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

Modelo de regressão gama-G em análise de sobrevivência

Elizabeth Mie Hashimoto

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Estatística e Experimen-
tação Agronômica

**Piracicaba
2013**

Elizabeth Mie Hashimoto
Bacharel em Estatística

Modelo de regressão gama-G em análise de sobrevivência

Orientador:

Prof. Dr. **EDWIN MOISES MARCOS ORTEGA**

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2013

DEDICATÓRIA

Aos meus queridos pais,

Agostinho Hashimoto e Eiko Kurihara Hashimoto,
fonte de toda minha sabedoria e base de todas as minhas conquistas. Otousan to okaasan, hontouni arigatou gozaimasu!

Ao meu irmão,

Fábio Jun Iti Hashimoto, *pela amizade, risadas, brigas e tudo mais que um grande e bom irmão pode fazer por uma irmã....*

Aos meus padrinhos,

Mauro Hitoshi Kubo e Meire Yayoi Kurihara Kubo.

A eles, dedico este trabalho com amor e carinho.

AGRADECIMENTOS

A Deus por sempre iluminar o meu caminho, para que eu pudesse prosseguir os meus estudos com sabedoria e perseverança.

Ao meu orientador Prof Edwin Moises Marcos Ortega, pela orientação, por acreditar no meu potencial acadêmico e por todo apoio que me permitiram crescer profissionalmente.

Aos meus pais Agostinho e Eiko Hashimoto, que me ensinaram a trabalhar com dedicação e por meio de um árduo trabalho, me deram condições de manter os meus estudos e motivação para terminar mais essa etapa. Ao meu irmão Jun, que esteve sempre ao lado dos nossos pais, os auxiliando, me propiciou desenvolver as minhas pesquisas, sem preocupações.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela concessão da bolsa de doutorado (Processo 2010/04496-2), que possibilitou a participação de eventos científicos nacionais e internacionais, a minha dedicação exclusiva às atividades de pós-graduação e a produção de artigos científicos.

Aos professores da banca de qualificação, Profa Dra Clarice Garcia Borges Demétrio, Prof Dr. Jorge Achcar e o Prof Dr. Gilberto Alvarenga Paula pelas sugestões e contribuições no desenvolvimento do trabalho.

A Carine Klauberg pela concessão dos dados de produção de óleo de copaíba utilizados no Capítulo 7 e pelas sugestões no desenvolvimento deste capítulo.

As secretárias do Departamento de Ciências Exatas, Luciane Brajão e Solange de Assis Paes Sabadin, aos auxiliares, Mayara Segatto e Rosni Honofre Aparecido Pinto e aos técnicos em informática, Eduardo Bonilha e Jorge Alexandre Wiendl, pela ajuda incessante e nos inúmeros momentos que pude contar com o auxílio e boa vontade de vocês.

Aos revisores anônimos da FAPESP por acreditarem e acompanharem o projeto.

Ao Prof Gauss Cordeiro, ao Prof Vicente Garibay Cancho e ao Prof Victor Hugo Lachos, por serem exemplo de referência na minha carreira acadêmica.

Aos professores do curso de Pós-graduação em Estatística e Experimentação Agrônômica pelos ensinamentos que contribuíram para minha formação acadêmica e em especial, ao Prof Décio Barbin pelas sugestões na escrita do trabalho.

A minha família de Campinas, em especial aos meus tios, Alécio Hashimoto, Emiko Hashimoto, Nelson Hirata e Shirley Hirata e ao meu primo/amigo, Gervásio Hirata, por todo o suporte e o acolhimento familiar durante esses seis últimos anos.

Aos meus professores de graduação da FCT/UNESP, Mario Hissamitsu Tarumoto e Olga Lyda Anglas Tarumoto, pelos conselhos e pelas sábias palavras de incentivo.

Aos amigos, Adriana Marques da Silva, Ana Julia Righetto, Ana Patrícia Peixoto, Andréia Meyer, Cris Rodrigues, Fabiane de Lima Silva, Fernanda Rizzato, Giovana Oliveira, Gláucia Ferrari, Izabela Oliveira, Juliana Fachini, Josi Rodrigues, Josi Padilha, Kuang Hongyu, Priscila Neves, Marcelino Pascoa, Mariana Urbano, Marisol Peña, Michele Barbosa, Renato Pereira, Ricardo Olinda, Sérgio Alarcón e Tiago Almeida, por sempre me incentivarem, aconselharem e por serem meu alicerce nos bons e maus momentos...

Aos amigos dos cursos de mestrado e doutorado do Programa de Pós-Graduação em Estatística e Experimentação Agronômica, em especial, Everton Batista, Guilherme Biz, Nilton Cruz, Pedro Cerqueira, Rodrigo Pescim, Simone Grego, Simone Sartorio e Veronika Rondinel a atenção e amizade.

Aos amigos de longa data..., André, Cíntia, Dani, Emely, Erik, Erika Saito, Gustavo, Hitoshi, Júlio Wada, Koiti, Leda, Marcela, Márcia, Márcio, Massumi, Missuzu, Paulinho, Renato, Rumi, Sônia Wada, Taciana, Takashi Moriya, Takashi Oyama, Takeshi, Thalita, Tiago, Wagner, Yoko, Yuri Wada e Yukie, pelas boas conversas e momentos de descontração.

As funcionárias da limpeza, Lu e Adriana pelo carinho e por sempre manterem o nosso lugar de estudo organizado.

As pessoas que contribuíram de alguma forma ou com um simples gesto para a realização do presente trabalho, pois na vida não conquistamos nada sozinhos e sempre precisamos do auxílio de outras pessoas para alcançar os nossos objetivos.

A todos vocês, “co-autores” deste trabalho, o meu muito obrigado!

Elizabeth Mie Hashimoto

“O problema com este mundo é que os tolos e os fanáticos estão sempre muito seguros de si mesmos, enquanto os mais sábios estão cheios de dúvidas”.

Bertrand Russell

“Você pode encarar um erro como uma besteira a ser esquecida, ou como um resultado que aponta uma nova direção”.

Steve Jobs

SUMÁRIO

RESUMO	11
ABSTRACT	13
1 INTRODUÇÃO	15
2 DISTRIBUIÇÕES DE PROBABILIDADE GAMA-G	19
2.1 Geradora de distribuição gama	19
2.2 Especificação do modelo gama-Weibull e gama-log-logística	20
2.2.1 Função quantílica	24
2.2.2 Momentos e momentos incompletos	26
2.2.3 Função geradora de momentos	31
2.2.4 Estimação de parâmetros por máxima verossimilhança	32
2.2.5 Estudo de simulação	34
2.3 Aplicação	36
3 MODELO DE REGRESSÃO LOCAÇÃO - ESCALA	41
3.1 Especificação do modelo log-gama-G	41
3.2 Estimação de parâmetros por máxima verossimilhança	45
3.3 Critério para seleção de modelos	47
3.4 Análise de sensibilidade	48
3.4.1 Influência global	49
3.4.2 Influência local	50
3.5 Análise de resíduos	53
3.5.1 Resíduo martingale	53
3.5.2 Resíduo componente do desvio	54
3.6 Estudo de simulação	55
3.7 Aplicação	61
4 MODELO DE REGRESSÃO PARA DADOS COM CENSURA INTERVALAR	69
4.1 Especificação do modelo log-gama-G	70
4.2 Estimação de parâmetros por máxima verossimilhança	72
4.3 Análise de sensibilidade	74
4.3.1 Influência global	74
4.3.2 Influência local	74
4.4 Análise de resíduos	78
4.4.1 Resíduo de martingale	78
4.4.2 Resíduo componente do desvio modificado	79
4.4.3 Estudo de simulação	80
4.5 Aplicação	83
5 MODELO DE LONGA-DURAÇÃO	91
5.1 Especificação dos modelos tempo-promoção	92

5.1.1 Modelo de tempo-promoção Poisson	92
5.1.2 Modelo de tempo-promoção geométrica	94
5.2 Estimação de parâmetros por máxima verossimilhança	95
5.3 Análise de sensibilidade	97
5.4 Aplicação	101
6 MODELO DE REGRESSÃO COM EFEITO ALEATÓRIO	111
6.1 Especificação do modelo de regressão com efeito aleatório log-gama-G	112
6.2 Estimação por máxima verossimilhança	113
6.3 Análise de sensibilidade	116
6.3.1 Influência global	117
6.3.2 Influência local	117
6.4 Aplicação	120
7 MODELO GAMA-G INFLACIONADO DE ZEROS	127
7.1 Especificação do modelo inflacionado de zeros	128
7.2 Estimação por máxima verossimilhança	129
7.3 Análise de sensibilidade	131
7.4 Aplicação	134
8 CONSIDERAÇÕES FINAIS	141
8.1 Pesquisas futuras	141
REFERÊNCIAS	143
APÊNDICES	153

RESUMO

Modelo de regressão gama-G em análise de sobrevivência

Dados de tempo de falha são caracterizados pela presença de censuras, que são observações que não foram acompanhadas até a ocorrência de um evento de interesse. Para estudar o comportamento de dados com essa natureza, distribuições de probabilidade são utilizadas. Além disso, é comum se ter uma ou mais variáveis explicativas associadas aos tempos de falha. Dessa forma, o objetivo geral do presente trabalho é propor duas novas distribuições utilizando a função geradora de distribuições gama, no contexto de modelos de regressão em análise de sobrevivência. Essa função possui um parâmetro de forma que permite criar famílias paramétricas de distribuições que sejam flexíveis para capturar uma ampla variedade de comportamentos simétricos e assimétricos. Assim, a distribuição Weibull e a distribuição log-logística foram modificadas, dando origem a duas novas distribuições de probabilidade, denominadas de gama-Weibull e gama-log-logística, respectivamente. Consequentemente, os modelos de regressão locação-escala, de longa-duração e com efeito aleatório foram estudados, considerando as novas distribuições de probabilidade. Para cada um dos modelos propostos, foi utilizado o método da máxima verossimilhança para estimar os parâmetros e algumas medidas de diagnóstico de influência global e local foram calculadas para encontrar possíveis pontos influentes. No entanto, os resíduos foram propostos apenas para os modelos locação-escala para dados com censura à direita e para dados com censura intervalar, bem um estudo de simulação para verificar a distribuição empírica dos resíduos. Outra questão explorada é a introdução dos modelos: gama-Weibull inflacionado de zeros e gama-log-logística inflacionado de zeros, para analisar dados de produção de óleo de copaíba. Por fim, diferentes conjunto de dados foram utilizados para ilustrar a aplicação de cada um dos modelos propostos.

Palavras-chave: Distribuição gama-G; Dados censurados; Modelo de longa-duração; Efeito aleatório; Função de verossimilhança marginal; Modelo inflacionado de zeros; Análise de diagnóstico

ABSTRACT

Gama-G regression model in survival analysis

Failure time data are characterized by the presence of censoring, which are observations that were not followed up until the occurrence of an event of interest. To study the behavior of the data of that nature, probability distributions are used. Furthermore, it is common to have one or more explanatory variables associated to failure times. Thus, the goal of this work is given to the generating of gamma distributions function in the context of regression models in survival analysis. This function has a shape parameter that allows create parametric families of distributions that are flexible to capture a wide variety of symmetrical and asymmetrical behaviors. Therefore, through the generating of gamma distributions function, the Weibull distribution and log-logistic distribution were modified to give two new probability distributions: gamma-Weibull and gamma-log-logistic. Additionally, location-scale regression models, long-term models and models with random effects were also studied, considering the new distributions. For each of the proposed models, we used the maximum likelihood method to estimate the parameters and some diagnostic measures of global and local influence were calculated for possible influential points. However, residuals have been proposed for data with right censoring and interval-censored data and a simulation study to verify the empirical distribution of the residuals. Another issue explored is the introduction of models: gamma-Weibull inflated zeros and gamma-log-logistic inflated zeros, to analyze production data copaiba oil. Finally, different data set are used to illustrate the application of each of the models.

