ 1 - Q\left(q^{-2}, q^{-2} \exp\left[q\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right)\right]\right), & \text{se } q < 0; \\ 1 - \Phi\left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right), & \text{se } q = 0, \end{cases} \quad (27)$$

em que $\Phi(\cdot)$ é a função de distribuição da normal padrão. Nesse caso, o logaritmo da função de verossimilhança pode ser escrito da seguinte forma:

$$l(\boldsymbol{\theta}) = \begin{cases} \sum_{i:\delta_i=1} \log\left[\frac{q}{\sigma\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}z_i - q^{-2}\exp(qz_i)\right\}\right] \\ + \sum_{i:\delta_i=0} \log\left[Q(q^{-2}, q^{-2}\exp\{qz_i\})\right], & \text{se } q > 0; \\ \sum_{i:\delta_i=1} \log\left[\frac{-q}{\sigma\Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp\left\{q^{-1}z_i - q^{-2}\exp(qz_i)\right\}\right] \\ + \sum_{i:\delta_i=0} \log\left[1 - Q(q^{-2}, q^{-2}\exp\{qz_i\})\right], & \text{se } q < 0; \\ \sum_{i:\delta_i=1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}z_i^2\right\} + \sum_{i:\delta_i=0} \log\left[1 - \Phi(z_i)\right], & \text{se } q = 0 \end{cases} \quad (28)$$

em que $z_i = \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}$. O vetor de parâmetros para este modelo, mais geral, é dado por $\boldsymbol{\theta}^T = (\sigma, q, \boldsymbol{\beta}^T)$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$.

A dificuldade na estimação dos parâmetros é ainda maior, pois têm-se $(p + 3)$ parâmetros a serem estimados e, então, usa-se o método da máxima verossimilhança perfilada. O desenvolvimento da análise inferencial, assim como influência local, leverage generalizado e resíduos são tratados por Ortega (2001).

2.3 Modelos de regressão log-gama generalizada com fração de cura

Em estudos de tempos de sobrevivência, assume-se que em algum momento o evento de interesse irá ocorrer para todos os indivíduos observados, quando se considera um período de acompanhamento suficientemente grande. Mas, em certos experimentos, encontra-se uma determinada parcela da população em que o evento de interesse não ocorre.

Em estudos de ensaios clínicos, para algumas doenças como o câncer, muitos pesquisadores consideram que um paciente está curado se não houver uma reincidência da doença num período de 5 a 10 anos após o tratamento, sendo este tempo determinado pela experiência do pesquisador. Entretanto, um de seus grandes interesses é a proporção de curados para uma possível indicação de qual tratamento seguir.

Pode-se identificar a existência de uma fração de sobreviventes (fração de cura) para um determinado conjunto de dados através do estimador produto-limite de Kaplan-Meier, que é um gráfico da função de sobrevivência empírica. Se o gráfico apresentar uma cauda que tende para um valor diferente de zero, tem-se a indicação da existência da fração de indivíduos curados (ou indivíduos de longa duração).

Como exemplo, considera-se um conjunto de observações referente a um ensaio clínico sobre reincidência de melanoma maligno (Ibrahim, 2001).

Como se pode observar na Figura 2, a função de sobrevivência empírica, para esses dados, não tende para 0 quanto t tende para infinito, e sim para uma certa probabilidade próxima de 0,5 que representa a probabilidade de cura para esses dados.

Figura 2 - Gráfico de Sobrevivência de Kaplan-Meier

O problema da fração de cura pode ser abordado de duas maneiras. Uma introduzida por Berkson e Gage (1952) e Maller e Zhou (1996) que propuseram a construção de uma função de sobrevivência populacional imprópria na forma de mistura, e a outra introduzida por Yakovlev et al. (1993) que propuseram uma nova classe de modelos de mistura com a estrutura de riscos competitivos, neste trabalho será abordada a primeira opção.

Para a modelagem desse tipo de dados de sobrevivência, consideram-se duas populações, uma formada por indivíduos de longa duração e outra formada por indivíduos que apresentam o evento de interesse. Se um indivíduo está em risco, então associa-se uma função de sobrevivência própria $S(t)$ com probabilidade $(1 - p)$; entretanto, se um indivíduo é de longa duração (curado), ele apresentará uma função de sobrevivência imprópria $G(t)$, na forma de mistura com probabilidade p , $p \in (0, 1)$. Isto é,

$$G(t) = p + (1 - p)S(t) \quad (29)$$

e que possui duas propriedades interessantes

$$\lim_{t \rightarrow \infty} G(t) = 1 \quad e \quad \lim_{t \rightarrow 0} G(t) = p$$

Por outro lado, seja b_i uma variável binária latente que indica se o i -ésimo indivíduo está, ou

não, em risco em relação ao evento de interesse, isto é,

$$b_i = \begin{cases} 1, & \text{se o indivíduo estiver em risco;} \\ 0, & \text{se o indivíduo estiver imune.} \end{cases}$$

Mas, se se considerarem que essas variáveis b_i são independentes e identicamente distribuídas, a probabilidade de um indivíduo estar em risco é $P(b_i = 1) = 1 - p_i$ e a probabilidade de um indivíduo estar imune é $P(b_i = 0) = p_i$.

Em metodologia será proposto, unindo as duas metodologias, o modelo de regressão log-gama generalizado com fração de cura.

2.4 Influência local

Ao se ajustar um modelo a um conjunto de dados deseja-se que as estimativas obtidas sejam resistentes a pequenas perturbações no modelo ou nas observações. Enquanto uma análise de resíduos tem como objetivo pesquisar problemas no modelo ajustado, como a presença de observações aberrantes e afastamentos sérios das suposições para a distribuição do erro, numa análise de influência assume-se o modelo como correto e estuda-se a robustez das conclusões a perturbações nos dados ou no modelo. Uma observação se diz influente quando produz alterações desproporcionais nos resultados da análise ao ser omitida no ajuste do modelo ou submetida a uma pequena perturbação.

A técnica de influência local, desenvolvida por Cook (1986) é uma das técnicas mais utilizadas atualmente dentro do contexto de metodologias para análise de diagnóstico. Considerando-se o modelo postulado como o modelo correto, compara-se as estimativas obtidas através desse modelo com as estimativas decorrentes de uma pequena perturbação.

2.4.1 Metodologia de influência local

Dado um conjunto de observações, seja $l(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança correspondente ao modelo postulado, onde $\boldsymbol{\theta}$ é um vetor da ordem $(p + 3) \times 1$ de parâmetros desconhecidos. Perturbações podem ser introduzidas no modelo através de um

vetor $\mathbf{w}^T = (w_1, w_2, \dots, w_m)$ pertencente a um subconjunto aberto Ω de $\mathbb{R}^m$. Geralmente, $\mathbf{w}$ pode refletir qualquer esquema de perturbação bem definida, por exemplo, $\mathbf{w}$ pode ser usado para introduzir uma menor modificação nas variáveis exliativas ou para perturbar a matriz de covariância nos erros, no modelo de regressão linear (GALEA; PAULA; BOLFARINE, 1997).

Supondo definido o esquema de perturbação, considera-se $l(\boldsymbol{\theta}|\mathbf{w})$ o logaritmo da função de verossimilhança perturbada. Será assumido que $l(\boldsymbol{\theta}|\mathbf{w})$ é duas vezes diferenciável em $(\boldsymbol{\theta}^T, \mathbf{w}^T)^T$ e que o modelo postulado está encaixado no modelo perturbado. Isto é, supõem-se que existe $\mathbf{w}_0 \in \Omega$ tal que $l(\boldsymbol{\theta}|\mathbf{w}_0) = l(\boldsymbol{\theta})$, para todo $\boldsymbol{\theta} \in \mathbb{R}^m$. Em geral, a dimensão m (do vetor de perturbação) está relacionada com a dimensão do vetor $\boldsymbol{\theta}$ ou com o tamanho da amostra, dependendo do tipo de perturbação.

No modelo perturbado, o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ será denotado por $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ e o objetivo agora é comparar $\hat{\boldsymbol{\theta}}$ com $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ quando $\mathbf{w}$ varia em Ω . Se a distância entre eles permanecer pequena quando $\mathbf{w}$ varia em Ω , têm-se uma indicação de que existe uma estabilidade no modelo ajustado no que diz respeito ao particular esquema de perturbação que foi utilizado, e portanto, ao correspondente aspecto da análise que esta sendo monitorado. Uma comparação direta entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ pode não ser simples, devido a diversos fatores tais como a diferença de escala, unidade de medida, erro de estimativa, etc.

Cook (1986) sugere que a comparação entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ seja feita através da função

$$LD(\mathbf{w}) = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\mathbf{w}})], \mathbf{w} \in \Omega$$

que é uma generalização do afastamento pela verossimilhança $LD_i = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{-i})]$ que é muito utilizada na análise de influência global. Desta forma, o sentido de distância entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\mathbf{w}}$ passa a depender da concavidade do logaritmo da função de verossimilhança.

Por outro lado, como $\hat{\boldsymbol{\theta}}$ é o estimador de máxima verossimilhança para $\boldsymbol{\theta}$ no modelo sem perturbação, segue-se que $LD(\mathbf{w}) \geq 0$, para todo $\mathbf{w} \in \Omega$. Além disso, como $LD(\mathbf{w}_0) = 0$, pode-se concluir que $\mathbf{w}_0$ é um ponto de mínimo local da função afastamento pela verossimilhança. Uma análise sobre o comportamento geométrico da função $LD(\mathbf{w})$ quando

$\mathbf{w}$ varia em Ω pode fornecer informações bastante relevantes a respeito de características importantes do modelo sob investigação.

Assim, a idéia de Cook (1986) consiste em estudar o comportamento da função $LD(\mathbf{w})$ numa vizinhança $\mathbf{w}_0$, que é o ponto em que as duas verossimilhanças são iguais. Para isso, considerou a seguinte superfície geométrica:

$$\alpha(\mathbf{w}) = \begin{pmatrix} \mathbf{w} \\ LD(\mathbf{w}) \end{pmatrix}$$

denominada gráfico de influência. A idéia principal é analisar como a superfície $\alpha(\mathbf{w})$ desvia-se de seu plano tangente em $\mathbf{w}_0$, ou seja, analisar o comportamento da função $LD(\mathbf{w})$ em $\mathbf{w}_0$. O procedimento consiste em selecionar um vetor unitário $\boldsymbol{\ell}$ e, então, considerar o gráfico de $LD(\mathbf{w}_0 + a\boldsymbol{\ell})$ em função de a , $a \in \mathbb{R}$. Esse gráfico é chamado de linha projetada. Desde que $LD(\mathbf{w}_0)$, $LD(\mathbf{w}_0 + a\boldsymbol{\ell})$ tem um mínimo local em $a = 0$. Cada linha projetada pode ser caracterizada pela curvatura normal $C_{\boldsymbol{\ell}}$ em torno de $a = 0$. Sugere-se considerar a direção $\boldsymbol{\ell}_{max}$ correspondente à maior curvatura $C_{\boldsymbol{\ell}_{max}}$. O gráfico das componentes de $\boldsymbol{\ell}_{max}$ revela os elementos que, sob pequenas perturbações, exercem notável influência sobre $LD(\mathbf{w})$.