Keywords: Gama-G distribution; Censored data; Long-term model; Random effect; Marginal likelihood; Zero inflated model; Diagnostic analysis

1 INTRODUÇÃO

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas décadas. A razão desse crescimento é a utilização de técnicas estatísticas em conjunto com computadores cada vez mais eficientes que permitem conduzir uma análise, levando em conta a informação de observações censuradas (COLOSIMO; GIOLO, 2006).

A presença das censuras é a principal característica de um conjunto de dados de tempo de falha. Nesse caso, assume-se que a variável resposta, denominada de tempo de falha, é o período até a ocorrência de um evento de interesse. Por outro lado, entende-se que as observações censuradas correspondem às unidades amostrais, nas quais o evento de interesse não ocorreu e por isso, somente uma informação parcial sobre o tempo é observado (KALBFLEISH; PRENTICE, 2002).

Além disso, é comum se ter uma ou mais covariáveis que estejam associadas à variável resposta, a fim de explicar parte da heterogeneidade presente nos tempos de falha. Por exemplo, em estudos de educação, é interessante verificar se a idade dos alunos de pós-graduação exerce alguma influência no tempo de titulação (HORTA; MORAES, 2005). Assim sendo, para estudar o efeito das covariáveis sobre a variável resposta, deve-se utilizar um modelo de regressão apropriado para dados censurados (SILVA et al., 2008).

Na literatura, encontra-se diversas metodologias para analisar dados com covariáveis. Entretanto, os modelos de regressão sob o enfoque paramétrico tem sido muito utilizados, pois complementam as técnicas não-paramétricas e permitem descrever o comportamento do tempo observado por meio de funções, bem como realizar estudos de simulações (MEEKER; ESCOBAR, 1998).

Dessa forma, os modelos paramétricos assumem que os dados seguem uma distribuição de probabilidade. No contexto análise de sobrevivência, as distribuições clássicas associadas à variável resposta podem ser: a distribuição exponencial, a distribuição Weibull, a distribuição log-normal ou a distribuição log-logística (LAWLESS, 2003).

Contudo, em algumas situações práticas, Aarset (1987) e Mudholkar, Srivastava e Kollia (1996) verificaram que a função de taxa de falha pode assumir formas não-monótonas, ou seja, na forma de banheira ou unimodal. Nesse caso, a distribuição Weibull e a distribuição exponencial não podem ser utilizadas, pois não apresentam essa característica. Por outro lado, a distribuição log-normal e a distribuição log-logística são apropriadas apenas em situações nas quais a função taxa de falha é unimodal.

Por essa razão, generalizar distribuições de probabilidade contínuas tem sido a proposta na literatura, com a finalidade de fornecer maior flexibilidade às distribuições transformadas e permitir que a função de taxa de falha possa descrever formas do tipo monótona, unimodal e de banheira. Além disso, a distribuição modificada pode ser utilizada para discriminação de modelos, pois possui como caso particular a distribuição a qual lhe deu origem.

Na literatura, encontram-se três formas de se transformar uma distribuição de probabilidade, são elas: a função geradora de distribuições beta generalizada, a função geradora de distribuições Kumaraswamy e a função geradora de distribuições McDonald. No entanto, essas funções geradoras de distribuições possuem de dois à três parâmetros que permitem flexibilizar as formas da função densidade de probabilidade da distribuição a ser modificada. Logo, dependendo da escolha da distribuição a ser generalizada, o novo modelo formulado pode apresentar expressões que sejam difíceis de serem tratadas ou apresentar complicações na obtenção das estimativas de máxima verossimilhança (SILVA; ORTEGA; CORDEIRO, 2010).

Em virtude do que foi exposto, existe a necessidade do estudo de outras formas de transformar distribuições de probabilidade, com a finalidade de torná-las mais flexíveis para capturar uma ampla variedade de comportamentos simétricos e assimétricos e ao mesmo tempo, que sejam utilizados poucos parâmetros para esta finalidade.

Outra contribuição esperada deste trabalho é o desenvolvimento de uma distribuição útil em aplicações na qual a variável resposta observada é não-negativa. Por exemplo, em estudos de tempo de titulação de alunos de pós-graduação, de seleção genética (GIOLO et al., 2004), de preparo de dietas para vacas leiteiras e de impactos ambientais e econômicos (Processo FAPESP 2011/04968-4).

Portanto, o objetivo deste trabalho é o estudo da função geradora de distribuições gama proposta por Zografos e Balakrishnan (2009), que modifica modelos clássicos de análise de sobrevivência. A vantagem dessa função geradora de distribuições é que ela possui apenas um parâmetro para transformar uma distribuição conhecida.

Mais especificamente, será considerada a distribuição Weibull e a distribuição log-logística para serem transformadas, pois são muito utilizadas na análise de sobrevivência e na literatura para serem modificadas (ABRAMS; ASHBY; ERRINGTON, 1996, LEUCOTRE; MABIKA; DERZKO, 2002, COLOSIMO; GIOLO, 2006, ASHKAR; AUCOIN, 2012). Entretanto, os resultados dessa pesquisa, também, poderão ser estendidos para outras distribuições de probabilidade.

Além do estudo da distribuição, pretende-se explorar os principais modelos de regressão paramétricos utilizados na análise de sobrevivência. Os modelos abordados neste estudo, serão: o modelo de locação-escala sob o enfoque de censura aleatória e sob o enfoque de censura intervalar, o modelo de tempo-promoção para analisar dados que se caracterizam pela presença de não são suscetíveis ao evento de interesse e o modelo com efeito aleatório, no qual os tempos são agrupados e dentro de cada grupo existe uma correlação entre as observações. Enfim, não necessariamente em uma área de análise de sobrevivência, mas em uma situação prática da área de floresta, estudar modelos inflacionados de zeros que possam ser utilizados para analisar dados contínuos com excesso de zeros.

Desse modo, o presente trabalho está organizado da seguinte maneira. No Capítulo 2, duas novas distribuições de probabilidade são obtidas a partir da função geradora de distribuições gama. Algumas caracterizações das distribuições de probabilidade, também, são

apresentadas. No Capítulo 3, são apresentados os modelos de regressão log-gama-Weibull e log-gama-logística, assim como um estudo de simulação para avaliar os resíduos dos novos modelos. No Capítulo 4, os modelos de locação-escala são novamente abordados, porém no contexto de censura intervalar, para verificar a eficiência de uma determinada dieta no tempo de ovulação de vacas leiteiras. No Capítulo 5, são apresentados os modelos de tempo-promoção, os quais, assumem que o número de causas responsáveis pela ativação do evento de interesse segue uma distribuição Poisson ou uma distribuição geométrica. No Capítulo 6, são introduzidos os modelos de regressão com efeito aleatório, supondo que a distribuição de probabilidade desses efeitos segue uma das distribuições propostas no Capítulo 3. O método de integração numérica, Gauss-Hermite, foi utilizado para maximizar a função de verossimilhança marginal desses modelos e para encontrar algumas medidas de influência global e local. No Capítulo 7, a idéia de modelos inflacionados de zeros é estendido para as distribuições contínuas, gama-Weibull e gama-logística. Esses modelos são utilizados para analisar dados de produção de óleo de copaíba. Por fim, no Capítulo 8, são apresentadas as principais considerações deste trabalho, bem como, algumas sugestões para pesquisas futuras, que podem ser desenvolvidas para dar continuidade ao estudo.

2 DISTRIBUIÇÕES DE PROBABILIDADE GAMA-G

2.1 Geradora de distribuição gama

Generalizações de distribuições de probabilidade contínuas têm sido muito estudadas nas últimas décadas. Uma das razões para generalizar uma distribuição conhecida é porque a sua forma generalizada flexibiliza a distribuição de tal forma que possa acomodar formas da função de taxa de falha que sejam não-monótonas (PASCOA et al., 2011). Por essa razão, uma linha de pesquisa importante na área de análise de sobrevivência é propor novas distribuições que permitam caracterizar dados de sobrevivência com função de taxa de falha monótona, unimodal e na forma de banheira (SILVA; ORTEGA; CORDEIRO, 2010). Desde então, inúmeros trabalhos com novas distribuições de probabilidade vem sendo desenvolvidos, por exemplo, Eugene, Lee e Famoye (2002), Nadarajah e Kotz (2006), Shawky e Abu-Zinadah (2008) e Pescim et al. (2011).

Na literatura encontram-se três formas de generalizar uma distribuição de probabilidade conhecida. Utilizando a função geradora de distribuição beta (EUGENE; LEE; FAMOYE, 2002) ou a função geradora de distribuição Kumaraswamy (CORDEIRO; de CASTRO, 2011) ou a função geradora de distribuição McDonald (ALEXANDER et al., 2012). A partir dessas funções geradoras de distribuições, várias distribuições têm sido generalizadas e a adequação dos modelos de probabilidade pode ser verificada por meio de critérios de seleção de modelos, comparando as curvas de sobrevivência obtidas pelo modelo paramétrico com a curva de sobrevivência de Kaplan-Meier (FACHINI; ORTEGA; LOUZADA-NETO, 2008) ou comparando a densidade da distribuição com o histograma obtido a partir dos dados (BARRETO-SOUZA; CRIBARI-NETO, 2009).

Entretanto, essas funções geradoras de distribuições apresentam entre dois a três parâmetros de forma, que auxilia na flexibilização do modelo, porém dificulta a obtenção das estimativas dos parâmetros que utilizam recursos numéricos para sua obtenção. Além disso, dependendo da distribuição a ser modificada, pode gerar uma distribuição com número de parâmetros elevado.

Por essa razão, pensando em desenvolver uma função geradora de distribuições utilizando apenas um parâmetro para modificar uma distribuição conhecida, Zografos e Balakrishnan (2009) apresentaram a função geradora de distribuição gama. Essa função geradora de distribuição de probabilidade é definida da seguinte forma, se $G(x)$ é uma função de distribuição acumulada de uma variável aleatória contínua, então, sua função de distribuição acumulada da função geradora de distribuição gama é dada por

$$F(x; \phi) = \frac{1}{\Gamma(\phi)} \gamma\{-\log[1 - G(x)], \phi\}, \quad (2.1)$$

em que $\gamma(w, \phi) = \int_0^w e^{-u} u^{\phi-1} du$ é a função gama incompleta (JOHNSON; KEMP; KOTZ,

2005).

Dessa forma, derivando a equação (2.1), obtém-se a sua função densidade de probabilidade expressa como

$$f(x; \phi) = \frac{g(x)}{\Gamma(\phi)} \{ -\log [1 - G(x)] \}^{\phi-1}, \quad (2.2)$$

em que $g(x) = dG(x)/dx$, $\phi > 0$ é o parâmetro de forma que flexibiliza a distribuição a ser modificada e o suporte de $f(x; \phi)$ é definido por $g(x)$. Observa-se que diferente das outras funções geradoras de distribuição encontradas na literatura, a função $f(x)$ apresenta expressões analíticas mais simples.

A função geradora de distribuição gama modifica qualquer distribuição de probabilidade contínua com função de distribuição acumulada $G(x)$ e a nova distribuição é denominada de gama-G, pois G identifica a distribuição de probabilidade que foi modificada e o termo gama indica a função geradora de distribuições proposta por Zografos e Balakrishnan (2009). Algumas propriedades dessa função geradora de distribuições são definidas a seguir.