Cook (1986) mostra que a curvatura normal na direção $\boldsymbol{\ell}$ pode expressa da seguinte forma:

$$C_{\boldsymbol{\ell}}(\boldsymbol{\theta}) = 2|\boldsymbol{\ell}^T \boldsymbol{\Delta}^T \ddot{\mathbf{L}}_{(\boldsymbol{\theta})}^{-1} \boldsymbol{\Delta} \boldsymbol{\ell}|,$$

em que $\ddot{\mathbf{L}}_{(\boldsymbol{\theta})}$ é a matriz de informação observada sob o modelo postulado e $\boldsymbol{\Delta}$ é a matriz $(p+3) \times n$ sendo

$$\boldsymbol{\Delta} = \{\Delta_{ij}\}_{i=1,\dots,n} \quad j=1,\dots,p+3,$$

com

$$\Delta_{ij} = \left. \frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{w})}{\partial \theta_j \partial w_i} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}} \quad \mathbf{w}=\mathbf{w}_0}$$

Assim, pode-se avaliar a influência que o esquema de perturbações considerado exerce sobre os componentes do modelo, tais como estimativas dos parâmetros e outros resultados da análise estatística. Segundo Cook (1986), a direção que produz maior mudança

local na estimativa dos parâmetros é dada por ℓ_{max} , que corresponde ao autovetor associado ao maior autovalor de $-\Delta^T \ddot{L}_{(\theta)}^{-1} \Delta$. O vetor ℓ_{max} é utilizado para identificar as observações que podem estar controlando propriedades importantes na análise dos dados.

Contudo seja a partição do vetor $\theta = (\gamma^T, \beta^T)$ onde $\gamma = (\mu, \sigma)^T$. Supondo que o interesse agora seja somente no vetor β , a curvatura normal na direção ℓ é dado por

$$C_{\ell}(\theta) = 2|\ell^T \Delta^T (\ddot{L}_{(\theta)}^{-1} - L_1) \Delta \ell|,$$

em que

$$L_1 = \begin{bmatrix} \ddot{L}_{\gamma\gamma} & 0 \\ 0 & 0 \end{bmatrix}$$

sendo $\ddot{L}_{\gamma\gamma}$ a matriz de informação observada para γ .

A influência local das observações em $\hat{\beta}$ pode ser avaliada considerando-se o gráfico de índices de ℓ_{max} para a matriz $|\Delta^T (\ddot{L}_{(\theta)}^{-1} - L_1) \Delta|$.

Por outro lado, Lesaffre e Verbeke (1998) sugerem avaliar a curvatura normal na direção da i -ésima observação que consiste na avaliação de $C_{\ell}(\theta)$ no vetor ℓ_i de dimensão $(n \times 1)$ formado por zeros com um na i -ésima posição.

Nesse caso, a curvatura normal chamada influência local total do i -ésimo indivíduo é dado por

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

no caso de se estar interessado em θ e

$$C_i(\beta) = 2|\Delta_i^T (\ddot{L}_{(\theta)}^{-1} - L_1) \Delta_i|$$

no caso de se estar interessado em β .

É possível perturbar o modelo proposto de diversas maneiras, porém, é importante escolher esquemas de perturbação e medidas de influência que permitam interpretações fáceis. Neste trabalho, em metodologia, serão utilizados três esquemas de perturbação (casos, resposta e covariáveis).

Alguns autores têm investigado influência local em análise de sobrevivência, dentre eles, Pettitt e Bin Daud (1989) têm investigado influência local em modelos de riscos proporcionais, Escobar e Meeker (1992) têm adaptado métodos de influência local em análises de regressão com dados censurados, Ortega et al. (2003) têm considerado o problema de avaliar influência local nos modelos de regressão log-gama generalizado com observações censuradas, Galea et al. (2004) descreve o cálculo das curvaturas para diferentes esquemas de perturbação no modelo Log-Birnbaum-Saunders e Ortega et al. (2006) têm aplicado influência local nos modelos de regressão Weibull-exponenciada.

2.5 Análise de resíduos e diagnósticos

Quando se está ajustando um modelo a um conjunto de dados, faz-se necessário avaliar a qualidade desse ajuste. Assim, ao se propor um modelo e realizar o ajuste, a análise de resíduos tem como uma de suas finalidades verificar eficientemente a qualidade desse ajuste. Além disso, possibilita a detecção de observações discrepantes (pontos aberrantes), a determinação da forma funcional de uma covariável e a verificação das pressuposições assumidas para o erro. A seguir, serão descritos dois resíduos propostos na literatura, o resíduo martingal e o resíduo deviance, que fornecem uma forma de verificar a adequação do modelo ajustado e auxiliam na detecção de observações atípicas.

2.5.1 Resíduos martingal

Os resíduos martingal nos modelos de regressão paramétricos são definidos por:

$$r_{M_i} = \delta_i - \hat{H}(t_i|x_i) \quad (30)$$

em que δ_i é a variável indicadora da falha e $\hat{H}(t_i|x_i)$ é a função de risco acumulada obtida do modelo ajustado.

Pode-se considerar que esses resíduos são uma estimativa do número de falhas em excesso observada nos dados mas não predito pelo modelo. Em geral, são utilizados para examinar a melhor forma funcional (linear, quadrática, etc) para uma dada covariável em um

modelo de regressão assumido para os dados sob estudo (COLOSSIMO; GIOLO, 2006).

2.5.2 Resíduos deviance

Nos modelos de regressão paramétricos os resíduos deviance podem ser escritos da seguinte forma:

$$r_{D_i} = \text{sinal}(r_{M_i}) \left[-2 \left(r_{M_i} + \delta_i \log(\delta_i - r_{M_i}) \right) \right]^{\frac{1}{2}} \quad (31)$$

em que δ_i é a variável indicadora da falha e $\hat{m}_i$ é o resíduo martingal.

Esses resíduos são uma tentativa de tornar os resíduos martingal mais simétricos em torno de zero, assim, se o modelo for apropriado, eles devem apresentar um comportamento aleatório em torno de zero (COLOSSIMO; GIOLO, 2006).

A metodologia de influência local assim como a de resíduos serão estendidas para o modelo de regressão log-gama generalizado com fração de cura.

3 METODOLOGIA

3.1 Material

O conjunto de dados obtido por Ibrahim et al. (2001) a ser utilizado neste trabalho faz parte de um ensaio sobre melanoma cutâneo (tipo de câncer maligno), para a avaliação do desempenho de tratamento pós-cirúrgico, com alta dose de determinada droga (interferon alfa-2b) a fim de prevenir a reincidência. O período de entrada dos pacientes no estudo foi de 1991 até 1995 e o acompanhamento foi realizado até 1998 e a variável resposta T representa o tempo até a morte do paciente. O tamanho original da amostra foi de $n = 417$ pacientes. A porcentagem de observações censuradas foi de 56%. Para cada paciente $i, i = 1, 2, \dots, n$, têm-se associadas as variáveis: t_i , tempo observado (em anos), δ_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=outra) e x_{i6} , espessura do tumor, em mm.

3.2 Métodos

O modelo de regressão log-gama generalizado com fração de cura que será apresentado a seguir como fonte inicial para poder modelar o conjunto de dados reais. Para estimação dos parâmetros será utilizado o método de máxima verossimilhança que exigirá algum procedimento numérico. Esses métodos serão implementados no pacote estatístico SAS utilizando o procedimento NLMIXED. O desenvolvimento da parte inferencial assim como a análise de resíduos e qualidade de ajuste também serão realizados.

3.2.1 Distribuição log-gama generalizada com fração de cura

Pelas características do modelo com fração de cura, têm-se duas possibilidades: o modelo considerando falhas e censuras e o modelo considerando falhas, censuras e covariá-

veis.

Sejam $y_1, y_2, \dots, y_n$ uma amostra aleatória em que $y_i = \log(t_i)$. Considerando que a variável aleatória Y tem distribuição log-gama generalizada dada pela expressão (15) com vetor de parâmetros $\boldsymbol{\theta} = (\mu, \sigma, q)^T$, a contribuição para a função de verossimilhança de um indivíduo que está sob risco, é:

$$P(b_i = 1)f(y_i; \boldsymbol{\theta}|b_i = 1) = (1 - p_i)f(y_i; \boldsymbol{\theta}|b_i = 1) \quad (32)$$

e considerando-se a distribuição log-gama generalizada estendida (15), tem-se:

$$\begin{cases} (1 - p_i) \frac{|q|}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[q z_i - \exp(q z_i) \right] \right\}, & \text{se } q \neq 0 \\ (1 - p_i) \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} (z_i)^2 \right\}, & \text{se } q = 0 \end{cases} \quad (33)$$

em que $z_i = \frac{y_i - \mu}{\sigma}$.

A contribuição para a função de verossimilhança de um indivíduo que está imune, é dada pela função de sobrevivência imprópria

$$G(y_i, \boldsymbol{\theta}) = P(b_i = 0) + P(b_i = 1)S(y_i; \boldsymbol{\theta}|b_i = 1) = p_i + (1 - p_i)S(y_i; \boldsymbol{\theta}), \quad (34)$$

em que $S(y_i; \boldsymbol{\theta})$ é a função de sobrevivência própria e considerando-se a distribuição log-gama generalizada estendida, tem-se:

$$\begin{cases} p_i + (1 - p_i)Q[q^{-2}, q^{-2}\exp(q z_i)], & \text{se } q > 0 \\ p_i + (1 - p_i)\left(1 - Q[q^{-2}, q^{-2}\exp(q z_i)]\right), & \text{se } q < 0 \\ p_i + (1 - p_i)[1 - \Phi(z_i)], & \text{se } q = 0. \end{cases} \quad (35)$$

Assim, utilizando-se as expressões (33) e (35) tem-se que a contribuição total

do i -ésimo indivíduo para a função de verossimilhança é dada por:

$$\left\{ \begin{array}{ll} \left[\begin{array}{l} (1-p_i) \frac{q}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[qz_i - \exp(qz_i) \right] \right\} \\ p_i + (1-p_i) Q[q^{-2}, q^{-2} \exp(qz_i)] \end{array} \right]^{1-\delta_i} & \text{se } q > 0 \\ \left[\begin{array}{l} (1-p_i) \frac{-q}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[qz_i - \exp(qz_i) \right] \right\} \\ p_i + (1-p_i) \left(1 - Q[q^{-2}, q^{-2} \exp(qz_i)] \right) \end{array} \right]^{1-\delta_i} & \text{se } q < 0 \\ \left[\begin{array}{l} (1-p_i) \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} (z_i)^2 \right\} \\ p_i + (1-p_i) \left\{ 1 - \Phi(z_i) \right\} \end{array} \right]^{1-\delta_i} & \text{se } q = 0. \end{array} \right. \quad (36)$$

em que δ_i é o indicador de censura.

Portanto, utilizando a expressão (36), o logaritmo da função de verossimilhança pode ser escrito como

$$l(\boldsymbol{\theta}) = \left\{ \begin{array}{ll} \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i)q}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[qz_i - \exp(qz_i) \right] \right\} \right] \\ + \sum_{i:\delta_i=0} \log \left[p_i + (1-p_i) Q[q^{-2}, q^{-2} \exp(qz_i)] \right], & \text{se } q > 0 \\ \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i)(-q)}{\sigma \Gamma(q^{-2})} (q^{-2})^{q^{-2}} \exp \left\{ q^{-2} \left[qz_i - \exp(qz_i) \right] \right\} \right] \\ + \sum_{i:\delta_i=0} \log \left[p_i + (1-p_i) \left(1 - Q[q^{-2}, q^{-2} \exp(qz_i)] \right) \right], & \text{se } q < 0 \\ \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i)}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} (z_i)^2 \right\} \right] \\ + \sum_{i:\delta_i=0} \log \left[p_i + (1-p_i) \left\{ 1 - \Phi(z_i) \right\} \right], & \text{se } q = 0 \end{array} \right. \quad (37)$$

em que $Q(., .)$ é a função integral gama incompleta definida na equação (19), $\Phi(.)$ é a função de distribuição acumulada normal padrão, δ_i é o indicador de cenuras e $\boldsymbol{\theta} = (\mu, \sigma, p_1, p_2, \dots, p_n)^T$ é o vetor de parâmetros. Assim, nota-se que o vetor de parâmetros é de dimensão $(n+3)$, e portanto, será necessário utilizar métodos iterativos para estimar os parâmetros.