- i. $E(X^r) = \int_0^1 [F^{-1}(y)]^r \frac{1}{\Gamma(\phi)} \{ -\ln[1 - F(x)] \}^{\phi-1} dy$, em que $y = F(x)$.
- ii. $E[\ln\{-\ln[1 - F(x)]\}] = \psi(\phi)$, em que $\psi(\cdot)$ denota a função digama, ou seja, a derivada do logaritmo da função gama (JOHNSON; KEMP; KOTZ, 2005).
- iii. $E[\ln\{f(x)\}] = E_Z[\ln\{f\{F^{-1}[1 - \exp(-Z)]\}\}]$, em que Z segue uma distribuição gama de um parâmetro, denotada por $Ga(\phi, 1)$, com função densidade de probabilidade dada por

$$f(z) = \frac{1}{\Gamma(\phi)} z^{\phi-1} \exp(-z), \quad z > 0$$

e $\phi > 0$ é o parâmetro de forma.

2.2 Especificação do modelo gama-Weibull e gama-log-logística

Na literatura estatística, mais especificamente, no contexto de análise de dados de sobrevivência, existe uma série de modelos probabilísticos que podem ser modificados. Entretanto, para este estudo, a distribuição Weibull e a distribuição log-logística foram consideradas para serem modificadas por meio da função geradora de distribuição gama, pois são distribuições muito utilizadas nas áreas médicas e de engenharia, como pode ser visto em, Kalbfleish e Prentice (2002), Collett (2003), Klein e Moeschberger (2003) e Lawless (2003). Porém, apesar dessas distribuições ocuparem uma posição de destaque na análise de dados, em algumas situações práticas elas podem não ser adequadas, como por exemplo, no caso em que o comportamento da taxa de falha dos dados seja na forma de banheira. A seguir, são apresentadas mais algumas razões para a escolha dessas duas distribuições, bem como a obtenção dos novos modelos por meio da função geradora de distribuição gama.

a. Distribuição gama-Weibull

Nas última décadas, vários estudos envolvendo a distribuição Weibull podem ser encontrados na literatura, dentre esses trabalhos, pode-se destacar o modelo de regressão com cópulas para analisar a dependência entre os tempos de vida bivariados sob o enfoque bayesiano apresentado por Romeo, Tanaka e Lima (2006). Giménez, Colosimo e Bolfarine (2006) consideraram a distribuição Weibull com erros de medidas. Klein (2008) caracterizou a resistência de tração de placa de alumínio utilizando a distribuição Weibull. Englehardt e Li (2011) utilizaram uma distribuição Weibull discreta para analisar a contagem microbiana em águas.

Dessa forma, a função densidade de probabilidade e a função de distribuição acumulada da distribuição Weibull, proposta por Weibull (1951), são expressões definidas por

$$g(x) = \frac{\alpha}{\lambda^\alpha} x^{\alpha-1} \exp \left[- \left(\frac{x}{\lambda} \right)^\alpha \right] \quad \text{e} \quad G(x) = 1 - \exp \left[- \left(\frac{x}{\lambda} \right)^\alpha \right], \quad (2.3)$$

respectivamente, em que $\lambda > 0$ é o parâmetro de escala e $\alpha > 0$ é o parâmetro de forma.

A sua popularidade em aplicações práticas se deve ao fato de possuir uma função de sobrevivência e uma função de taxa de falha com expressões matematicamente simples (COLOSIMO; GIOLO, 2006). Porém, a distribuição Weibull é ideal apenas em situações na qual a função de taxa de falha é monótona. Por essa razão, assim como Eugene, Lee e Famoye (2002), a distribuição Weibull será modificada para que o acréscimo de um parâmetro de forma possa permitir que a característica dos dados possa ser melhor representada pela distribuição de probabilidade.

Assim, como a distribuição Weibull está sendo considerada, a nova distribuição de probabilidade é denominada de gama-Weibull e é obtida substituindo as expressões definidas em (2.3) na função geradora de distribuições gama (2.2). Portanto, a função densidade de probabilidade da distribuição gama-Weibull é expressa por

$$f(x) = \frac{\alpha}{\lambda^\phi \Gamma(\phi)} x^{\phi-1} \exp \left[- \left(\frac{x}{\lambda} \right)^\alpha \right], \quad x > 0, \quad (2.4)$$

em que $\phi > 0$ e $\alpha > 0$ são parâmetros de forma e $\lambda > 0$ é o parâmetro de escala. Observou-se também, que a nova distribuição em (2.4) representa a função densidade de probabilidade da distribuição gama generalizada proposta por Stacy (1962). Dessa forma, combinando a função geradora de distribuição gama com a distribuição Weibull é possível demonstrar como Stacy (1962) obteve a função densidade de probabilidade da distribuição gama generalizada.

Em seguida, utilizando as condições de existência de uma função densidade de probabilidade definidas por Mood, Graybill e Boes (1974), verifica-se que a função (2.4) da distribuição gama-Weibull é uma função densidade de probabilidade, pois satisfaz as seguintes condições

- i. $f(x) \geq 0$ para todo x , como mostra a Figura 2.1a.
- ii. $\int_{-\infty}^{\infty} f(x)dx = \frac{\alpha}{\lambda^{\phi}\Gamma(\phi)} \int_0^{\infty} x^{\phi\alpha-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right] dx = 1$.

Por outro lado, a função de distribuição acumulada da nova distribuição pode ser obtida integrando a função densidade de probabilidade (2.4) ou por meio da substituição da função de distribuição acumulada em (2.3) na expressão (2.1). Dessa forma, a função de distribuição acumulada e a função de sobrevivência da distribuição gama-Weibull são

$$F(x) = \frac{1}{\Gamma(\phi)} \gamma\left[\left(\frac{x}{\lambda}\right)^{\alpha}, \phi\right] \quad \text{e} \quad S(x) = 1 - F(x) = 1 - \frac{1}{\Gamma(\phi)} \gamma\left[\left(\frac{x}{\lambda}\right)^{\alpha}, \phi\right], \quad (2.5)$$

respectivamente.

A função de taxa de falha é expressa da seguinte forma

$$h(x) = \frac{\alpha}{\lambda^{\phi}\Gamma(\phi)} \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right] \left\{1 - \frac{1}{\Gamma(\phi)} \gamma\left[\left(\frac{x}{\lambda}\right)^{\alpha}, \phi\right]\right\}^{-1}.$$

A Figura 2.1a ilustra algumas possíveis formas da função densidade de probabilidade (2.4) para determinados valores dos parâmetros. Uma característica da distribuição gama-Weibull é que sua função de taxa de falha pode ser monótona, em forma de banheira ou unimodal dependendo dos valores dos parâmetros, como mostra a Figura 2.1c. Por outro lado, é possível observar que quando os valores do parâmetro ϕ aumentam, a função de sobrevivência também aumenta, para $\alpha = 2$ e $\lambda = 1$ (Figura 2.1b).

Figura 2.1 – Gráficos da distribuição gama-Weibull para alguns valores de ϕ , λ e α . (a) Função densidade de probabilidade, (b) Função de sobrevivência e (c) Função de taxa de falha

b. Distribuição gama-log-logística

A distribuição log-logística foi apresentada por Bennett (1983) como uma alternativa à distribuição Weibull (COLOSIMO; GIOLO, 2006). Na literatura pode-se encontrar inúmeros trabalhos envolvendo essa distribuição, por exemplo, Shoukri, Mian e Tracy (1988)

estudaram a distribuição log-logística para descrever os fenômenos que envolvem precipitação na região do Canadá. Shayan, Ayatollahi e Zare (2011) propuseram uma distribuição log-logística com quatro parâmetros, o que permite a nova distribuição descrever dados que tenham como característica função de taxa de falha na forma de banheira. Ashkar e Aucoin (2012) realizaram um estudo de comparação entre a distribuição log-normal e a distribuição log-logística para analisar o fenômeno que descreve a movimentação das águas no planeta. Santana et al. (2012) apresentaram uma generalização da distribuição log-logística e suas propriedades. Zare, Doostfatemeheh e Rezaianzadeh (2012) utilizaram a distribuição log-logística para analisar dados de pacientes com câncer de mama na região sul do Irã.

Assim, a função densidade de probabilidade e a função de distribuição acumulada da distribuição log-logística, segundo Colosimo e Giolo (2006), são definidas por

$$g(x) = \frac{\alpha}{\lambda^\alpha} x^{\alpha-1} \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-2} \quad \text{e} \quad G(x) = 1 - \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-1}, \quad (2.6)$$

respectivamente, em que $\lambda > 0$ é o parâmetro de escala e $\alpha > 0$ é o parâmetro de forma.

Por meio dos trabalhos citados anteriormente, é possível observar que a distribuição log-logística tem sido bastante utilizada na análise de dados para descrever fenômenos naturais e prever os eventos futuros, como é o caso da movimentação das águas no oceano (ASHKAR; AUCCOIN, 2012). Entretanto, em algumas situações, os dados observados podem apresentar um comportamento incomum e dessa forma, a distribuição log-logística não seria adequada para descrever e prever o fenômeno em estudo.

Por esse motivo, da mesma forma que a distribuição Weibull, a distribuição log-logística será modificada utilizando a função geradora de distribuição gama, com o objetivo de torná-la flexível para descrever o comportamento dos dados observados.

Assim, como a distribuição log-logística está sendo considerada, a nova distribuição de probabilidade é denominada de gama-log-logística e é obtida substituindo as expressões definidas em (2.6) na função geradora de distribuições gama (2.2). Portanto, a função densidade de probabilidade da distribuição gama-log-logística é dada por

$$f(x) = \frac{\alpha}{\lambda^\alpha \Gamma(\phi)} x^{\alpha-1} \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-2} \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right] \right\}^{\phi-1}, \quad x > 0, \quad (2.7)$$

em que $\phi > 0$ e $\alpha > 0$ são os parâmetros de forma e $\lambda > 0$ é o parâmetro de escala.

Verifica-se também que a distribuição gama-log-logística é uma função densidade de probabilidade, pois satisfaz as seguintes condições

- i. $f(x) \geq 0$ para todo x , como mostra a Figura 2.2a.
- ii. $\int_{-\infty}^{\infty} f(x) dx = \frac{\alpha}{\lambda^\alpha \Gamma(\phi)} \int_0^{\infty} x^{\alpha-1} \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-2} \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right] \right\}^{\phi-1} dx = 1.$

A função de distribuição acumulada da nova distribuição, pode ser obtida integrando a função densidade de probabilidade (2.7) ou por meio da substituição da função

de distribuição acumulada da log-logística (2.6) na expressão (2.1). Assim, a função de distribuição acumulada e a função de sobrevivência da distribuição gama-log-logística são dadas por

$$F(x) = \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right], \phi \right\} \quad \text{e} \quad S(x) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right], \phi \right\} \quad (2.8)$$

respectivamente.

Por outro lado, a função de taxa de falha é expressa da seguinte forma

$$h(x) = \frac{\alpha}{\lambda^\alpha} x^{\alpha-1} \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-2} \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right] \right\}^{\phi-1} \gamma^{-1} \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right], \phi \right\}.$$

A Figura 2.2a ilustra algumas possíveis formas da função densidade de probabilidade (2.7) para determinados valores dos parâmetros. Por meio da Figura 2.2b, observa-se que, para o parâmetro $\lambda = 1$, quando o valor parâmetro ϕ aumenta, a curva da função de sobrevivência também aumenta. Entretanto, quando o valor do parâmetro α aumenta, a curva da função de sobrevivência cresce até o ponto $x = 1$ e depois decresce. Por outro lado, a função de taxa de falha apresenta o comportamento unimodal e crescente (Figura 2.2c).