3.2.2 Modelo de regressão log-gama generalizada com fração de cura

O interesse em se introduzirem covariáveis no modelo de regressão log-gama generalizado estendido com fração de cura está em detectar se existe ou não influência (positiva ou negativa) das variáveis regressoras na proporção de indivíduos curados ou imunes. Farewell (1977) e Miller (1981) utilizam o modelo logito para incluir covariáveis no modelo, isto é, covariáveis na probabilidade de um indivíduo estar imune da seguinte forma:

$$p_i = P(b_i = 0) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}. \quad (38)$$

Assim, substituindo a equação (38) no logaritmo da função de verossimilhança (37), observa-se que o vetor de parâmetros é $\boldsymbol{\theta} = (\mu, \sigma, \beta_0, \beta_1, \dots, \beta_p)^T$ e $\mathbf{x}_i^T = (x_{i0}, x_{i1}, \dots, x_{ip})$ é o vetor de covariáveis associadas ao i -ésimo indivíduo.

Para se obterem os estimadores de máxima verossimilhança seria preciso derivar $l(\boldsymbol{\theta})$ em relação ao vetor de parâmetros $\boldsymbol{\theta}$ e q . Contudo, dificuldades existentes na derivação de $l(\boldsymbol{\theta})$ em relação ao parâmetro q é ainda maior neste caso em que admitimos fração de cura. Então, será utilizado o método descrito por Lawless (2003). Ele propõe considerar duas etapas, onde a primeira etapa considera o parâmetro q fixo, de modo que para um dado valor de q encontramos os valores $\tilde{\mu}(q)$, $\tilde{\sigma}(q)$, $\tilde{\beta}_0(q)$, $\dots$, $\tilde{\beta}_p(q)$, que maximizam $l(\boldsymbol{\theta})$. Feito isso, para alguns valores de q , pode-se calcular o valor da função de verossimilhança para esses valores de q e, então, localizar $\hat{q}$ que leva à verossimilhança máxima. Nessa parte do trabalho será utilizado no *software* SAS basicamente o procedimento NLMIXED, cujo programa será descrito no Anexo A.

Podemos também calcular estimativas para a variância de $\hat{\boldsymbol{\theta}}$ e construir testes de hipóteses para os parâmetros utilizando o fato que $\hat{\boldsymbol{\theta}}$ tem distribuição assintótica normal multivariada sob certas condições de regularidade com média $\boldsymbol{\theta}$ e matriz de covariância $I^{-1}(\boldsymbol{\theta})$ com $I(\boldsymbol{\theta}) = -E[\ddot{\mathbf{L}}(\boldsymbol{\theta})]$ tal que $\ddot{\mathbf{L}}(\boldsymbol{\theta}) = \left\{ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}$, isto é,

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

Visto que o cálculo de $I(\boldsymbol{\theta})$ (denominada matriz de informação de Fischer) não

é possível devido a presença de censuras, pode-se utilizar alternativamente a matriz $-\ddot{\mathbf{L}}_{(\boldsymbol{\theta})}$ avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ (denominada matriz de informação observada), a qual é uma estimativa consistente de $I(\boldsymbol{\theta})$. Neste caso, a matriz $\ddot{\mathbf{L}}_{(\boldsymbol{\theta})}$ tem a seguinte forma:

$$\ddot{\mathbf{L}}_{(\boldsymbol{\theta})} = \begin{pmatrix} \ddot{L}_{\mu\mu} & \ddot{L}_{\mu\sigma} & \ddot{L}_{\mu\beta} \\ \cdot & \ddot{L}_{\sigma\sigma} & \ddot{L}_{\sigma\beta} \\ \cdot & \cdot & \ddot{L}_{\beta\beta} \end{pmatrix}$$

sendo que cada elemento da matriz $\ddot{\mathbf{L}}_{(\boldsymbol{\theta})}$ será definido na Apêndice A.

Assim, o intervalo de confiança assintótico para β_j , $j = 0, \dots, p$ considerando $(1 - \alpha)100\%$ de confiança será dado por

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\hat{Var}(\hat{\beta}_j)}. \quad (39)$$

Para a construção de um teste de hipótese para os parâmetros também será utilizada a diagonal principal da matriz $-\ddot{\mathbf{L}}_{(\boldsymbol{\theta})}^{-1}$ como uma estimativa para a variância dos parâmetros, logo a estatística para testar as hipóteses

$$\begin{cases} H_0 : \beta_j = \beta_{j0} \\ H_a : \beta_j \neq \beta_{j0} \end{cases}$$

será dada por $z = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \sim N(0, 1)$.

3.2.3 Influência local

Neste momento será aplicada a influência local no modelo de regressão log-gama generalizado com fração de cura, introduzida na seção 2.3, para fazer uma análise de influência no modelo. Serão três os esquemas de perturbação utilizados: Ponderação de casos, perturbação das respostas e perturbação de uma covariável.

3.2.3.1 Ponderação de casos

O modelo ajustado será perturbado através da introdução de pesos em cada elemento do logaritmo da função de verossimilhança, definindo

$$l(\boldsymbol{\theta}|\mathbf{w}) = \sum_{i=1}^n w_i l_i(\boldsymbol{\theta}).$$

Neste caso, o vetor correspondente à não perturbação é o vetor $\mathbf{w}_0 = (1, 1, \dots, 1)^T$. Para o modelo log-gama generalizado estendido têm-se:

(i) $q > 0$

Neste caso, o logaritmo da função de verossimilhança tem a seguinte forma:

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{w}) = & \sum_{i:\delta_i=1} w_i \log \left[\frac{(1-p_i)q}{\Gamma(q^{-2})\sigma} (q^{-2})^{q^{-2}} \exp\{q^{-1}z_i - q^{-2}\exp(qz_i)\} \right] \\ & + \sum_{i:\delta_i=0} w_i \log \left[p_i + (1-p_i)Q[q^{-2}, q^{-2}\exp\{qz_i\}] \right] \end{aligned} \quad (40)$$

onde $0 \leq w_i \leq 1$ e $\mathbf{w} = (1, \dots, 1)^T$. Seja $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \dots, \boldsymbol{\Delta}_{p+2})^T$. Então, os elementos do vetor $\boldsymbol{\Delta}_1$ têm a seguinte forma

$$\Delta_{1i} = \begin{cases} (\hat{q}\hat{\sigma})^{-1}[-1 + \exp\{\hat{q}\hat{z}_i\}] & \forall i : \delta_i = 1 \\ (1 - \hat{p}_i)\hat{u}_i^{\hat{q}-2} \exp\{-\hat{u}_i\}\hat{q} \left[\hat{h}_{i1}\hat{\sigma}\Gamma(\hat{q}^{-2}) \right]^{-1} & \forall i : \delta_i = 0. \end{cases}$$

Por outro lado, os elementos do vetor $\boldsymbol{\Delta}_2$ podem ser dados por

$$\Delta_{2i} = \begin{cases} -(\hat{\sigma})^{-1} - (\hat{q}\hat{\sigma})^{-1}\hat{z}_i[1 - \exp\{\hat{q}\hat{z}_i\}] & \forall i : \delta_i = 1 \\ (1 - \hat{p}_i)\hat{u}_i^{\hat{q}-2} \exp\{-\hat{u}_i\}\hat{z}_i\hat{q} \left[\hat{h}_{i1}\hat{\sigma}\Gamma(\hat{q}^{-2}) \right]^{-1} & \forall i : \delta_i = 0. \end{cases}$$

Os elementos do vetor $\boldsymbol{\Delta}_j$, for $j = 3, \dots, p+2$, podem ser expressos como

$$\Delta_{ji} = \begin{cases} -x_{ij}\hat{p}_i(1 - \hat{p}_i)^{-1}[1 + \exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}]^{-1} & \forall i : \delta_i = 1 \\ x_{ij}\hat{p}_i[1 - Q(\hat{q}^{-2}, \hat{u}_i)][1 + \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})]^{-1}[\hat{h}_{i1}]^{-1} & \forall i : \delta_i = 0. \end{cases}$$

(ii) $q < 0$

Neste caso, o logaritmo da função de verossimilhança pode ser expresso da seguinte forma

$$l(\boldsymbol{\theta}|\mathbf{w}) = \sum_{i:\delta_i=1} w_i \log \left[\frac{(1-p_i)(-q)}{\Gamma(q^{-2})\sigma} (q^{-2})^{q^{-2}} \exp\{q^{-1}z_i - q^{-2}\exp(qz_i)\} \right] \\ + \sum_{i:\delta_i=0} w_i \log \left[p_i + (1-p_i) \left\{ 1 - Q[q^{-2}, q^{-2}\exp\{qz_i\}] \right\} \right] \quad (41)$$

onde $0 \leq w_i \leq 1$ e $\mathbf{w} = (1, \dots, 1)^T$.

Então, os elementos do vetor $\boldsymbol{\Delta}_1$ são

$$\Delta_{1i} = \begin{cases} (\hat{q}\hat{\sigma})^{-1}[-1 + \exp\{\hat{q}\hat{z}_i\}] & \forall i : \delta_i = 1 \\ (1 - \hat{p}_i)\hat{u}_i^{\hat{q}-2} \exp\{-\hat{u}_i\}(-\hat{q}) \left[\hat{h}_{i2}\hat{\sigma}\Gamma(\hat{q}^{-2}) \right]^{-1} & \forall i : \delta_i = 0. \end{cases}$$

Por outro lado, os elementos do vetor $\boldsymbol{\Delta}_2$ podem ser dados por

$$\Delta_{2i} = \begin{cases} -(\hat{\sigma})^{-1} - (\hat{q}\hat{\sigma})^{-1}\hat{z}_i[1 - \exp\{\hat{q}\hat{z}_i\}] & \forall i : \delta_i = 1 \\ (1 - \hat{p}_i)\hat{u}_i^{\hat{q}-2} \exp\{-\hat{u}_i\}\hat{z}_i(-\hat{q}) \left[\hat{h}_{i2}\hat{\sigma}\Gamma(\hat{q}^{-2}) \right]^{-1} & \forall i : \delta_i = 0. \end{cases}$$

Os elementos do vetor $\boldsymbol{\Delta}_j$, para $j = 3, \dots, p+2$, podem ser expressos como

$$\Delta_{ji} = \begin{cases} -x_{ij}\hat{p}_i(1 - \hat{p}_i)^{-1}[1 + \exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}]^{-1} & \forall i : \delta_i = 1 \\ x_{ij}\hat{p}_i Q(\hat{q}^{-2}, \hat{u}_i)[1 + \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})]^{-1}[\hat{h}_{i2}]^{-1} & \forall i : \delta_i = 0. \end{cases}$$

(iii) $q=0$

Neste caso, o logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\theta}|\mathbf{w}) = \sum_{i:\delta_i=1} w_i \log \left[\frac{(1-p_i)}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{z_i^2}{2}\right\} \right] + \sum_{i:\delta_i=0} w_i \log \left[p_i + (1-p_i) \{1 - \Phi(z_i)\} \right] \quad (42)$$

onde $0 \leq w_i \leq 1$ e $\mathbf{w} = (1, \dots, 1)^T$.