Figura 2.2 – Gráficos da distribuição gama-log-logística para alguns valores de ϕ , λ e α . (a) Função densidade de probabilidade, (b) Função de sobrevivência e (c) Função de taxa de falha

2.2.1 Função quantílica

Segundo Zografos e Balakrishnan (2009), se X é uma variável aleatória contínua com função densidade de probabilidade (2.2), então, a variável aleatória $Z = -\log[1 - F(x)]$ segue uma distribuição $Ga(\phi, 1)$. Logo, a variável aleatória $X = F^{-1}(1 - e^{-Z})$ segue uma função geradora de distribuições gama.

Dessa forma, a função quantílica das duas novas distribuição são obtidas da seguinte forma:

i. Distribuição gama-Weibull

A função quantílica da distribuição gama-Weibull é dada por

$$x = Q(u) = \lambda Z^{1/\alpha}, \quad (2.9)$$

em que $Z = Q_\phi(u)$ denota a função quantílica da gama com parâmetro ϕ .

ii. Distribuição gama-log-logística

A função quantílica da distribuição gama-log-logística é dada por

$$x = Q(u) = \lambda [\exp(Z) - 1]^{1/\alpha}, \quad (2.10)$$

em que $Z = Q_\phi(u)$ denota a função quantílica da gama com parâmetro ϕ .

Essas funções podem ser utilizadas em estudos de simulação via método de Monte Carlo para gerar amostras provenientes de uma dada distribuição de probabilidade. Uma outra aplicação da função quantílica é para o cálculo da assimetria de Bowley apresentada por Kenney e Keeping (1962) e definida por

$$B = \frac{Q(3/4) - 2Q(1/2) + Q(1/4)}{Q(3/4) - Q(1/4)}$$

e a curtose de Moors apresentada por Moors (1988) dada pela seguinte quantidade

$$M = \frac{Q(7/8) - Q(5/8) - Q(3/8) + Q(1/8)}{Q(6/8) - Q(2/8)},$$

em que $Q(\cdot)$ é uma função quantílica.

Essas medidas são menos sensíveis à presença de observações discrepantes e existem para distribuições na qual não é possível obter os momentos. Para a distribuição t de Student com 10 e 5 graus de liberdade, essas medidas são zero (Bowley) e entre 1,277 a 1,327 (Moors), respectivamente. Para a distribuição normal padrão, essas medidas são zero (Bowley) e 1,2331 (Moors) (Alexander et al., 2011).

A Figura 2.3 mostra as representações gráficas da assimetria de Bowley (Figura 2.3a) e da curtose de Moors (Figura 2.3b) como função do parâmetro α para alguns valores do parâmetro ϕ e para $\lambda = 3, 0$. Observa-se que os valores da assimetria de Bowley crescem quando os valores do parâmetro α aumentam e os valores da curtose de Moors diminuem quando os valores do parâmetro α aumentam.

Por outro lado, considerando $\lambda = 3, 0$; a assimetria de Bowley e a curtose de Moors da distribuição gama-log-logística foram obtidas como função do parâmetro α para alguns valores do parâmetro ϕ . Essas medidas apresentam o mesmo comportamento da distribuição gama-Weibull, porém o decaimento da assimetria e da curtose da distribuição gama-log-logística é mais suave em relação à distribuição gama-Weibull, como ilustra a Figura 2.4.

Figura 2.3 – Assimetria e curtose da distribuição gama-Weibull como função do parâmetro α , para alguns valores de ϕ e $\lambda = 3$

Figura 2.4 – Assimetria e curtose da distribuição gama-log-logística como função do parâmetro α , para alguns valores de ϕ e $\lambda = 3$

2.2.2 Momentos e momentos incompletos

O r -ésimo momento de uma variável aleatória contínua, apresentado por Johnson, Kemp e Kotz (2005), é expresso por

$$\mu'_r = E(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx, \quad (2.11)$$

em que r é um número inteiro e positivo e $f(x)$ é a função densidade de probabilidade da variável aleatória X . Dessa forma, os momentos das duas novas distribuições são determinados por:

i. Distribuição gama-Weibull

Para a distribuição gama-Weibull, substituindo a $f(x)$ definida em (2.4) na expressão (2.11), tem-se

$$\mu'_r = \frac{\alpha}{\lambda^{\alpha\phi}\Gamma(\phi)} \int_0^\infty x^{r+\alpha\phi-1} \exp\left[-\left(\frac{x}{\lambda}\right)^\alpha\right] dx.$$

Portanto, após algumas mudanças de variáveis, o r -ésimo momento da distribuição gama-Weibull é dado por

$$\mu'_r = \frac{\lambda^r}{\Gamma(\phi)} \Gamma(r\alpha^{-1} + \phi). \quad (2.12)$$

A partir do r -ésimo momento (2.12) é possível determinar a média e a variância da distribuição gama-Weibull. Dessa forma, essas medidas são expressas, respectivamente, por

$$E(X) = \mu'_1 = \frac{\lambda}{\Gamma(\phi)} \Gamma(\alpha^{-1} + \phi)$$

e

$$\text{var}(X) = E(X^2) - E^2(X) = \frac{\lambda^2}{\Gamma(\phi)} \left[\Gamma(2\alpha^{-1} + \phi) - \frac{\Gamma(\alpha^{-1} + \phi)^2}{\Gamma(\phi)} \right].$$

ii. Distribuição gama-log-logística

Para a distribuição gama-log-logística, substituindo a $f(x)$ definida em (2.7) na expressão (2.11), tem-se

$$\mu'_r = \frac{\alpha}{\lambda^\alpha \Gamma(\phi)} \int_0^\infty x^r x^{\alpha-1} \left[1 + \left(\frac{x}{\lambda}\right)^\alpha\right]^{-2} \left\{ \log \left[1 + \left(\frac{x}{\lambda}\right)^\alpha\right] \right\}^{\phi-1} dx.$$

Portanto, após algumas mudanças de variáveis, o r -ésimo momento da distribuição gama-log-logística é dado por

$$\mu'_r = \frac{\lambda^r}{\Gamma(\phi)} \int_0^1 \left(\frac{1-u}{u}\right)^{r/\alpha} \left[\log \left(\frac{1}{u}\right) \right]^{\phi-1} du, \quad (2.13)$$

em que $u = \left[1 + \left(\frac{x}{\lambda}\right)^\alpha\right]^{-1}$.

A partir do r -ésimo momento (2.13) é possível determinar a média e a variância da distribuição gama-log-logística. Essas medidas são, respectivamente, expressas por

$$E(X) = \mu'_1 = \frac{\lambda}{\Gamma(\phi)} \int_0^1 \left(\frac{1-u}{u}\right)^{1/\alpha} \left[\log \left(\frac{1}{u}\right) \right]^{\phi-1} du$$

e

$$\begin{aligned} \text{var}(X) = E(X^2) - E^2(X) = & \frac{\lambda^2}{\Gamma(\phi)} \left\{ \int_0^1 \left(\frac{1-u}{u}\right)^{2/\alpha} \left[\log \left(\frac{1}{u}\right) \right]^{\phi-1} du - \right. \\ & \left. \frac{1}{\Gamma(\phi)} \left[\int_0^1 \left(\frac{1-u}{u}\right)^{1/\alpha} \left[\log \left(\frac{1}{u}\right) \right]^{\phi-1} du \right]^2 \right\}. \end{aligned}$$

Por meio dos momentos, também é possível determinar o coeficiente de assimetria e o coeficiente de curtose das novas distribuições de probabilidade (MAGALHÃES, 2011). Essas medidas podem ser utilizadas para caracterizar o comportamento das novas distribuições. Assim, supondo a existência do terceiro momento, coeficiente de assimetria, denotado por α_3 , indica o grau de assimetria da distribuição de probabilidade e é definido por

$$\alpha_3 = \frac{E[(X - \mu)^3]}{\sigma^3}.$$

Por outro lado, o coeficiente de curtose, denotado por α_4 , mede a intensidade dos picos da distribuição de probabilidade e é definido por

$$\alpha_4 = \frac{E[(X - \mu)^4]}{\sigma^4},$$

em que se supõe a existência do quarto momento.

A Figura 2.5 mostra as representações gráficas do coeficiente de assimetria (Figura 2.5a) para $\lambda = 0,5$ e do coeficiente de curtose (Figura 2.5b) para $\lambda = 1,0$ como função do parâmetro α para alguns valores do parâmetro ϕ . Por meio dessas figuras, observa-se que quando o parâmetro α aumenta a distribuição gama-Weibull é mais alta e concentrada do que a distribuição normal padrão, ou seja, tem “caudas pesadas”. Por outro lado, quando o valor do parâmetro ϕ diminuiu a distribuição gama-Weibull vai se tornando assimétrica à direita.

Figura 2.5 – Assimetria e curtose da distribuição gama-Weibull como função do parâmetro α , para alguns valores de ϕ . (a) Para $\lambda = 0,5$ e (b) Para $\lambda = 1,0$

Considerando, a distribuição gama-log-logística, a Figura 2.6 mostra os gráficos do coeficiente de assimetria (Figura 2.6a) e do coeficiente de curtose (Figura 2.6b) para $\lambda = 1,0$ como função do parâmetro α para alguns valores do parâmetro ϕ . Assim como a distribuição gama-Weibull, a distribuição gama-log-logística apresenta “caudas mais pesadas” quando os

valores do parâmetro α aumentam. Em relação à assimetria, observa-se que a distribuição é assimétrica à direita independente dos valores de ϕ e α .

Figura 2.6 – Assimetria e curtose da distribuição gama-log-logística como função do parâmetro α , para alguns valores de ϕ e para $\lambda = 1, 0$

Uma outra medida de interesse, são os momentos incompletos de uma distribuição de probabilidade. Em geral, os momentos incompletos são utilizadas no contexto de economia para determinar as medidas de desigualdade da distribuição de renda de uma população (BUTLER; McDONALD, 1989). Assim, o r -ésimo momento incompleto de uma variável aleatória contínua X , apresentada por Butler e McDonald (1989), é definido por

$$T_r(y) = E(X^r | X \leq y) = \int_0^y x^r f(x) dx, \quad (2.14)$$

em que r é um número inteiro e positivo e $f(x)$ é a função densidade de probabilidade da variável aleatória X .

Portanto, o r -ésimo momento incompleto das novas distribuições é obtido da seguinte forma:

i. Distribuição gama-Weibull

Para a distribuição gama-Weibull, substituindo a função densidade de probabilidade (2.4) na expressão (2.14), o r -ésimo momento incompleto da distribuição gama-Weibull é definido por

$$T_r(y) = \frac{\lambda^r}{\Gamma(\phi)} \gamma \left[\left(\frac{y}{\lambda} \right)^\alpha, r\alpha^{-1} + \phi \right],$$

em que $\gamma(x, \phi)$ denota a função gama incompleta.

ii. Distribuição gama-log-logística

Para a distribuição gama-log-logística, substituindo a função densidade de probabilidade (2.7) na expressão (2.14), o r -ésimo momento incompleto da distribuição gama-log-logística é definido por

$$T_r(y) = \frac{\lambda^r}{\Gamma(\phi)} \int_0^v \left(\frac{1-u}{u} \right)^{r/\alpha} \left[\log \left(\frac{1}{u} \right) \right]^{\phi-1} du,$$

em que $v = u = \left[1 + \left(\frac{y}{\lambda} \right)^\alpha \right]^{-1}$.

Assim, por meio dos momentos incompletos, é possível determinar as curvas de Lorenz e Bonferroni. Essas curvas são muito utilizadas em estudos de economia e demografia, pois permitem descrever a distribuição de renda em uma sociedade.