Assim, os elementos do vetor $\boldsymbol{\Delta}_1$ são dados da seguinte forma

$$\Delta_{1i} = \begin{cases} \hat{\sigma}^{-1}\hat{z}_i & \forall i : \delta_i = 1 \\ (1 - \hat{p}_i)\phi(\hat{z}_i)\hat{\sigma}^{-1}\hat{h}_{i3}^{-1} & \forall i : \delta_i = 0. \end{cases}$$

Por outro lado, os elementos do vetor $\boldsymbol{\Delta}_2$ podem ser escritos como

$$\Delta_{2i} = \begin{cases} \hat{\sigma}^{-1}(-1 + \hat{z}_i^2) & \forall i : \delta_i = 1 \\ (1 - \hat{p}_i)\phi(\hat{z}_i)\hat{z}_i\hat{\sigma}^{-1}\hat{h}_{i3}^{-1} & \forall i : \delta_i = 0. \end{cases}$$

Os elementos do vetor Δ_j , para $j = 3, \dots, p+2$, podem ser expressos como

$$\Delta_{ji} = \begin{cases} -x_{ij}\hat{p}_i(1-\hat{p}_i)[1 + \exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}] & \forall i : \delta_i = 1 \\ x_{ij}\hat{p}_i\Phi(\hat{z}_i)[1 + \exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}]\hat{h}_{i3}^{-1} & \forall i : \delta_i = 0. \end{cases}$$

3.2.3.2 Perturbação da variável resposta

Considera-se que cada y_i é perturbado como $y_{iw} = y_i + w_i S_y$, onde S_y pode ser a estimativa do desvio padrão da variável Y e $w_i \in \mathbb{R}$.

Aqui, o vetor de não perturbação é dado por $\mathbf{w}_0 = (0, \dots, 0)^T$. Adotando o modelo em estudo temos:

(i) $q > 0$

O logaritmo da função de verossimilhança perturbada pode ser expressa como

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{w}) &= \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i)q}{\Gamma(q^{-2})\sigma} (q^{-2})^{q^{-2}} \exp\{q^{-1}z_i^* - q^{-2}\exp(qz_i^*)\} \right] \\ &\quad + \sum_{i:\delta_i=0} \log \left[p_i + (1-p_i)Q[q^{-2}, q^{-2}\exp\{qz_i^*\}] \right] \end{aligned} \quad (43)$$

onde $z_i^* = \frac{(y_i + w_i S_y) - \mu}{\sigma}$. Assim, os elementos do vetor Δ_1 tem a seguinte forma

$$\Delta_{1i} = \begin{cases} \hat{q}^{-1}\hat{\sigma}^{-2}\exp\{\hat{q}\hat{z}_i\}S_y & \forall i : \delta_i = 1 \\ \frac{(1-\hat{p}_i)\hat{q}S_y}{\hat{\sigma}\Gamma(\hat{q}^{-2})} \left[\frac{\hat{u}_i^{\hat{q}^{-2}}(\hat{q}^{-1}-\hat{u}_i)\exp\{-\hat{u}_i\}\hat{q}}{\hat{\sigma}\hat{h}_{i1}} + \frac{\hat{u}_i^{2\hat{q}^{-2}}\hat{q}(1-\hat{p}_i)\exp\{-2\hat{u}_i\}}{\hat{\sigma}\Gamma(\hat{q}^{-2})\hat{h}_{i1}^2} \right] & \forall i : \delta_i = 0. \end{cases}$$

Por outro lado, os elementos do vetor Δ_2 podem ser dados por

$$\Delta_{2i} = \begin{cases} -(\hat{q}\hat{\sigma}^2)^{-1}S_y[-1 + \exp\{\hat{q}\hat{z}_i\}(1 + \hat{z}_i\hat{q})] & \forall i : \delta_i = 1 \\ \frac{(1-\hat{p}_i)\hat{q}S_y}{\hat{\sigma}^2\Gamma(\hat{q}^{-2})} \left\{ \frac{\hat{u}_i^{\hat{q}^{-2}}[\hat{z}_i(\hat{q}^{-1}-\hat{u}_i)\exp\{-\hat{u}_i\}\hat{q} + \exp\{-\hat{u}_i\}]}{\hat{h}_{i1}} + \frac{\hat{u}_i^{2\hat{q}^{-2}}\exp\{-2\hat{u}_i\}\hat{z}_i(1-\hat{p}_i)\hat{q}}{\Gamma(\hat{q}^{-2})\hat{h}_{i1}^2} \right\} & \forall i : \delta_i = 0. \end{cases}$$

Os elementos do vetor Δ_j , para $j = 3, \dots, p+2$, podem ser expressos como

$$\Delta_{ji} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{x_{ij}\hat{p}_i\hat{u}_i^{\hat{q}^{-2}}\exp\{-\hat{u}_i\}\hat{q}S_y}{\hat{\sigma}\Gamma(\hat{q}^{-2})[1+\exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}]} \left\{ \hat{h}_i^{-1} + \hat{h}_i^{-2}(1-\hat{p}_i) \left[1 - Q(\hat{q}^{-2}, \hat{u}_i) \right] \right\} & \forall i : \delta_i = 0. \end{cases}$$

(ii) $q < 0$

Neste caso, o logaritmo da função de verossimilhança perturbada é expresso como

$$l(\boldsymbol{\theta}|\mathbf{w}) = \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i)(-q)}{\Gamma(q^{-2})\sigma} (q^{-2})^{q^{-2}} \exp\{q^{-1}z_i^* - q^{-2}\exp(qz_i^*)\} \right] \\ + \sum_{i:\delta_i=0} \log \left[p_i + (1-p_i) \left\{ 1 - Q[q^{-2}, q^{-2}\exp\{qz_i^*\}] \right\} \right]. \quad (44)$$

Então os elementos do vetor $\boldsymbol{\Delta}_1$ têm a seguinte forma

$$\Delta_{1i} = \begin{cases} \hat{q}^{-1}\hat{\sigma}^{-2}\exp\{\hat{q}\hat{z}_i\}S_y & \forall i : \delta_i = 1 \\ \frac{(1-\hat{p}_i)(-\hat{q})S_y}{\hat{\sigma}\Gamma(\hat{q}^{-2})} \left[\frac{\hat{u}_i^{\hat{q}^{-2}}(\hat{q}^{-1}-\hat{u}_i\exp\{-\hat{u}_i\}\hat{q})}{\hat{\sigma}\hat{h}_{i2}} - \frac{\hat{u}_i^{2\hat{q}^{-2}}\hat{q}(1-\hat{p}_i)\exp\{-2\hat{u}_i\}}{\hat{\sigma}\Gamma(\hat{q}^{-2})\hat{h}_{i2}^2} \right] & \forall i : \delta_i = 0. \end{cases}$$

Por outro lado, os elementos do vetor $\boldsymbol{\Delta}_2$ podem ser expressos como

$$\Delta_{2i} = \begin{cases} -(\hat{q}\hat{\sigma}^2)^{-1}S_y[1 - \exp\{\hat{q}\hat{z}_i\}(1 + \hat{z}_i\hat{q})] & \forall i : \delta_i = 1 \\ \frac{(1-\hat{p}_i)(-\hat{q})S_y}{\hat{\sigma}^2\Gamma(\hat{q}^{-2})} \left\{ \frac{\hat{u}_i^{\hat{q}^{-2}}[z_i(\hat{q}^{-1}-\hat{u}_i\exp\{-\hat{u}_i\}\hat{q}) + \exp\{-\hat{u}_i\}]}{\hat{h}_{i2}} + \frac{\hat{u}_i^{2\hat{q}^{-2}}\exp\{-2\hat{u}_i\}\hat{z}_i(1-\hat{p}_i)\hat{q}}{\Gamma(\hat{q}^{-2})\hat{h}_{i2}^2} \right\} & \forall i : \delta_i = 0. \end{cases}$$

Os elementos do vetor $\boldsymbol{\Delta}_j$, for $j = 3, \dots, p+2$, podem ser expressos como

$$\Delta_{ji} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{x_{ij}\hat{p}_i\hat{u}_i^{\hat{q}^{-2}}\exp\{-\hat{u}_i\}(-\hat{q})S_y}{\hat{\sigma}\Gamma(\hat{q}^{-2})[1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]} \left\{ \hat{h}_i^{-1} + \hat{h}_i^{-2}(1-\hat{p}_i)Q(\hat{q}^{-2}, \hat{u}_i) \right\} & \forall i : \delta_i = 0. \end{cases}$$

(iii) $q=0$

Neste caso, o logaritmo da função de verossimilhança perturbada é dada por

$$l(\boldsymbol{\theta}|\mathbf{w}) = \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i)}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{z_i^{2*}}{2}\right\} \right] + \sum_{i:\delta_i=0} \log \left[p_i + (1-p_i) \{1 - \Phi(z_i^*)\} \right]. \quad (45)$$

Então os elementos do vetor $\boldsymbol{\Delta}_1$ têm a seguinte forma

$$\Delta_{1i} = \begin{cases} \hat{\sigma}^{-2}S_y & \forall i : \delta_i = 1 \\ \frac{(1-\hat{p}_i)\phi(\hat{z}_i)S_y}{\hat{\sigma}^2} \left[-\hat{z}_i\hat{h}_{i3}^{-1} + \phi(\hat{z}_i)(1-\hat{p}_i)\hat{h}_{i3}^{-2} \right] & \forall i : \delta_i = 0. \end{cases}$$

Por outro lado, os elementos do vetor Δ_2 podem ser dados como

$$\Delta_{2i} = \begin{cases} 2\hat{\sigma}^{-2}\hat{z}_i S_y & \forall i : \delta_i = 1 \\ \frac{(1-\hat{p}_i)\phi(\hat{z}_i)S_y}{\hat{\sigma}^2} \left[(1-\hat{z}_i^2)\hat{h}_{i3}^{-1} + \phi(\hat{z}_i)(1-\hat{p}_i)\hat{z}_i\hat{h}_{i3}^{-2} \right] & \forall i : \delta_i = 0. \end{cases}$$

Os elementos do vetor Δ_j , for $j = 3, \dots, p+2$, podem ser expressos como

$$\Delta_{ji} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{x_{ij}\hat{p}_i\phi(\hat{z}_i)S_y}{\hat{\sigma}[1+\exp\{\mathbf{x}_i^T\hat{\beta}\}]} \left[\hat{h}_{i3}^{-1} + \Phi(\hat{z}_i)(1-\hat{p}_i)\hat{h}_{i3}^{-2} \right] & \forall i : \delta_i = 0. \end{cases}$$

3.2.3.3 Perturbação de uma covariável

Agora considera-se uma perturbação na t -ésima covariável, X_t , fazendo $x_{itw} = x_{it} + w_i S_t$, onde S_t é um fator de escala, $w_i \in \mathbb{R}$. Neste caso, o vetor de não perturbação é $\mathbf{w}_0 = (0, \dots, 0)^T$. Para o modelo em estudo têm-se:

(i) $q > 0$

Neste caso, o logaritmo da função de verossimilhança perturbada é dado por

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{w}) &= \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i^*)q}{\Gamma(q^{-2})\sigma} (q^{-2})^{q^{-2}} \exp\{q^{-1}z_i - q^{-2}\exp(qz_i)\} \right] \\ &\quad + \sum_{i:\delta_i=0} \log \left[p_i^* + (1-p_i^*)Q[q^{-2}, q^{-2}\exp\{qz_i\}] \right] \end{aligned} \quad (46)$$

onde $pi^* = \frac{\exp(\mathbf{x}_i^{*T}\hat{\beta})}{1+\exp(\mathbf{x}_i^{*T}\hat{\beta})}$ and $\mathbf{x}_i^{*T} = \beta_1 + \beta_2 x_{i2} + \dots + \beta_t(x_{it} + w_i S_t) + \dots + \beta_p x_{ip}$.