As curvas de Lorenz e Bonferroni apresentadas por Cordeiro et al. (2012) são escritas como função dos momentos incompletos e são expressas, respectivamente, por

$$L(\pi) = \frac{T_1(q)}{\mu'_1} \quad \text{e} \quad B(\pi) = \frac{T_1(q)}{(\pi \mu'_1)},$$

em que μ'_1 é o primeiro momento, $T_1(\cdot)$ é o primeiro momento incompleto e $q = Q(\pi)$ é a função quantílica (2.9) ou (2.10) a uma dada probabilidade π .

As Figuras 2.7-2.8 ilustram as curvas de Lorenz e Bonferroni para alguns valores de ϕ para as duas distribuições propostas.

Por meio dessas figuras é possível observar que à medida que os valores do parâmetro ϕ diminuem, as curvas de Lorenz e Bonferroni tendem a se aproximar da desigualdade máxima de renda, no contexto de estudos econômicos, para ambas as distribuições de probabilidade.

Figura 2.7 – Curvas de Bonferroni e Lorenz como função de π para $\lambda = 1, 0$; $\alpha = 0, 6$ e alguns valores de ϕ da distribuição gama-Weibull

Figura 2.8 – Curvas de Bonferroni e Lorenz como função de π para $\lambda = 1, 0$; $\alpha = 2, 0$ e alguns valores de ϕ da distribuição gama-logística

2.2.3 Função geradora de momentos

Em Johnson, Kemp e Kotz (2005), a função geradora de momentos de uma variável aleatória contínua X é definida por

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx, \quad (2.15)$$

em que $f(x)$ é a função densidade de probabilidade da variável aleatória X . Assim, utilizando essa expressão, a função geradora de momentos das novas distribuições, é obtida da seguinte maneira:

i. Distribuição gama-Weibull

Para a distribuição gama-Weibull, substituindo a expressão (2.4) na função geradora de momentos (2.15), obtém-se

$$M_X(t) = \frac{\alpha}{\lambda^{\alpha\phi}\Gamma(\phi)} \int_0^{\infty} e^{tx} x^{\alpha\phi-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{\alpha}\right] dx. \quad (2.16)$$

Mas, sabendo que e^{tx} pode ser escrito como

$$e^{tx} = \sum_{m=0}^{\infty} \frac{(tx)^m}{m!},$$

e após alguns cálculos matemáticos, é possível escrever a integral em (2.16) como uma função gama. Portanto, a função geradora de momentos que caracteriza a distribuição gama-Weibull é dada por

$$M_X(t) = \frac{1}{\Gamma(\phi)} \sum_{m=0}^{\infty} \frac{(\lambda t)^m}{m!} \Gamma(m\alpha^{-1} + \phi).$$

ii. Distribuição gama-log-logística

Para a distribuição gama-log-logística, substituindo a expressão (2.7) na função geradora de momentos (2.15), obtém-se

$$M_X(t) = \frac{\alpha}{\lambda^\alpha \Gamma(\phi)} \int_0^\infty e^{tx} x^{\alpha-1} \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-2} \left\{ \log \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right] \right\}^{\phi-1} dx.$$

Mas, sabendo que e^{tx} pode ser escrito como

$$e^{tx} = \sum_{m=0}^{\infty} \frac{(tx)^m}{m!}$$

e após alguns cálculos matemáticos, a função geradora de momentos que caracteriza a distribuição gama-log-logística é dada por

$$M_X(t) = \frac{1}{\Gamma(\phi)} \sum_{m=0}^{\infty} \frac{(\lambda t)^m}{m!} \int_0^1 \left(\frac{1-u}{u} \right)^{m/\alpha} \left[\log \left(\frac{1}{u} \right) \right]^{\phi-1} du,$$

em que $u = \left[1 + \left(\frac{x}{\lambda} \right)^\alpha \right]^{-1}$.

2.2.4 Estimação de parâmetros por máxima verossimilhança

O método de estimação por máxima verossimilhança é o método de estimação mais utilizado em análise de sobrevivência, pois permite que a informação dos tempos censurados seja incorporado no processo de estimação, ou seja, a função de verossimilhança leva em conta a contribuição dos indivíduos que foram falhas e os indivíduos que foram censurados (LAWLESS, 2003).

Dessa forma, seja $x_1, \dots, x_n$ uma amostra aleatória de uma distribuição de probabilidade e sejam $c_1, \dots, c_n$ os tempos de censura aleatória, então, os dados observados são definidos por $t_i = \min(x_i, c_i)$ e um indicador de censura $\delta_i = I(x_i \leq c_i)$. Portanto, assumindo que as censuras são aleatórias e não-informativas, a função de verossimilhança é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n [f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}], \quad (2.17)$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos, $f(t_i)$ é a função densidade de probabilidade e $S(t_i)$ é a função de sobrevivência.

Assim, o estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ do vetor de parâmetros $\boldsymbol{\theta}$ é calculado maximizando o logaritmo da função de verossimilhança dada na expressão (2.17). Esses estimadores são obtidos resolvendo o seguinte sistema de equações não-lineares

$$U(\boldsymbol{\theta}) = \frac{\partial \log[L(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = 0.$$

Dessa forma, as funções utilizadas para resolver o sistema de equações acima e encontrar as estimativas de máxima verossimilhança dos parâmetros da distribuição gama-Weibull e da distribuição gama-log-logística são:

i. Distribuição gama-Weibull

Seja $(t_1, \delta_1), \dots, (t_n, \delta_n)$ um conjunto de dados de tamanho n , em que t_i é a variável resposta e δ_i é o indicador de censura (0=tempo de censura, 1=tempo de falha), para $i = 1, \dots, n$. Então, o logaritmo da função de verossimilhança para dados censurados é obtido substituindo (2.4) e a função de sobrevivência (2.5) na expressão (2.17) e é escrito da seguinte forma

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r \log[\Gamma(\phi)] + r\alpha\phi \log(\lambda) + \alpha\phi \sum_{i=1}^n \delta_i \log(t_i) - \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\left(\frac{t_i}{\lambda} \right)^\alpha, \phi \right] \right\}, \quad (2.18)$$

em que $\boldsymbol{\theta} = (\phi, \lambda, \alpha)^\top$ é o vetor de parâmetros desconhecidos e r é o número que observações que não foram censuradas.

A função escore é obtida, calculando-se a derivada do logaritmo da função de verossimilhança (2.18) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Assim, tem-se que

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + r\alpha \log(\lambda) + \alpha \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n (1 - \delta_i) [\dot{A}_i]_\phi,$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} = \frac{r\alpha\phi}{\lambda} + \sum_{i=1}^n (1 - \delta_i) [\dot{A}_i]_\lambda \quad \text{e}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} = \frac{r}{\alpha} + r\phi \log(\lambda) + \phi \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n (1 - \delta_i) [\dot{A}_i]_\alpha,$$

em que $S(t_i) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\left(\frac{t_i}{\lambda} \right)^\alpha, \phi \right]$, $A_i = \log[S(t_i)]$, $[\dot{A}_i]_\phi = \frac{\partial A_i}{\partial \phi}$, $[\dot{A}_i]_\lambda = \frac{\partial A_i}{\partial \lambda}$ e $[\dot{A}_i]_\alpha = \frac{\partial A_i}{\partial \alpha}$.

ii. Distribuição gama-log-logística

Seja $(t_1, \delta_1), \dots, (t_n, \delta_n)$ um conjunto de dados de tamanho n , em que t_i é a variável resposta e δ_i é o indicador de censura (0=tempo de censura, 1=tempo de falha), para $i = 1, \dots, n$. Então, o logaritmo da função de verossimilhança para dados censurados é obtido substituindo (2.7) e a função de sobrevivência (2.8) na expressão (2.17) e é escrito da seguinte forma

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r \log[\Gamma(\phi)] + r\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i) - 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\} + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right], \phi \right] \right\}, \quad (2.19)$$

em que $\boldsymbol{\theta} = (\phi, \lambda, \alpha)^\top$ é o vetor de parâmetros desconhecidos e r é o número de observações que não foram censuradas.

A função escore é obtida, calculando-se a derivada do logaritmo da função de verossimilhança (2.19) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Assim, tem-se que

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\} + \sum_{i=1}^n (1 - \delta_i) [\dot{A}_i]_\phi,$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} = & \frac{r\alpha}{\lambda} + \frac{2\alpha}{\lambda^{\alpha+1}} \sum_{i=1}^n \delta_i t_i^\alpha \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right]^{-1} - \frac{\alpha(\phi-1)}{\lambda^{\alpha+1}} \sum_{i=1}^n \delta_i t_i^\alpha \log^{-1} \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \times \\ & \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right]^{-1} + \sum_{i=1}^n (1 - \delta_i) [\dot{A}_i]_\lambda \quad \text{e} \end{aligned}$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} = & \frac{r}{\alpha} + r \log(\lambda) + \phi \sum_{i=1}^n \delta_i \log(t_i) + \frac{\phi-1}{\lambda^\alpha} \sum_{i=1}^n t_i^\alpha \log \left(\frac{t_i}{\lambda} \right) \log^{-1} \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \times \\ & \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right]^{-1} + \sum_{i=1}^n (1 - \delta_i) [\dot{A}_i]_\alpha, \end{aligned}$$

em que $S(t_i) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right], \phi \right\}$, $A_i = \log[S(t_i)]$, $[\dot{A}_i]_\phi = \frac{\partial A_i}{\partial \phi}$, $[\dot{A}_i]_\lambda = \frac{\partial A_i}{\partial \lambda}$ e $[\dot{A}_i]_\alpha = \frac{\partial A_i}{\partial \alpha}$.

Entretanto, em algumas situações não é possível resolver analiticamente esse sistema de equações, por essa razão, algoritmos computacionais como de *Newton-Raphson* ou *Quase-Newton* são utilizados para encontrar os estimadores de máxima verossimilhança. Neste trabalho, foi utilizada a sub-rotina `MaxBFGS` da linguagem matricial de programação `OX` versão 6.20 (DOORNIK, 2007).

2.2.5 Estudo de simulação

Para avaliar as estimativas de máxima verossimilhança da distribuição gama-Weibull e da distribuição gama-log-logística, um estudo de simulação foi realizado. Os tempos de sobrevivência $t_1, \dots, t_n$ provenientes da distribuição gama-Weibull (2.4) e da distribuição gama-log-logística (2.7) foram gerados a partir do método da aceitação-rejeição, maiores detalhes sobre o método podem ser encontrados em Ross (2006) e Bonat et al. (2012).

Assim, para amostras de tamanho $n = 100$, $n = 300$ e $n = 500$, para a distribuição gama-Weibull, os parâmetros foram fixados em $\phi = 0,5$ e $1,5$; $\alpha = 0,6$, $\alpha = 0,6$ e $\lambda = 1,0$. Por outro lado, para a distribuição gama-log-logística, os parâmetros foram fixados em $\phi = 0,8$ e $1,5$; $\alpha = 0,6$, $\alpha = 0,6$ e $\lambda = 1,0$.

Os tempos de sobrevivência foram gerados a partir do seguinte algoritmo (Zeviani, 2012) adaptado neste trabalho para dados censurados.

- i. Gerar $y \sim \text{uniforme}(0, b_1)$.
- ii. Gerar $u \sim \text{uniforme}(0, b_2)$.
- iii. Se $u \leq f(y) \Rightarrow t^* = y$, em que $f(\cdot)$ é a função densidade de probabilidade (2.4) ou (2.7).
- iv. Gerar $c \sim \text{uniforme}(0, \tau)$, em que τ foi ajustado até obter as porcentagens de censura à direita 10% e 30%.
- v. Fazer $t = \min(t^*, c)$ e $\delta = I(t^* \leq c)$ como sendo o indicador de censura.
- vi. Caso contrário, retorne ao passo i.