Assim, os elementos do vetor Δ_1 são expressos como

$$\Delta_{1i} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{\hat{u}_i^{\hat{q}-2} \exp\{-\hat{u}_i\}(-\hat{q})\hat{p}_i(1-\hat{p}_i)\hat{\beta}_t S_x}{\hat{\sigma}\Gamma(\hat{q}^{-2})} \left\{ \hat{h}_{i1}^{-1} + (1-\hat{p}_i) \left[1 - Q(\hat{q}^{-2}, \hat{u}_i) \right] \hat{h}_{i1}^{-2} \right\} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor Δ_2 são expressos como

$$\Delta_{2i} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{\hat{u}_i^{\hat{q}-2} \exp\{-\hat{u}_i\}\hat{z}_i(-\hat{q})\hat{p}_i(1-\hat{p}_i)\hat{\beta}_t S_x}{\hat{\sigma}\Gamma(\hat{q}^{-2})} \left\{ \hat{h}_{i1}^{-1} + (1-\hat{p}_i) \left[1 - Q(\hat{q}^{-2}, \hat{u}_i) \right] \hat{h}_{i1}^{-2} \right\} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor Δ_j , para $j = 1, \dots, p$ e $j \neq t$, são dados por

$$\Delta_{ji} = \begin{cases} -x_{ij}\hat{p}_i^2\hat{\beta}_t S_x (1-\hat{p}_i)^{-1} [1 + \exp\{\mathbf{x}_i^T\hat{\beta}\}]^{-1} & \forall i : \delta_i = 1 \\ -x_{ij}\hat{p}_i^2(1-\hat{p}_i)\hat{\beta}_t S_x \left[1 - Q(\hat{q}^{-2}, \hat{u}_i) \right]^2 [1 + \exp\{\mathbf{x}_i^T\hat{\beta}\}]^{-1} \hat{h}_{i1}^{-2} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor Δ_t são dados como

$$\Delta_{ti} = \begin{cases} -\hat{p}_i S_x [x_{it} \hat{\beta}_t (1 - \hat{p}_i) + 1] & \forall i : \delta_i = 1 \\ x_{it} \hat{p}_i (1 - \hat{p}_i) [1 - Q(\hat{q}^{-2}, \hat{u}_i)] \\ \left\{ (1 - \hat{\beta}_t \hat{p}_i S_x) \hat{h}_{i1}^{-1} - \hat{p}_i (1 - \hat{p}_i) \hat{\beta}_t S_x [1 - Q(\hat{q}^{-2}, \hat{u}_i)] \hat{h}_{i1}^{-2} \right\} & \forall i : \delta_i = 0. \end{cases}$$

(ii) $q < 0$

Neste caso, o logaritmo da função de verossimilhança perturbada pode ser expressa como

$$\begin{aligned} l(\theta | \mathbf{w}) &= \sum_{i: \delta_i = 1} \log \left[\frac{(1 - p_i^*)(-q)}{\Gamma(q^{-2}) \sigma} (q^{-2})^{q^{-2}} \exp\{q^{-1} z_i - q^{-2} \exp(q z_i)\} \right] \\ &\quad + \sum_{i: \delta_i = 0} \log \left[p_i^* + (1 - p_i^*) \left\{ 1 - Q[q^{-2}, q^{-2} \exp\{q z_i\}] \right\} \right] \end{aligned} \quad (47)$$

Assim, os elementos do vetor Δ_1 são expressos como

$$\Delta_{1i} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{\hat{u}_i^{q^{-2}} \exp\{-\hat{u}_i\} \hat{q} \hat{p}_i (1 - \hat{p}_i) \hat{\beta}_t S_x}{\partial \Gamma(\hat{q}^{-2})} \left\{ \hat{h}_{i2}^{-1} + (1 - \hat{p}_i) Q(\hat{q}^{-2}, \hat{u}_i) \hat{h}_{i2}^{-2} \right\} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor Δ_2 são

$$\Delta_{2i} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ \frac{\hat{u}_i^{q^{-2}} \exp\{-\hat{u}_i\} \hat{z}_i \hat{q} \hat{p}_i (1 - \hat{p}_i) \hat{\beta}_t S_x}{\partial \Gamma(\hat{q}^{-2})} \left\{ \hat{h}_{i2}^{-1} + (1 - \hat{p}_i) Q(\hat{q}^{-2}, \hat{u}_i) \hat{h}_{i2}^{-2} \right\} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor Δ_j , para $j = 1, \dots, p$ e $j \neq t$, são da seguinte forma

$$\Delta_{ji} = \begin{cases} -x_{ij} \hat{p}_i^2 \hat{\beta}_t S_x (1 - \hat{p}_i)^{-1} [1 + \exp\{\mathbf{x}_i^T \hat{\beta}\}]^{-1} & \forall i : \delta_i = 1 \\ -x_{ij} \hat{p}_i^2 (1 - \hat{p}_i) \hat{\beta}_t S_x [Q(\hat{q}^{-2}, \hat{u}_i)]^2 [1 + \exp\{\mathbf{x}_i^T \hat{\beta}\}]^{-1} \hat{h}_{i2}^{-2} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor Δ_t são dados por

$$\Delta_{ti} = \begin{cases} -\hat{p}_i S_x [x_{it} \hat{\beta}_t (1 - \hat{p}_i) + 1] & \forall i : \delta_i = 1 \\ x_{it} \hat{p}_i (1 - \hat{p}_i) Q(\hat{q}^{-2}, \hat{u}_i) \left\{ (1 - \hat{\beta}_t \hat{p}_i S_x) \hat{h}_{i2}^{-1} - \hat{p}_i (1 - \hat{p}_i) \hat{\beta}_t S_x Q(\hat{q}^{-2}, \hat{u}_i) \hat{h}_{i2}^{-2} \right\} & \forall i : \delta_i = 0. \end{cases}$$

(iii) $q=0$

Neste caso, o logaritmo da função de verossimilhança perturbada é dado da seguinte forma

$$l(\boldsymbol{\theta}|\mathbf{w}) = \sum_{i:\delta_i=1} \log \left[\frac{(1-p_i^*)}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{z_i^2}{2}\right\} \right] + \sum_{i:\delta_i=0} \log \left[p_i^* + (1-p_i^*)\{1 - \Phi(z_i)\} \right] \quad (48)$$

onde $0 \leq w_i \leq 1$ e $\mathbf{w} = (1, \dots, 1)^T$.

Assim, os elementos do vetor $\boldsymbol{\Delta}_1$ são expressos como

$$\Delta_{1i} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ -\hat{\sigma}^{-1} \hat{p}_i (1 - \hat{p}_i) \hat{\beta}_t S_x \phi(\hat{z}_i) \left[\hat{h}_{i3}^{-1} - (1 - \hat{p}_i) \Phi(\hat{z}_i) \hat{h}_{i3}^{-2} \right] & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor $\boldsymbol{\Delta}_2$ são expressos como

$$\Delta_{2i} = \begin{cases} 0 & \forall i : \delta_i = 1 \\ -\hat{\sigma}^{-1} \hat{z}_i \hat{p}_i (1 - \hat{p}_i) \hat{\beta}_t S_x \phi(\hat{z}_i) \left[\hat{h}_{i3}^{-1} - (1 - \hat{p}_i) \Phi(\hat{z}_i) \hat{h}_{i3}^{-2} \right] & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor $\boldsymbol{\Delta}_j$, para $j = 1, \dots, p$ e $j \neq t$, são dados da seguinte forma

$$\Delta_{ji} = \begin{cases} -x_{ij} \hat{p}_i^2 (1 - \hat{p}_i)^{-1} \hat{p}_i \hat{\beta}_t S_x & \forall i : \delta_i = 1 \\ -x_{ij} \hat{p}_i^2 (1 - \hat{p}_i) \hat{\beta}_t S_x \left[\Phi(\hat{z}_i) \right]^2 \left[1 + \exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\} \right]^{-1} \hat{h}_{i3}^{-1} & \forall i : \delta_i = 0, \end{cases}$$

os elementos do vetor $\boldsymbol{\Delta}_t$ são dados por

$$\Delta_{ti} = \begin{cases} -\hat{p}_i S_x [x_{it} \hat{\beta}_t (1 - \hat{p}_i) + 1] & \forall i : \delta_i = 1 \\ x_{it} p_i (1 - p_i) \Phi(z_i) \left[(1 - \beta_t S_x p_i) h_{i3}^{-1} - p_i (1 - p_i) \beta_t S_x \Phi(z_i) h_{i3}^{-2} \right] & \forall i : \delta_i = 0. \end{cases}$$

Onde

$$\hat{h}_{i1} = \left[\hat{p}_i + (1 - \hat{p}_i) Q(\hat{q}^{-2}, \hat{u}_i) \right], \quad \hat{h}_{i2} = \left[\hat{p}_i + (1 - \hat{p}_i) \left\{ 1 - Q(\hat{q}^{-2}, \hat{u}_i) \right\} \right],$$

$$\hat{h}_{i3} = \left[\hat{p}_i + (1 - \hat{p}_i) \left\{ 1 - \Phi(\hat{z}_i) \right\} \right], \quad \hat{u}_i = \hat{q}^{-2} \exp\{\hat{q} \hat{z}_i\}, \quad \hat{z}_i = \frac{y_i - \hat{\mu}}{\hat{\sigma}}$$

$$\hat{p}_i = \frac{\exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})}$$

$\Phi(\cdot)$ denota a função de distribuição da normal padrão, $\phi(\cdot)$ a densidade da normal padrão e $Q(\cdot)$ é a função integral gama incompleta.

3.2.4 Análise de resíduos

Neste momento será realizada a análise de resíduos para o modelo de regressão log-gama generalizado com fração de cura na presença de covariáveis, os resíduos utilizados serão aqueles introduzidos na seção 2.4.

3.2.4.1 Resíduos martingal

No resíduo martingal em (30) têm-se ele em função da função de risco acumulada, entretando é sabido que a mesma pode ser escrita da seguinte forma

$$H(t_i|x_i) = -\log(S(t_i)) \quad (49)$$

onde $S(t_i)$ é a função de sobrevivência. Mas neste modelo a função de sobrevivência a ser utilizada será a função de sobrevivência imprópria $G(t)$, logo temos que o resíduo martingal pode ser escrito como

(i) Para $q > 0$

$$r_{M_i} = \begin{cases} 1 + \log\left\{\hat{p} + (1 - \hat{p})Q[q^{-2}, q^{-2}\exp(\hat{q}\hat{z}_i)]\right\}, & \forall i : \delta_i = 1 \\ \log\left\{\hat{p} + (1 - \hat{p})Q[q^{-2}, q^{-2}\exp(\hat{q}\hat{z}_i)]\right\}, & \forall i : \delta_i = 0 \end{cases}$$

(ii) Para $q < 0$

$$r_{M_i} = \begin{cases} 1 + \log\left\{\hat{p} + (1 - \hat{p})\left(1 - Q[q^{-2}, q^{-2}\exp(\hat{q}\hat{z}_i)]\right)\right\}, & \forall i : \delta_i = 1 \\ \log\left\{\hat{p} + (1 - \hat{p})\left(1 - Q[q^{-2}, q^{-2}\exp(\hat{q}\hat{z}_i)]\right)\right\}, & \forall i : \delta_i = 0 \end{cases}$$

(iii) Para $q = 0$

$$r_{M_i} = \begin{cases} 1 + \log\left\{\hat{p} + (1 - \hat{p})\left(1 - \Phi(\hat{z}_i)\right)\right\}, & \forall i : \delta_i = 1 \\ \log\left\{\hat{p} + (1 - \hat{p})\left(1 - \Phi(\hat{z}_i)\right)\right\}, & \forall i : \delta_i = 0 \end{cases}$$

3.2.4.2 Resíduos deviance

Substituindo o resíduo martingal no resíduo deviance como em (31) têm-se:

(i) Para $q > 0$

$$r_{D_i} = \begin{cases} \text{sinal}(r_{M_i})\sqrt{2}\left\{-1 - \log\left[G_1(y_i, \hat{\boldsymbol{\theta}})\log\{G_1(y_i, \hat{\boldsymbol{\theta}})^{-1}\}\right]\right\}^{1/2}, & \forall i : \delta_i = 1 \\ \text{sinal}(r_{M_i})\sqrt{2}\left[\log G_1(y_i, \hat{\boldsymbol{\theta}})^{-1}\right]^{1/2}, & \forall i : \delta_i = 0 \end{cases}$$

onde $G_1(y_i, \hat{\boldsymbol{\theta}}) = \hat{p} + (1 - \hat{p})Q[q^{-2}, q^{-2}\exp(\hat{q}\hat{z}_i)]$.