Dessa forma, para cada combinação de n , ϕ , α , λ e porcentagens de censuras, 1000 amostras aleatórias foram geradas por meio de simulações de Monte Carlo e para cada uma das amostras foi obtido as estimativas de máxima verossimilhança dos parâmetros da distribuição gama-Weibull e da distribuição gama-log-logística. A partir dessas 1000 estimativas de cada parâmetro das duas distribuições, foram calculadas as médias e o erro quadrático médio.

Tabela 2.1 – Média e erro quadrático médio (em parênteses) das estimativas dos parâmetros da distribuição gama-Weibull

n	Parâmetro	$\phi = 0,5$			$\phi = 1,5$		
		0%	10%	30%	0%	10%	30%
100	ϕ	0,5568 (0,1229)	0,5799 (0,2050)	0,6058 (0,3340)	2,1267 (5,3015)	2,0615 (5,1436)	2,0082 (5,0089)
	α	2,3661 (5,7427)	2,3328 (1,7131)	2,6842 (6,3489)	2,1773 (0,8957)	2,2314 (1,8864)	2,4046 (1,9313)
	λ	2,9421 (0,5649)	2,9079 (0,6348)	2,8720 (0,8032)	2,8887 (1,6270)	2,9230 (1,5219)	3,0265 (1,7374)
300	ϕ	0,5106 (0,0189)	0,5101 (0,0219)	0,5148 (0,0314)	1,6789 (0,7524)	1,6520 (0,6159)	1,6772 (0,9003)
	α	2,0719 (0,1556)	2,0845 (0,1766)	2,1401 (0,3615)	2,0405 (0,1969)	2,0652 (0,2303)	2,0961 (0,3196)
	λ	2,9908 (0,1677)	2,9887 (0,1797)	2,9753 (0,2412)	2,9483 (0,6125)	2,9664 (0,6367)	2,9763 (0,7742)
500	ϕ	0,5032 (0,0099)	0,5071 (0,0114)	0,5127 (0,0167)	1,5768 (0,2286)	1,5646 (0,2617)	1,5917 (0,3862)
	α	2,0491 (0,0867)	2,0422 (0,0922)	2,0619 (0,1610)	2,0284 (0,1138)	2,0472 (0,1245)	2,0604 (0,1728)
	λ	3,0014 (0,0941)	2,9900 (0,1018)	2,9696 (0,1366)	2,9776 (0,3462)	2,9984 (0,3654)	2,9921 (0,4776)

Tabela 2.2 – Média e erro quadrático médio (em parênteses) das estimativas dos parâmetros da distribuição gama-log-logística

n	Parâmetro	$\phi = 0,8$			$\phi = 1,5$		
		0%	10%	30%	0%	10%	30%
100	ϕ	0,4360 (0,0167)	0,4591 (0,0189)	0,5563 (0,3473)	1,8837 (9,8490)	2,0722 (13,5900)	2,1919 (21,7350)
	α	2,4236 (0,2892)	2,3713 (0,2629)	2,1758 (0,2196)	2,4013 (0,5592)	2,4186 (0,6353)	2,4570 (0,7782)
	λ	3,0747 (0,3595)	2,9761 (0,3865)	2,7082 (0,6162)	3,7747 (2,7686)	3,6141 (2,5868)	3,6012 (2,9004)
300	ϕ	0,4073 (0,0115)	0,4241 (0,0093)	0,4825 (0,0092)	1,2750 (0,7020)	1,2860 (0,3768)	1,3098 (0,4893)
	α	2,4327 (0,2256)	2,3716 (0,1812)	2,1816 (0,1048)	2,2116 (0,0804)	2,2035 (0,0647)	2,2062 (0,0779)
	λ	3,1957 (0,1598)	3,1053 (0,1479)	2,8789 (0,1988)	3,7701 (1,2210)	3,6782 (1,0817)	3,6670 (1,1379)
500	ϕ	0,4052 (0,0108)	0,4216 (0,0083)	0,4712 (0,0056)	1,2302 (0,1080)	1,2403 (0,1054)	1,2583 (0,1043)
	α	2,4258 (0,2054)	2,3757 (0,1675)	2,1852 (0,0766)	2,1909 (0,0445)	2,1920 (0,0454)	2,1911 (0,0473)
	λ	3,1960 (0,1223)	3,1103 (0,0874)	2,9276 (0,1215)	3,7452 (0,9082)	3,7159 (0,8630)	3,6695 (0,8484)

Por meio das Tabelas 2.1-2.2, é possível observar que o erro quadrático médio dos parâmetros da distribuição gama-Weibull e da distribuição gama-log-logística aumentam quando a porcentagem de censura aumenta e por outro lado, o erro quadrático diminui quando

o tamanho da amostra aumenta. Observa-se, também, que o erro quadrático é maior para valores de ϕ menores que um.

2.3 Aplicação

A fim de exemplificar o ajuste da distribuição gama-Weibull e da distribuição gama-log-logística a dados censurados, considerou-se um estudo com gatos domésticos descrito em Petrie e Watson (2009). O conjunto de dados apresentado na Tabela 2.3, é referente a um experimento conduzido por Elliott et al. (2000) com o objetivo de avaliar a eficiência de duas dietas no tratamento de insuficiência renal crônica felina. No estudo, 50 gatos domésticos diagnosticados com insuficiência renal crônica foram considerados e foi observado para cada animal o tempo (em dias) a partir do diagnóstico até a morte devido à doença e o sinal + indica o tempo censurado.

Tabela 2.3 – Tempo (em dias) de vida de gatos domésticos com insuficiência renal crônica

367	511	535	252	166	173	368	797	576	189
191	264	166	252	236	158	1029+	532	419	674+
534	201	970	586	581	954	1036	334	202	247
973	677	633	345	685	320	478	259	267	142
1240+	1218+	707	1057+	980+	801	1227+	911+	766	411

Inicialmente, para verificar o comportamento dos tempos de vida dos gatos, foram construídos a curva TTT e a curva de Kaplan-Meier, como ilustra a a Figura 2.9.

Figura 2.9 – TTT-plot (a) e estimativas de Kaplan-Meier da função de sobrevivência (b) para os dados de gatos domésticos

Por meio da Figura 2.9, pode-se observar que a curva TTT indica que a função de taxa de falha dos tempos de vida dos gatos domésticos apresenta forma crescente, pois a curva é

côncava e a curva de sobrevivência estimada pelo método de Kaplan-Meier indicando que não há presença de um patamar acima de zero. Assim, para esse conjunto de dados, ajustou-se a distribuição gama-Weibull (2.4) e a distribuição gama-log-logística (2.7).

Nesse caso, para determinar as estimativas dos parâmetros, consideraram-se os logaritmos das funções de verossimilhança (2.18) e (2.19). Utilizando a sub-rotina `MaxBGFS` da linguagem matricial de programação `Oct` versão 6.20, as estimativas de máxima verossimilhança dos parâmetros desses modelos foram obtidas considerando como valores iniciais $\phi = 1,000$; $\lambda = 673,129$ e $\alpha = 1,519$ obtidos a partir do ajuste da distribuição Weibull no pacote estatístico `R` versão 2.15.1. As estimativas de máxima verossimilhança dos parâmetros dos modelos ajustados, bem como, os erros padrões (entre parênteses) são apresentados na Tabela 2.4.

Desse modo, após encontrar as estimativas dos parâmetros de cada modelo, as seguintes estatísticas foram utilizadas para selecionar o modelo mais adequado para analisar o conjunto de dados de gatos com problema de insuficiência renal:

i. CrITÉrio de informação de Akaike (AIC)

$$AIC = -2l(\theta) + 2p, \quad (2.20)$$

em que p é o número de parâmetros da distribuição e $l(\theta)$ é o logaritmo da função de verossimilhança.

ii. CrITÉrio de informação bayesiano (BIC)

$$BIC = -2l(\theta) + p \log(n), \quad (2.21)$$

em que n é o tamanho da amostra.

Assim, os menores valores dessas estatísticas, indicam o modelo mais adequado, dentre os estudados. Dessa forma, para as novas distribuições, bem como para os seus casos particulares, as estatísticas AIC e BIC são apresentadas na Tabela 2.4.

Tabela 2.4 – EMV dos parâmetros dos modelos ajustados para os dados de gatos domésticos, os erros padrões (entre parênteses) e as estatísticas AIC e BIC

Modelo	ϕ	λ	α	AIC	BIC
Gama-Weibull	22,4796 (3,0890)	0,0100 (0,0021)	0,2866 (0,0130)	619,9138	625,6499
Gama-log-logística	6,6022 (4,9141)	64,0773 (47,7523)	3,2159 (1,3464)	615,8317	621,5677
Weibull	—	673,1290 (69,1334)	1,519 (0,1904)	623,7458	627,5699
Log-logística	—	477,1090 (54,4274)	2,2044 (0,2765)	619,1527	622,9767

Então, por meio da Tabela 2.4, observa-se a distribuição gama-log-logística apresenta os menores valores de AIC e BIC, indicando que seria o modelo mais adequado, dentre os

modelos ajustados. Da mesma forma, a curva de sobrevivência estimada por cada modelo *versus* a curva de sobrevivência estimada pelo método de Kaplan-Meier, representadas na Figura 2.10, também indicam um melhor ajuste da distribuição gama-log-logística, pois nos tempos iniciais e finais, a curva apresenta melhor acompanhamento, apesar de ser mínima a diferença.

Figura 2.10 – Curvas de sobrevivência estimadas pelos modelos ajustados *versus* a curva de sobrevivência estimada pelo método de Kaplan-Meier para os dados de gatos domésticos

Diante do que foi exposto, a distribuição gama-log-logística foi considerada para analisar o conjunto de dados de tempo de vida de gatos diagnosticados com insuficiência renal crônica. Portanto, o modelo é dado por

$$S(x) = 1 - \frac{1}{\Gamma(6, 6022)} \gamma \left\{ \log \left[1 + \left(\frac{x}{64, 0773} \right)^{3,2159} \right]; 6, 6022 \right\}, \quad x > 0, \quad (2.22)$$

em que X é a variável aleatória que representa o tempo de vida dos animais.

Assim, a partir do modelo ajustado (2.22), é possível obter as seguintes informações sobre o conjunto de dados de gatos com insuficiência renal crônica:

- i. O tempo mediano, denotado por $\hat{t}_{0,5}$, pode ser estimado por meio da função quantílica (2.10). Dessa forma, tem-se

$$\hat{t}_{0,5} = 64, 0773 [\exp(\hat{Q}_{6,6022}(0, 5)) - 1]^{1/3,2159} = 450, 2576;$$

em que $\hat{Q}_{6,6022}(0, 5) = 6, 2720$ é a função quantílica da distribuição gama de um parâmetro.

Logo, o tempo mediano de vida dos gatos em tratamento de insuficiência renal é aproximadamente de 450 dias.

- ii. A chance de um gato estar livre da doença depois de 400 dias é estimada por meio da função de sobrevivência (2.22). Assim,

$$\hat{S}(400) = 1 - \frac{1}{\Gamma(6, 6022)} \gamma \left\{ \log \left[1 + \left(\frac{400}{64, 0773} \right)^{3,2159} \right] ; 6, 6022 \right\} = 0, 56.$$

Portanto, a chance de um gato estar livre da doença depois de 400 dias após o tratamento é de cerca de 56%, ou seja, os animais que permanecem sem a doença até 400 dias, tem grandes chances de se curarem ou sobreviverem.

3 MODELO DE REGRESSÃO LOCAÇÃO - ESCALA

A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) é responsável por avaliar de três em três anos, os cursos de pós-graduação do país com o objetivo de manter os cursos com nível de excelência e contribuir para a formação de pesquisadores (HORTA; MORAES, 2005). De acordo com a CAPES, uma das exigências é que os cursos de mestrados sejam feitos em dois anos e os cursos de doutorado em quatro anos (MOREIRA et al., 2010).