(ii) Para $q < 0$

$$r_{D_i} = \begin{cases} \text{sinal}(r_{M_i})\sqrt{2}\left\{-1 - \log\left[G_2(y_i, \hat{\boldsymbol{\theta}})\log\{G_2(y_i, \hat{\boldsymbol{\theta}})^{-1}\}\right]\right\}^{1/2}, & \forall i : \delta_i = 1 \\ \text{sinal}(r_{M_i})\sqrt{2}\left[\log G_2(y_i, \hat{\boldsymbol{\theta}})^{-1}\right]^{1/2}, & \forall i : \delta_i = 0 \end{cases}$$

onde $G_2(y_i, \hat{\boldsymbol{\theta}}) = \hat{p} + (1 - \hat{p})\left(1 - Q[q^{-2}, q^{-2}\exp(\hat{q}\hat{z}_i)]\right)$.

(iii) Para $q = 0$

$$r_{D_i} = \begin{cases} \text{sinal}(r_{M_i})\sqrt{2}\left\{-1 - \log\left[G_3(y_i, \hat{\boldsymbol{\theta}})\log\{G_3(y_i, \hat{\boldsymbol{\theta}})^{-1}\}\right]\right\}^{1/2}, & \forall i : \delta_i = 1 \\ \text{sinal}(r_{M_i})\sqrt{2}\left[\log G_3(y_i, \hat{\boldsymbol{\theta}})^{-1}\right]^{1/2}, & \forall i : \delta_i = 0 \end{cases}$$

onde $G_3(y_i, \hat{\boldsymbol{\theta}}) = \hat{p} + (1 - \hat{p})\left(1 - \Phi(\hat{z}_i)\right)$.

3.2.5 Impacto das observações influentes

Para verificar o impacto das observações detectadas como influentes, os parâmetros serão novamente estimados, porém sem as observações em questão. Sejam, respectivamente, $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}^0$ os estimadores de máxima verossimilhança do modelo obtidos através dos dados com e sem as observações influentes. Lee, Lu e Song (2006) definem as seguintes quantidades para medir a diferença entre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}^0$:

$$TRC = \sum_{i=1}^{n_p} \left| \frac{\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_i^0}{\hat{\boldsymbol{\theta}}_i} \right| \quad e \quad MRC = \max_i \left| \frac{\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_i^0}{\hat{\boldsymbol{\theta}}_i} \right|$$

onde TRC é a mudança total relativa, MRC é a mudança relativa máxima e n_p é o número de parâmetros. O afastamento da verossimilhança é dado por $LD_I(\boldsymbol{\theta}) = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(I)})\}$, onde $\hat{\boldsymbol{\theta}}_{(I)}$ denota o estimador de máxima verossimilhança para $\boldsymbol{\theta}$ depois do conjunto (I) de observações influentes ter sido removido (ver, COOK, PEÑA e WEISBERG, 1988)

Para detectar o impacto das observações influentes, repete-se a análise retirando a mesma quantidade de observações aleatoriamente da amostra, porém observações não influentes, para depoisas medidas TRC e MRC serem comparadas.

4 RESULTADOS E DISCUSSÃO

Como ilustração, será apresentado aqui a aplicação dos resultados desenvolvidos no capítulo 3, mediante um conjunto de dados reais introduzidos na seção 3.1. Foi implementado uma macro no *software* SAS para encontrar os estimadores de máxima verossimilhança, os quais estão no Anexo A.

Na Figura 3 é apresentado o gráfico da função de sobrevivência estimada de Kaplan-Meier, onde pode-se observar a existência de uma significativa fração de sobreviventes.

Figura 3 - Gráfico de Kaplan Meier

4.1 Resultados dos estimadores de máxima verossimilhança

Ajustando o modelo log-gama generalizado estendido com fração de cura via método da máxima verossimilhança perfilada pode-se observar através da Figura 4 que quando $q = 0,25$, $l(\hat{\mu}(q), \hat{\sigma}(q), \hat{\beta}_0(q), \dots, \hat{\beta}_p(q))$ atinge o valor máximo. Assim, assume-se este valor como a estimativa de máxima verossimilhança para o parâmetro q .

Figura 4 - Gráfico do método da máxima verossimilhança perfilada

A tabela 1 apresenta os valores das estimativas de máxima verossimilhança para os demais parâmetros quando $\hat{q} = 0,25$

Tabela 1 - Estimativas de Máxima Verossimilhança

Parametros	Estimativas	Erro padrão	p-valor	Intervalo de Confiança
μ	0,6529	0,0934	<0,0001	(0,4692 ; 0,8366)
σ	0,7789	0,0648	<0,0001	(0,6514 ; 0,9062)
β_0	2,4047	0,6554	0,0003	(1,1165 ; 3,6929)
β_1	-0,1590	0,2427	0,5128	(-0,6360 ; 0,3181)
β_2	-0,0158	0,0093	0,0912	(-0,0341 ; 0,0025)
β_3	-0,6016	0,1406	<0,0001	(-0,8780 ; -0,3253)
β_4	0,1999	0,2493	0,4231	(-0,2902 ; 0,6900)
β_5	-0,1367	0,3661	0,7091	(-0,8564 ; 0,5830)
β_6	-0,0765	0,0481	0,1131	(-0,1712 ; 0,8201)
q	0,2500			
p	0,4837			

A proporção de cura estimada foi $\hat{p} = 0,4837$ e a única variável significativa é x_3 (nódulo), considerando um nível de 5% de significância. Assim, uma interpretação direta pode ser descrita da seguinte forma: a medida que a variável nódulo aumenta, a proporção de curados tende a diminuir. Também pode-se observar que a variável idade é significativa

quando se considera um nível 10% de significância.

Para estudar a sensibilidade do modelo proposto, faremos um estudo de diagnóstico e resíduos baseado na teoria de influência local e resíduos.

4.2 Análise de influência local

A seguir apresenta-se os resultados da análise de influência local considerando os esquemas: ponderação de casos, perturbação da variável resposta e perturbação de uma covariável.

4.2.1 Ponderação de casos

De acordo com o esquema de ponderação de casos têm-se que o valor $C_{\ell_{max}} = 1,6566$ foi obtido como a curvatura máxima. Na Figura 5 é apresentado o gráfico do autovetor correspondente a $C_{\ell_{max}}$ e a influencia total C_i é mostrada na Figura 6. As observações que mais se destacam das demais são 47, 68, 176 e 341. A observação 47 representa o indivíduo que tem a maior idade dentro do conjunto dos indivíduos de dados censurados.

Figura 5 - Gráfico de ℓ_{max} para θ (ponderação de casos)

Figura 6 - Influência local total C_i (ponderação de casos)

4.2.2 Perturbação da variável resposta

Considerando perturbações nos tempos de sobrevivência observado, o valor da curvatura máxima calculado foi $C_{\ell_{max}} = 21,334$. Na Figura 7 é apresentado o gráfico para $|\ell_{max}|$ versus o índice das observações, onde pode-se observar que os pontos 23, 47 e 176 são os que se destacam dos demais. Na Figura 8 apresenta-se a influência local total (C_i) onde destacam-se os mesmos pontos da Figura 7. Através da análise desses dados, pode-se observar que as observações em destaque se referem aos pacientes com os menores tempos de sobrevivência, todos não censurados.

Figura 7 - Gráfico de ℓ_{max} para θ (perturbação da variável resposta)

Figura 8 - Influência local total C_i (perturbação da variável resposta)

4.2.3 Perturbação de uma covariável

Continuando com a análise de influência, nesta seção serão apresentados os resultados das perturbações feitas nas covariáveis idade (x_2) e espessura do tumor (x_6). Para a perturbação da covariável idade foi obtido o valor $C_{\ell_{max}} = 1,0375$ para a curvatura máxima,

e para a perturbação da covariável tumor foi obtido o valor $C_{\ell_{max}} = 1,5753$. Os respectivos gráficos para $|\ell_{max}|$ assim como para a influência local total C_i contra os índices das observações são apresentados nas Figuras 9, 10, 11 e 12. De onde observa-se que nenhum ponto se destaca.

Figura 9 - Gráfico de ℓ_{max} para θ (perturbação da covariável idade)

Figura 10 - Influência local total C_i (perturbação da covariável idade)

Figura 11 - Gráfico de ℓ_{max} para θ (perturbação da covariável espessura do tumor)

Figura 12 - Influência local total C_i (perturbação da covariável espesura do tumor)

4.3 Análise de resíduos

Para detectar possíveis *outliers* bem como desvios das suposições do modelo log-gama generalizado com indivíduos de longa duração, nas Figura 13 e 14, respectivamente, são apresentados os gráficos do resíduo martingal (r_{Mi}) e do resíduo deviance (r_{Di}) versus o

índice das observações.

Pela análise gráfica do resíduo martingal e deviance, pode-se observar o comportamento aleatório dos dados e que todos os pontos estão bem modelados. Também pode-se notar uma tendência a formação de dois grupos, este fato deve-se a utilização da ligação logito para a introdução de covariáveis na proporção de curados. Problemas como este, também podem ser observados regressão logística. Para maiores detalhes, consultar McCullagh et al. (1989), Hosmer et al. (1989), entre outros.

Figura 13 - Resíduo Martingal(r_{M_i})

Figura 14 - Resíduo Deviance(r_{D_i})

Assim, baseado nos resultados de influência local e de resíduos, a decisão foi pela retirada de três observações (47, 68, 176) por serem as que mais se destacam em relação aos outros.

4.4 Impacto das observações influentes

Para verificar o impacto das observações detectadas como influentes, os parâmetros serão novamente estimados, porém sem as observações 47, 68 e 176.

Assim, é possível verificar que $TRC = 1,2504$, $MRC = 0,3222$ e $LD_{(I)} = -10,3$. Para detectar o impacto das observações influentes, repete-se a análise retirando a mesma quantidade de observações aleatoriamente, porém observações não influentes, neste caso, estão sendo retiradas as observações 37, 94 e 165. E encontra-se os seguintes resultados: $TRC = 0,2547$, $MRC = 0,1032$ e $LD_{(I)} = -5,8$. Assim, comparando esses resultados pode-se observar que não existe uma mudança significativa entre as medidas TRC , MRC e $LD_{(I)}$.

Na Tabela 2 serão apresentados os estimadores de máxima verossimilhança para os parâmetros com a retirada das observações 47, 68 e 176. A fração de cura média estimada foi de 0,4863 e pode-se notar que a covariável espessura do tumor (x6) é significativa para o modelo. Assim, pode-se interpretar que a medida que a espessura do tumor aumenta, a fração de pacientes curados diminui significativamente.

Tabela 2 - Estimativas de Máxima Verossimilhança (retirando as observações 47, 68 e 176)

Parametros	Estimativas	Erro padrão	p-valor	Intervalo de Confiança
μ	0,6281	0,08559	<0,0001	(0,4599 ; 0,7964)
σ	0,7625	0,06018	<0,0001	(0,6442 ; 0,8808)
β_0	2,5727	0,6491	<0,0001	(1,2697 ; 3,8487)
β_1	-0,1252	0,2412	0,6038	(-0,5993 ; 0,3488)
β_2	-0,0153	0,00923	0,0976	(-0,0335 ; 0,0028)
β_3	-0,6349	0,1396	<0,0001	(-0,9093 ; -0,3604)
β_4	0,1410	0,2478	0,5698	(-0,3462 ; 0,6281)
β_5	-0,1694	0,2659	0,6435	(-0,8886 ; 0,5497)
β_6	-0,1011	0,04846	0,0375	(-0,1964 ; -0,0059)
q	0,2300			
p	0,4863			

4.5 Qualidade de ajuste

Para verificar a qualidade de ajuste, plota-se a função de sobrevivência de Kaplan-Meier e a função estimada pelo modelo log-gama generalizado com fração de cura (ver, Figura 15). Observa-se que o modelo esta bem ajustado.

Figura 15 - Função de sobrevivência do modelo log-gama generalizada, Curva de Kaplan-Meier

Através da comparação dos modelos propostos (considerando e sem considerar possíveis pontos influentes), pode-se observar que não houve uma mudança significativa na estimação da proporção de curados e nas estimativas dos parâmetros. Assim, considera-se que o modelo inicial seja mais adequado pois consegue modelar todas as observações. O único ganho ao se considerar o modelo com a retirada das observações 47, 68 e 176, seria que a variável x_6 (espessura do tumor) passa a ser significativa.

5 Considerações Finais

5.1 Conclusões

Neste trabalho modificamos o modelo de regressão log-gama generalizado para poder incluir indivíduos de longa duração. Na proposta considerada, tomamos como base a modelagem paramétrica log-linear para os tempos de vida. A função logito mostrou-se uma boa função de ligação para poder incluir covariáveis no modelo proposto. A parte inferencial é feita utilizando a distribuição assintótica dos estimadores de máxima verossimilhança. Também discutimos neste trabalho a aplicação da teoria de influência local (COOK, 1986 e THOMAS; COOK, 1990) e um estudo de resíduos baseado no resíduo martingal e deviance no modelo de regressão log-gama generalizado com indivíduos de longa duração para poder avaliar o modelo proposto. Obtivemos as matrizes necessárias para a aplicação da técnica, considerando vários tipos de perturbações nos elementos dos dados e no modelo. Aplicando estes resultados em um conjunto de dados, obtivemos indicações de que o modelo proposto é robusto a perturbações feitas no modelo e nos dados, os estimadores de máxima verossimilhança não mudaram significativamente. Os resultados das aplicações indicam que o uso da técnica de influência local assim como a análise de resíduos no modelo de regressão com indivíduos de longa duração podem ser bastante útil na detecção de possíveis pontos influentes. Para poder medir a qualidade de ajuste usamos os resíduos martingal e deviance, os quais mostram que o ajuste do modelo está correto. Também plotamos a função de sobrevivência de Kaplan-Meier com a função de sobrevivência do modelo proposto que indica um bom ajuste do modelo.

5.2 Perspectivas para trabalhos futuros

Algumas investigações que acredita-se ser de interesse para ampliar e dar continuidade aos resultados contidos neste trabalho são:

- (i) Pode-se incluir covariáveis simultaneamente no tempo e na proporção de curados, isto é, no parâmetro de locação pode-se ter $\mu = \mathbf{x}_i^T \boldsymbol{\gamma}$ e na proporção de curados $p_i =$

$\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$, assim pode-se saber quais são as covariáveis que aceleram ou desaceleram o tempo de vida e quais covariáveis influenciam positiva ou negativamente na proporção de curados. Porém deve-se ter cuidado com a identificabilidade do modelo pois o número de parâmetros nesse caso será de $(2(p+1)+1)$, isto é, $\boldsymbol{\theta} = (\sigma, \beta_0, \beta_1, \dots, \beta_p, \gamma_1, \gamma_2, \dots, \gamma_p)^T$.

- (ii) No lugar da função de ligação logito pode-se utilizar outras funções de ligação, tais como probito e complemento log-log e fazer uma comparação entre elas para observar o comportamento dessas funções.
- (iii) O desenvolvimento e implementação computacional de outros métodos de estimação, tais como, estimador bayesiano e o método de jackknife.
- (iv) O desenvolvimento de uma análise de sensibilidade do modelo utilizando medidas como, influência global, influência local, leverage generalizado, dentre outras.
- (v) Outra possibilidade seria estudar a distribuição dos resíduos propostos via simulação Monte Carlo, e assim poder utilizar o gráfico de envelopes como uma boa medida de qualidade.

REFERÊNCIAS

- BELSEY, D.A.; KUH, E.; WELSC, R. E. **Regression diagnostics**. New York: John Wiley, 1980. 292 p.
- BERKSON, J.; GAGE, R.P. Survival curve for cancer patients following treatment, **Journal of the American Statistical Association**, Alexandria, v. 47, p. 501-515, 1952.
- COOK, R. D. Detection of influential observations in linear regression, **Technometrics**, Wisconsin, v. 19, p. 15-18, 1977.
- COOK, R. D. Assesment of local influence (with discussion), **Journal of the Royal Statistical Society**, London, v. 48, p. 133-169, 1986.
- COOK, R. D.; PEÑA, D.; WEISBERG, S. The likelihood displacement: a unifying principle for influence, **Communications in Statistics, Part Theory and Methods**, Philadelphia, v. 17, p. 623-640, 1988.
- COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**, São Paulo, Edgard Blücher, 2006. 392 p.
- ESCOBAR, L. A.; MEEKER, W. Q. Assessing influence in regression analysis with censored data. **Biometrics**, Washington, v. 48, p. 507-528, 1992.
- FAREWELL, V. T. A model for a binary variable with time-censored observations, **Biometrika**, London, v. 64, p. 43-46, 1977.
- FAREWELL, V. T.; PRENTICE, R. L. A Study of distributional shape in life testing, **Technometrics**, Wisconsin, v. 19, p. 69-75, 1977.
- GALEA, M.; PAULA, G. A.; BOLFARINE, H. Local influence in elliptical linear regression models, **The Statistician**, Oxford, v. 46, p. 839-843, 1997.
- HAGER, H. W.; BRAIN, L. J. Inferential procedures for the generalized gamma distributions, **Journal of the American Statistical Association**, Alexandria, v. 65, p. 1601-1609, 1970.
- HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. New York: John Wiley, 1989. 307 p.
- IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. **Bayesian survival analysis**, New York: Springer, 2001. 479 p.
- LAWLESS, J. F. Inference in the generalized gamma and log gamma distributions, **Technometrics**, Wisconsin, v. 22, n. 3, p. 409-419, 1980.

LAWLESS, J. F. **Statistical models and methods for lifetime data**, 2nd ed, New York: Wiley, 2003. 630 p.

LEE, S. Y.; LU, B.; SONG, X. Y. Assessing local influence for nonlinear structural equation models with ignorable missing data, **Computational Statistics and Data Analysis**, New York, v. 50, p. 1356-1377, 2006.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, Washington, v. 55, p. 1281-1285, 1998.

MALLER, R.; ZHOU, X. **Survival analysis with long-term survivors**. New York, Wiley. 1996. 278 p.

MC CULLAGH, P.; NELDER, J. A. **Generalized linear models**, 2nd ed, London: Chapman and Hall, 1989. 511 p.

MILLER, R. G. **Survival analysis**, New York: Wiley, 1981. 238 p.

ORTEGA, E. M. M. **Análise de influência e resíduos em modelos de regressão log-gama generalizados**. 2001. 159p. Tese (Doutorado em Estatística) - Instituto de Matemática, Estatística da Universidade de São Paulo, 2001.

ORTEGA, E. M. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. **Computational Statistics e Data Analysis**, New York, v. 42, p. 165-186, 2003.

ORTEGA, E. M. M.; CANCHO, V. G.; PAULA, G. A. Influence diagnostics in exponentiated-Weibull regression models with censored data. **Statistics and Operation Reserch Transactions**, Espanha, v. 30, (to appear), 2006.

PETTITT, A. N.; BIN DAUD, I. Case-weight measures of influence for proportional hazards regression. **Applied Statistics**, New York, v. 38, p. 51-67, 1989.

PRENTICE, R. L. A log gamma model and its maximum likelihood estimation, **Biometrika**, London, v. 61, n. 3, p. 539-544, 1974.

STACY, E. W. A Generalization of the gamma distribution, **Annals of Mathematical Statistics**, Michigan, v. 33, p. 1187-1192, 1962.

THOMAS, W.; COOK, R. D. Assessing influence on predictions from generalized linear models. **Technometrics**, Wisconsin, v. 32, p. 59-65, 1990.

YAKOVLEV, A.; ASSELAIN, B.; BARDOU, V., FOURQUET, A. HOANG, T. ROCHEFEDIERE; TSODIKOV, A. A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer. **Biometrie et analyse de Données Spatio-Temporelles**, Paris, v. 12, p. 66-82, 1993.

YU, B.; TIWARI, R. C.; CRONIN, K. A.; FEUER, E. J. Cure fraction estimation from the mixture cure models for grouped survival data. **Statistics in Medicine**, New York, v. 23, p. 1733-1747, 2004.

APÊNDICE

Apêndice A

Aqui serão apresentadas as segundas derivadas que compõem a matriz de informação observada.