Por essa razão, os programas de pós-graduação tem estudado forma de incentivar os alunos antes ou dentro do prazo estipulado. Para verificar a melhor estratégia a ser tomada, um determinado programa de pós-graduação observou o tempo que seus alunos de doutorado levavam para titular, em função do sexo e da idade no momento da primeira matrícula.

Uma maneira de estudar o efeito dessas variáveis explicativas (sexo e idade dos alunos) sobre o tempo de titulação é por meio de um modelo de regressão locação-escala, também conhecido como, modelo de tempo de vida acelerado. Esses modelos assumem que a variável resposta pertence a uma família de distribuições, que se caracteriza por um parâmetro de locação e um parâmetro de escala. Maiores detalhes sobre essa classe de modelos de regressão podem ser encontrados em Cox e Oakes (1984), Kalbfleisch e Prentice (2002) e Lawless (2003).

Contudo, por se tratar de modelos paramétricos, supõe-se que os tempos de titulação sejam uma amostra proveniente de alguma distribuição de probabilidade. Dessa forma, no contexto de análise de sobrevivência, algumas distribuições tem sido utilizadas como alternativa para analisar dados censurados. Por exemplo, Leiva et al. (2007) realizaram um estudo de diagnóstico em um modelo de regressão log-Birnbaum-Saunders. Carrasco, Ortega e Paula (2008) propuseram o modelo de regressão considerando uma modificação da distribuição Weibull. Ortega, Cordeiro e Hashimoto (2011) propuseram o modelo de regressão beta-Weibull. Silva, Ortega e Paula (2011) propuseram um modelo de regressão log-Burr XII.

Assim, utilizando a mesma abordagem adotada nesses trabalhos, as duas novas distribuições de probabilidade propostas no capítulo anterior, serão escritas na forma de distribuições que pertencem aos modelos de locação-escala. Dessa modo, é possível estudar a influência das covariáveis sobre o tempo de titulação de alunos de doutorado.

3.1 Especificação do modelo log-gama-G

Para obter uma distribuição que pertence aos modelos de locação-escala, considere a transformação de variáveis aleatórias $Y = \log(T)$, em que T denota o tempo de falha ou sobrevivência e utilizando o método da transformação do jacobiano (MOOD; GRAYBILL; BOES, 1974). Assim, considerando as reparametrizações $\alpha = 1/\sigma$ e $\lambda = \exp(\mu)$, para as duas novas distribuições de probabilidade tem-se que:

i. Log-gama-Weibull

Se T é uma variável aleatória com distribuição gama-Weibull definida em (2.4), então

$$f(y) = \frac{1}{\sigma \Gamma(\phi)} \exp \left[\frac{\phi(y - \mu)}{\sigma} - \exp \left(\frac{y - \mu}{\sigma} \right) \right], \quad -\infty < y < \infty, \quad (3.1)$$

em que $\phi > 0$ é o parâmetro de forma, $\sigma > 0$ é o parâmetro de escala e $-\infty < \mu < \infty$ é o parâmetro de locação. Portanto, Y segue uma distribuição log-gama-Weibull.

A Figura 3.1 ilustra algumas formas da função densidade de probabilidade (3.1) para $\mu = 0$, $\sigma = 1$ e diferentes valores de ϕ . Por meio dessa figura, observa-se que à medida que o valor do parâmetro ϕ aumenta, a forma da distribuição se torna menos achatada, ou seja, leptocúrtica.

Figura 3.1 – Formas da função densidade de probabilidade da distribuição log-gama-Weibull

O r -ésimo momento da distribuição log-gama-Weibull é obtido substituindo a $f(y)$ definida em (3.1) na expressão (2.11). Assim,

$$\mu'_r = \frac{1}{\sigma \Gamma(\phi)} \int_{-\infty}^{\infty} y^r \exp \left[\frac{\phi(y - \mu)}{\sigma} - \exp \left(\frac{y - \mu}{\sigma} \right) \right] dy.$$

Após algumas mudanças de variáveis e utilizando uma expansão binomial, o r -ésimo momento da distribuição log-gama-Weibull é dado por

$$\mu'_r = \frac{1}{\Gamma(\phi)} \sum_{i=0}^r \binom{r}{i} \sigma^i \mu^{r-i} \int_0^{\infty} [\log(u)]^i u^{\phi-1} \exp(-u) du, \quad (3.2)$$

em que $u = \exp \left(\frac{y - \mu}{\sigma} \right)$.

A partir do r -ésimo momento (3.2) é possível determinar a média e a variância da distribuição log-gama-Weibull, que são expressos, respectivamente, por

$$E(X) = \mu'_1 = \frac{1}{\Gamma(\phi)} \sum_{i=0}^1 \binom{1}{i} \sigma^i \mu^{1-i} \int_0^\infty [\log(u)]^i u^{\phi-1} \exp(-u) du = \mu + \sigma\psi(\phi)$$

e

$$\text{var}(X) = E^2(X) = \sigma^2\psi(1, \phi) + \sigma^2\psi^2(\phi) = 2\mu\sigma\psi(\phi),$$

em que $\psi(\cdot)$ é a função digama, $\psi(n, \phi)$ é a função poligama e n sendo um número inteiro.

ii. Log-gama-logística

Se T é uma variável aleatória com distribuição gama-log-logística definida em (2.7), então

$$f(y) = \frac{1}{\sigma\Gamma(\phi)} \exp\left(\frac{y-\mu}{\sigma}\right) \left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2} \left\{\log\left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\phi-1} \quad (3.3)$$

em que $\phi > 0$, $\sigma > 0$ e $-\infty < \mu < \infty$. Portanto, Y segue uma distribuição log-gama-logística.

A Figura 3.2 ilustra algumas formas da função densidade de probabilidade (3.3) para $\mu = 0$, $\sigma = 1$ e diferentes valores de ϕ . Por meio dessa figura, observa-se que à medida que o valor do parâmetro ϕ diminui a forma da função densidade de probabilidade é mais achatada.

Figura 3.2 – Formas da função densidade de probabilidade da distribuição log-gama-logística

O r -ésimo momento da distribuição log-gama-logística é obtido substituindo a $f(y)$ definida em (3.3) na expressão (2.11). Assim,

$$\mu'_r = \frac{1}{\sigma\Gamma(\phi)} \int_{-\infty}^{\infty} y^r \exp\left(\frac{y-\mu}{\sigma}\right) \left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2} \left\{ \log\left[1 + \exp\left(\frac{y-\mu}{\sigma}\right)\right] \right\}^{\phi-1} dy.$$

Após algumas mudanças de variáveis e utilizando uma expansão binomial, o r -ésimo momento da distribuição log-gama-logística é dado por

$$\mu'_r = \frac{1}{\Gamma(\phi)} \sum_{i=0}^r \binom{r}{i} \sigma^i \mu^{r-i} \int_0^1 \left[\log\left(\frac{1-u}{u}\right) \right]^i \left[\log\left(\frac{1}{u}\right) \right]^{\phi-1} du, \quad (3.4)$$

em que $u = [1 + \exp(\frac{y-\mu}{\sigma})]^{-1}$.

A partir do r -ésimo momento (3.4) é possível determinar a média e a variância da distribuição log-gama-logística e que são expressos, respectivamente, por

$$E(X) = \mu'_1 = \mu + \frac{\sigma}{\Gamma(\phi)} \int_0^1 \left[\log\left(\frac{1-u}{u}\right) \right] [-\log(u)]^{\phi-1} du$$

e

$$\text{var}(X) = \mu'_2 - (\mu'_1)^2,$$

em que μ'_1 e μ'_2 são obtidos por meio da expressão (3.4).

Após ter definido a distribuição da variável resposta $Y = \log(T)$, na qual pertence à família de distribuições que se caracteriza pelo fato de ter um parâmetro de escala $\sigma(\sigma > 0)$ e parâmetro de locação $\mu(-\infty < \mu < \infty)$ que depende do vetor de covariáveis $\mathbf{x}$, denotado por $\mu(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, em que $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})$ é o vetor de covariáveis associado a i -ésima observação.

Então, o modelo de regressão que relaciona a variável resposta Y a um vetor de covariáveis $\mathbf{x}$ é um modelo escrito da seguinte forma

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma z_i, \quad i = 1, \dots, n, \quad (3.5)$$

em que y_i é o logaritmo dos tempos de sobrevivência para a i -ésima observação e z_i é uma variável aleatória que assume uma das seguintes distribuições:

i. Log-gama-Weibull

$$f(z) = \frac{1}{\Gamma(\phi)} \exp[\phi z - \exp(z)], \quad -\infty < z < \infty, \quad (3.6)$$

em que $\phi > 0$ é o parâmetro de forma desconhecido.

Portanto, a função densidade de probabilidade e a função de sobrevivência de $Y|\mathbf{x}$ assumem as formas

$$f(y|\mathbf{x}) = \frac{1}{\sigma\Gamma(\phi)} \exp\left[\frac{\phi(y - \mathbf{x}^\top \boldsymbol{\beta})}{\sigma} - \exp\left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma}\right)\right] \quad -\infty < y < \infty \quad (3.7)$$

e

$$S(y|\mathbf{x}) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right), \phi \right], \quad (3.8)$$

respectivamente.

ii. Log-gama-logística

$$f(z) = \frac{1}{\Gamma(\phi)} \exp(z) [1 + \exp(z)]^{-2} \{\log[1 + \exp(z)]\}^{\phi-1}, \quad -\infty < z < \infty, \quad (3.9)$$

em que $\phi > 0$ é o parâmetro de forma desconhecido.

Portanto, a função densidade de probabilidade e a função de sobrevivência de $Y|\mathbf{x}$ assumem as formas

$$f(y|\mathbf{x}) = \frac{1}{\sigma \Gamma(\phi)} \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \left[1 + \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \right]^{-2} \times \\ \left\{ \log \left[1 + \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \right] \right\}^{\phi-1}, \quad -\infty < y < \infty \quad (3.10)$$

e

$$S(y|\mathbf{x}) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \right], \phi \right\}, \quad (3.11)$$

respectivamente.

Observa-se que esse modelo é log-linear para T . Logo, é um modelo de regressão linear para Y . Outra característica desse modelo é que o vetor de covariáveis $\mathbf{x}$ tem efeito multiplicativo em T .

3.2 Estimação de parâmetros por máxima verossimilhança

Para estimar os parâmetros dos modelos de regressão foi utilizado o método da máxima verossimilhança, descrito no Capítulo 2, pois permite levar em conta a informação das observações censuradas. Assim, para um conjunto de dados que consiste de n observações, $(y_1, \delta_1, \mathbf{x}_1), \dots, (y_n, \delta_n, \mathbf{x}_n)$, em que $y_i = \log(t_i)$, δ_i é o indicador de censura e $\mathbf{x}_i$ é o vetor de covariáveis associados a i -ésima observação, para $i = 1, \dots, n$, o logaritmo da função de verossimilhança para cada um dos modelos é:

i. Log-gama-Weibull

$$l(\boldsymbol{\theta}) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \frac{\phi}{\sigma} \sum_{i=1}^n \delta_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - \sum_{i=1}^n \delta_i \exp \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) + \\ \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right), \phi \right] \right\}, \quad (3.12)$$

em que $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, r é o número que observações que não foram censuradas e $\gamma(x, \phi)$ é a função gama incompleta.