- $q > 0$

$$\begin{aligned} \mathbf{L}_{\mu\mu} = & -\frac{1}{q^2\sigma^2} \sum_{i:\delta_i=1} u_i + \\ & \sum_{i:\delta_i=0} \left\{ \frac{(1-p_i)q^2}{\sigma^2\Gamma(q^{-2})} u_i^{q^{-2}+1} \exp\{-u_i\} \left[\frac{q^{-2}+1-u_i}{h_{i1}} - \frac{(1-p_i)u_i^{q^{-2}+1} \exp\{-u_i\}}{\Gamma(q^{-2})h_{i1}^2} \right] \right\} \end{aligned}$$

$$\begin{aligned} \mathbf{L}_{\mu\sigma} = & \frac{r}{q\sigma^2} - \frac{q}{\sigma^2} \sum_{i:\delta_i=1} u_i - \frac{q^2}{\sigma^2} \sum_{i:\delta_i=1} u_i z_i + \\ & \sum_{i:\delta_i=0} \left\{ \frac{(1-p_i)q}{\sigma^2\Gamma(q^{-2})} u_i^{q^{-2}+1} \exp\{-u_i\} \right. \\ & \left. \left[\frac{(q^{-2}+1)qz_i - qu_i z_i - 1}{h_{i1}} - \frac{(1-p_i)qu_i^{q^{-2}+1} \exp\{-u_i\} z_i}{\Gamma(q^{-2})h_{i1}^2} \right] \right\} \end{aligned}$$

$$\begin{aligned} \mathbf{L}_{\sigma\sigma} = & \frac{r}{\sigma^2} + \frac{2}{q\sigma^2} \sum_{i:\delta_i=1} z_i - \frac{q^2}{\sigma^2} \sum_{i:\delta_i=1} z_i^2 u_i - \frac{2q}{\sigma^2} \sum_{i:\delta_i=1} z_i u_i + \\ & \sum_{i:\delta_i=0} \left\{ \frac{(1-p_i)q}{\sigma^2\Gamma(q^{-2})} u_i^{q^{-2}+1} \exp\{-u_i\} z_i \right. \\ & \left. \left[\frac{(q^{-2}+1)qz_i - qu_i z_i + 2}{h_{i1}} - \frac{(1-p_i)qu_i^{q^{-2}+1} \exp\{-u_i\} z_i}{\Gamma(q^{-2})h_{i1}^2} \right] \right\} \end{aligned}$$

$$\mathbf{L}_{\mu\beta} = \frac{q}{\sigma\Gamma(q^{-2})} \sum_{i:\delta_i=0} \frac{x_{ij}(u_i)^{q^{-2}+1} \exp\{-u_i\} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{h_{i1}^2 [\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]^2}$$

$$\mathbf{L}_{\sigma\beta} = \frac{q}{\sigma\Gamma(q^{-2})} \sum_{i:\delta_i=0} \frac{x_{ij}z_i(u_i)^{q^{-2}+1} \exp\{-u_i\} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{h_{i1}^2 [\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]^2}$$

$$\begin{aligned} \mathbf{L}_{\beta\beta} = & \sum_{i:\delta_i=1} -\frac{x_{ij}x_{is} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{(1-p_i)^2 [1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]^4} \left\{ (1-p_i) \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \right]^2 + \right. \\ & \left. \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} - 2(1-p_i) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \right] \right\} + \end{aligned}$$

$$\sum_{i:\delta_i=0} \frac{x_{ij}x_{is} \left[1 - Q(q^{-2}, u_i)\right] \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{h_{i1}^2 \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right]^4} \left\{ h_{i1} \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right]^2 - \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \left[1 - Q(q^{-2}, u_i)\right] - 2h_{i1} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right] \right\}$$

• $q < 0$

$$\begin{aligned} \mathbf{L}_{\mu\mu} &= \frac{1}{q^2 \sigma^2} \sum_{i:\delta_i=1} u_i \\ &+ \sum_{i:\delta_i=0} \left\{ \frac{(1-p_i)q^2}{\sigma^2 \Gamma(q^{-2})} u_i^{q^{-2}+1} \exp\{-u_i\} \left[\frac{-q^{-2} - 1 + u_i}{h_{i2}} - \frac{(1-p_i)u_i^{q^{-2}+1} \exp\{-u_i\}}{\Gamma(q^{-2})h_{i2}^2} \right] \right\} \end{aligned}$$

$$\begin{aligned} \mathbf{L}_{\mu\sigma} &= \frac{1}{q\sigma^2} - \frac{q}{\sigma^2} \sum_{i:\delta_i=1} u_i - \frac{q^2}{\sigma^2} \sum_{i:\delta_i=1} u_i z_i + \\ &\sum_{i:\delta_i=0} \left\{ \frac{(1-p_i)q}{\sigma^2 \Gamma(q^{-2})} u_i^{q^{-2}+1} \exp\{-u_i\} \right. \\ &\left. \left[\frac{-(q^{-2}+1)qz_i + qu_i z_i - 1}{h_{i2}} - \frac{(1-p_i)qu_i^{q^{-2}+1} \exp\{-u_i\} z_i}{\Gamma(q^{-2})h_{i2}^2} \right] \right\} \end{aligned}$$

$$\begin{aligned} \mathbf{L}_{\sigma\sigma} &= \frac{r}{\sigma^2} = \frac{2}{q\sigma^2} \sum_{i:\delta_i=1} z_i - \frac{q^2}{\sigma^2} z_i^2 u_i - \frac{2q}{\sigma^2} z_i u_i + \\ &\sum_{i:\delta_i=0} \left\{ \frac{(1-p_i)q}{\sigma^2 \Gamma(q^{-2})} u_i^{q^{-2}+1} \exp\{-u_i\} z_i \right. \\ &\left. \left[\frac{-(q^{-2}+1)qz_i + qu_i z_i - 2}{h_{i2}} - \frac{(1-p_i)qu_i^{q^{-2}+1} \exp\{-u_i\} z_i}{\Gamma(q^{-2})h_{i2}^2} \right] \right\} \end{aligned}$$

$$\mathbf{L}_{\mu\beta} = \frac{-q}{\sigma \Gamma(q^{-2})} \sum_{i:\delta_i=0} \frac{x_{ij}(u_i)^{q^{-2}+1} \exp\{-u_i\} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{h_{i2}^2 [\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]^2}$$

$$\mathbf{L}_{\sigma\beta} = \frac{-q}{\sigma \Gamma(q^{-2})} \sum_{i:\delta_i=0} \frac{x_{ij}z_i(u_i)^{q^{-2}+1} \exp\{-u_i\} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{h_{i2}^2 [\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]^2}$$

$$\begin{aligned} \mathbf{L}_{\beta\beta} &= \sum_{i:\delta_i=1} -\frac{x_{ij}x_{is} \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{(1-p_i)^2 [1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]^4} \left\{ (1-p_i) \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right]^2 + \right. \\ &\left. \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} - 2(1-p_i) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \left[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}\right] \right\} + \end{aligned}$$

$$\sum_{i:\delta_i=0} \frac{x_{ij}x_{is}Q(q^{-2}, u_i)\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{h_{i2}^{*2}\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]^4} \left\{ h_{i2}\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]^2 - \right. \\ \left. \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}Q(q^{-2}, u_i) + 2h_{i2}\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right] \right\}$$

• $q = 0$

$$\mathbf{L}_{\mu\mu} = -\frac{r}{\sigma^2} + \frac{1}{\sigma} \sum_{i:\delta_i=0} (1-p_i)\phi(z_i)h_{i3}^{-2}\left[z_i h_{i3} - (1-p_i)\phi(z_i)\right]$$

$$\mathbf{L}_{\mu\sigma} = -\frac{2}{\sigma^2} \sum_{i:\delta_i=1} z_i + \sum_{i:\delta_i=0} (1-p_i)h_{i3}^{-2}\phi(z_i)\sigma^{-2}\left\{\left[-2+z_i^2\right]h_{i3} - (1-p_i)z_i\phi(z_i)\right\}$$

$$\mathbf{L}_{\sigma\sigma} = \frac{r}{\sigma^2} - \frac{2}{\sigma} \sum_{i:\delta_i=1} z_i^2 + \\ \sum_{i:\delta_i=0} (1-p_i)h_{i3}^{-2}z_i^2\sigma^{-2}\left\{\left[-3z_i^{-1}\phi(z_i) + z_i\sigma^{-1}\exp\left\{-\frac{z_i^2}{2}\right\}\right]h_{i3} - [\phi(z_i)]^2\right\}$$

$$\mathbf{L}_{\mu\beta} = -\frac{1}{\sigma} \sum_{i:\delta_i=0} \frac{x_{ij}p_i\phi(z_i)}{[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}]} \left[h_{i3}^{-1} + (1-p_i)h_{i3}^{-2}\Phi(z_i)\right]$$

$$\mathbf{L}_{\sigma\beta} = -\frac{1}{\sigma} \sum_{i:\delta_i=0} \frac{x_{ij}p_i z_i \phi(z_i)}{[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}]} \left[h_{i3}^{-1} + (1-p_i)h_{i3}^{-2}\Phi(z_i)\right]$$

$$\mathbf{L}_{\beta\beta} = \sum_{i:\delta_i=1} (1-p_i)^{-2}\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]^{-4}\left[x_{ij}x_{is}\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\left\{(1-p_i)\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]^2 + \right. \right. \\ \left. \left. \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\} - 2(1-p_i)\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]\right\}\right] \\ + \sum_{i:\delta_i=0} \left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]^{-4}h_{i3}^{-2}\left[\Phi(z_i)x_{ij}x_{is}\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\left\{\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]^2 h_{i3} - \right. \right. \\ \left. \left. 2\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\left[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}\right]h_{i3} + \Phi(z_i)\right\}\right]$$

onde

$$h_{i1} = \left[p_i + (1-p_i)Q(q^{-2}, u_i)\right], \quad h_{i2} = \left[p_i + (1-p_i)\left\{1 - Q(q^{-2}, u_i)\right\}\right],$$

$$h_{i3} = \left[p_i + (1-p_i)\left\{1 - \Phi(z_i)\right\}\right], \quad u_i = q^{-2}\exp\{qz_i\}, \quad z_i = \frac{y_i - \mu}{\sigma}$$

$$p_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$$

$i = 1, 2, \dots, n$, $j, s = 0, 1, 2, \dots, p$, $\Phi(\cdot)$ denota a distribuição acumulada da normal padrão, $\phi(\cdot)$ função densidade da normal padrão e $Q(\cdot)$ é a função integral gama incompleta.

ANEXO

Anexo A

A seguir, serão apresentados os programas no SAS, proc NLMIXED, para a estimação dos parâmetros.

```

/*quando q é positivo*/
%macro nlmixed(q);
  proc nlmixed data=cancer;
    parms mi=0.8038 sigma=0.42783 beta0=0.69344 beta1=0.26576 beta2=-0.00201
    beta3=-0.13938 beta4=-0.11956 beta5=-0.08441 beta6=0.01458;
    bounds sigma>0;
    q=&q/10; *(0,2];
    xbeta=(x0*beta0)+(trat*beta1)+(idade*beta2)+(nodulo*beta3)+(sexo*beta4)+
    (perfor*beta5)+(breslow*beta6);
    p=exp(xbeta)/(1+exp(xbeta));
    a= (1-p)*(q/(sigma*GAMMA(q**(-2)) ) )*(q**(-2))*(q**(-2));
    b= q**(-1)*((tempo1-mi)/sigma);
    c= q**(-2)*exp(q*((tempo1-mi)/sigma));
    d= p+(1-p)*(1-(probgam( q**(-2)*exp(q*((tempo1-mi)/sigma)),q**(-2))));
    lv=censura*log(a*exp(b-c))+(1-censura)*log(d);
    model tempo1 ~ general(lv);
  run;
%mend;
options mtrace mprint;
%macro analisa();
  %do q = 2 %to 20 %by 2;
    %nlmixed(&q);
  %end;
%mend;

%analisa();

/*quando q é negativo*/
%macro nlmixed(q);
  proc nlmixed data=cancer;
    parms mi=0.8038 sigma=0.42783 beta0=0.69344 beta1=0.26576 beta2=-0.00201
    beta3=-0.13938 beta4=-0.11956 beta5=-0.08441 beta6=0.01458;
    bounds sigma>0;
    q=&q/(-10); *[-2,0);
    xbeta=(x0*beta0)+(trat*beta1)+(idade*beta2)+(nodulo*beta3)+(sexo*beta4)+
    (perfor*beta5)+(breslow*beta6);
    p=exp(xbeta)/(1+exp(xbeta));
    a= (1-p)*(-q/(sigma*GAMMA(q**(-2)) ) )*(q**(-2))*(q**(-2));

```

```

        b= q**(-1)*((tempo1-mi)/sigma);
        c= q**(-2)*exp(q*((tempo1-mi)/sigma));
        d= p+(1-p)*((probgam( q**(-2)*exp(q*((tempo1-mi)/sigma)),q**(-2))));
        lv=censura*log(a*exp(b-c))+(1-censura)*log(d);
        model tempo1 ~ general(lv);
    run;
%mend;
options mtrace mprint;
%macro analisa();
    %do q = 2 %to 20 %by 2;
        %nlmixed(&q);
    %end;
%mend;

%analisa();

```