A função escore é obtida, calculando-se a derivada do logaritmo da função de verossimilhança (3.12) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Dessa forma

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + \frac{1}{\sigma} \sum_{i=1}^n \delta_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - \sum_{i=1}^n (1 - \delta_i) \frac{J[\exp(z_i), \phi, 1] - \psi(\phi)\gamma[\exp(z_i), \phi]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]},$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = -\frac{r}{\sigma} - \frac{\phi}{\sigma} \sum_{i=1}^n \delta_i z_i + \frac{1}{\sigma} \sum_{i=1}^n \delta_i z_i \exp(z_i) + \frac{1}{\sigma} \sum_{i=1}^n (1 - \delta_i) \frac{z_i \exp[\phi z_i - \exp(z_i)]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]} \quad \text{e}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = -\frac{\phi}{\sigma} \sum_{i=1}^n \delta_i x_{ij} + \frac{1}{\sigma} \sum_{i=1}^n \delta_i x_{ij} \exp(z_i) + \frac{1}{\sigma} \sum_{i=1}^n (1 - \delta_i) \frac{x_{ij} \exp[\phi z_i - \exp(z_i)]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]},$$

em que $z_i = \frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}$, $\psi(\cdot)$ é a função digama, $J(x, \phi, 1) = \int_0^x u^{\phi-1} e^{-u} \log(u) du$ e $j = 0, 1, \dots, p$.

ii. Log-gama-logística

$$\begin{aligned} l(\boldsymbol{\theta}) = & -r \log(\sigma) - r \log[\Gamma(\phi)] + \frac{1}{\sigma} \sum_{i=1}^n \delta_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - 2 \sum_{i=1}^n \delta_i \log \left[1 + \right. \\ & \left. \exp \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \exp \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right] \right\} + \\ & \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\log \left[1 + \exp \left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right], \phi \right] \right\} \end{aligned} \quad (3.13)$$

em que $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, r é o número que observações que não foram censuradas e $\gamma(x, \phi)$ é a função gama incompleta. A função escore é obtida, calculando-se a derivada do logaritmo da função de verossimilhança (3.13) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Dessa forma

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + \frac{1}{\sigma} \sum_{i=1}^n \delta_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - \sum_{i=1}^n (1 - \delta_i) \frac{J[\exp(z_i), \phi, 1] - \psi(\phi)\gamma[\exp(z_i), \phi]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]},$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = -\frac{r}{\sigma} - \frac{\phi}{\sigma} \sum_{i=1}^n \delta_i z_i + \frac{1}{\sigma} \sum_{i=1}^n \delta_i z_i \exp(z_i) + \frac{1}{\sigma} \sum_{i=1}^n (1 - \delta_i) \frac{z_i \exp[\phi z_i - \exp(z_i)]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]} \quad \text{e}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = -\frac{\phi}{\sigma} \sum_{i=1}^n \delta_i x_{ij} + \frac{1}{\sigma} \sum_{i=1}^n \delta_i x_{ij} \exp(z_i) + \frac{1}{\sigma} \sum_{i=1}^n (1 - \delta_i) \frac{x_{ij} \exp[\phi z_i - \exp(z_i)]}{\Gamma(\phi) - \gamma[\exp(z_i), \phi]},$$

em que $z_i = \frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}$, $\psi(\cdot)$ é a função digama, $J(x, \phi, 1) = \int_0^x u^{\phi-1} e^{-u} \log(u) du$ e $j = 0, 1, \dots, p$.

Portanto, para maximizar os logaritmos das funções de verossimilhança (3.12) e (3.13) e assim, encontrar as estimativas dos parâmetros, foi utilizada a sub-rotina `MaxBFGS` da linguagem matricial de programação `Oct` versão 6.2.0 (DOORNIK, 2007). Pois, nesse caso, não é possível resolver analiticamente o sistema de equações, representados pela função `score`.

No caso do tamanho da amostra ser grande e sob certas condições gerais de regularidade da função de verossimilhança, os intervalos de confiança e testes de hipóteses podem ser obtidos utilizando o fato de que a distribuição assintótica dos estimadores de máxima verossimilhança é a distribuição normal com média θ e matriz de covariância igual ao inverso da matriz de informação. Portanto, a matriz de covariância assintótica é igual a $\mathbf{I}^{-1}(\theta)$ com $\mathbf{I}(\theta) = E[\ddot{\mathbf{L}}(\theta)]$, em que $\ddot{\mathbf{L}}(\theta) = - \left[\frac{\partial^2 \log[L(\theta)]}{\partial \theta \partial \theta^\top} \right]$.

Como não é possível determinar a matriz de informação de Fisher $\mathbf{I}(\theta)$ devido a presença de observações censuradas, uma forma alternativa é utilizar a matriz de informação observada $\ddot{\mathbf{L}}(\theta)$, avaliada no estimador de máxima verossimilhança $\theta = \hat{\theta}$. A aproximação normal assintótica de $\hat{\theta}$ é expressa como

$$\hat{\theta}^\top \sim N_{(p+2)}[\theta^\top; \ddot{\mathbf{L}}(\theta)^{-1}],$$

em que $\ddot{\mathbf{L}}(\theta)$ é a matriz de informação observada $(p+2)(p+2)$.

Assim, o intervalo de confiança assintótico para β_j , com $j = 1, \dots, p$, considerando um nível de confiança de $(1 - \alpha)100\%$, é expresso por

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\hat{\text{Var}}(\hat{\beta}_j)},$$

em que $z_{\alpha/2}$ é o valor que excede com probabilidade $\alpha/2$ a distribuição normal padrão.

3.3 Critério para seleção de modelos

Sabendo que os modelos de regressão log-gama-Weibull e log-gama-logística não são encaixados e que existe a presença de observações censuradas, os critérios para escolher o modelo mais adequado entre os dois modelos propostos para analisar um conjunto de dados, são apresentados a seguir:

i. Critério de informação de Akaike (AIC_{SUR})

Na presença de observações censuradas, a estatística AIC em 2.20 pode não ser adequada. Por essa razão, Liang e Zou (2008) propuseram uma estatística AIC modificada para dados censurados e essa estatística é expressa por

$$AIC_{SUR} = AIC + \frac{2(p+2)(p+3)}{n-p-3}, \quad (3.14)$$

em que p é o número de parâmetros da distribuição, n é o tamanho da amostra e AIC é a estatística de AIC do modelo de regressão.

ii. Teste de hipótese para modelos não encaixados

Por outro lado, no caso de modelos não encaixados, o teste de hipótese apresentado por Silva, Ortega e Cordeiro (2009) pode ser utilizado para identificar o modelo mais adequado. Assim, o teste de hipótese supõe que F_θ é um modelo com função densidade de probabilidade $f(y; \theta)$ e G_γ é outro modelo com função densidade de probabilidade $g(y; \gamma)$. Então, as hipóteses de interesses para testar modelos não encaixados são

$$\begin{cases} H_0 : F_\theta \text{ e } G_\gamma \text{ são modelos equivalentes } \textit{versus} \\ H_a : \text{O modelo } F_\theta \text{ é mais adequado do que o modelo } G_\gamma \end{cases}$$

Logo, a estatística do teste é definida por

$$T = \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \log \left(\frac{f(y_i | \mathbf{x}_i, \hat{\theta})}{g(y_i | \mathbf{x}_i, \hat{\gamma})} \right) \right] \div \left\{ \frac{1}{n} \sum_{i=1}^n \left[\log \left(\frac{f(y_i | \mathbf{x}_i, \hat{\theta})}{g(y_i | \mathbf{x}_i, \hat{\gamma})} \right) \right]^2 - \left[\frac{1}{n} \sum_{i=1}^n \log \left(\frac{f(y_i | \mathbf{x}_i, \hat{\theta})}{g(y_i | \mathbf{x}_i, \hat{\gamma})} \right) \right]^2 \right\}, \quad (3.15)$$

em que n é o tamanho da amostra, θ é o vetor de parâmetros do modelo F_θ e γ é o vetor de parâmetros do modelo G_γ . A estatística (3.15) é uma distância entre os dois modelos, medida em termos do critério de informação de Kullback-Liebler e converge em distribuição para uma distribuição normal(0,1) sobre a hipótese nula (SILVA; ORTEGA; CORDEIRO, 2009). Logo, rejeita-se a hipótese nula (H_0), ao nível de significância α se $|T| > z_{\alpha/2}$. Nesse caso, conclui-se que o modelo F_θ é mais adequado do que o modelo G_γ .

3.4 Análise de sensibilidade

Uma etapa importante após a estimação dos parâmetros é o diagnóstico de anomalias do modelo ajustado, por exemplo, detectar observações influentes que podem causar importantes distorções nos resultados da análise. Essa etapa é conhecida como análise de sensibilidade. Denota-se por i -ésimo caso, a combinação das observações (y_i, δ_i) com os respectivos valores das covariáveis. Cook (1977) introduziu a medida de influência, denominada por distância de Cook, para avaliar a influência de um caso nas estimativas dos parâmetros do modelo de regressão linear. Essa distância é baseada na eliminação de casos, que é um enfoque comum para estudar o efeito de remover um caso da análise, assim é possível determinar quais indivíduos podem ser influentes. Uma desvantagem da eliminação de casos é que ela não permite dizer em que aspecto específico do modelo o caso tem influência. Medidas de diagnóstico baseadas na deleção de casos, tais como distância de Cook ou o afastamento de verossimilhança, têm sido aplicadas em vários modelos estatísticos, por exemplo, Lu e Song (2006), Poon e Poon (2007) e Jung (2008).

Uma metodologia alternativa, para contornar esse problema, foi desenvolvido por Cook (1986) que propõe avaliar se um modelo é relativamente estável sob pequenas perturbações, ao invés de usar eliminação de casos. Essa metodologia é denominada de influência local, na qual consiste em avaliar como os resultados da análise são modificados sob pequenas perturbações no modelo.

Diversos autores têm focado o estudo de influência local em modelos de análise de sobrevivência para identificar observações influentes na presença de censuras. Por exemplo, Pettitt e Bin Daud (1989) investigaram a influência local em modelo de regressão com riscos proporcionais, Escobar e Meeker (1992) adaptaram métodos de influência local para modelos de regressão com dados censurados. Além desses, existem vários trabalhos que consideraram influência local em modelos de regressão com dados censurados para uma específica distribuição, tais como, Ortega, Bolfarine e Ortega (2003) que trabalharam com a distribuição log-gama generalizada, Ortega, Paula e Bolfarine (2008) que consideraram a distribuição log-Weibull exponenciada e Cancho, Ortega e Paula (2010) que propuseram uma análise de sensibilidade em um modelo log-Birnbaum Saunders sob o enfoque bayesiano.

3.4.1 Influência global

Sabendo que a amostra observada é formada por $(Y_l, \delta_l, \mathbf{x}_l)$ para cada indivíduo $l = 1, \dots, n$, em que Y_l é o logaritmo da variável resposta, δ_l é o indicador de censura e $\mathbf{x}_l$ é o vetor de covariáveis, então, a eliminação de um indivíduo (ou caso) em um modelo de regressão é especificada por

$$Y_l = \mathbf{x}_l^\top \boldsymbol{\beta} + \sigma Z_l, \quad l = 1, \dots, n, \quad l \neq i, \quad (3.16)$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos, σ é o parâmetro escala e Z_l é uma variável aleatória.

Agora, considere que uma quantidade com subscrito i , significa a quantidade original com o i -ésimo caso eliminado. Assim, para o modelo (3.16), o logaritmo da função de verossimilhança de $\boldsymbol{\theta}$ excluindo o i -ésimo caso é denotado por $l_{(i)}(\boldsymbol{\theta})$, em que $\hat{\boldsymbol{\theta}}_{(i)}$ é o estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ calculada maximizando-se $l_{(i)}(\boldsymbol{\theta})$. Para avaliar a influência do i -ésimo caso da estimativa de máxima verossimilhança $\hat{\boldsymbol{\theta}}$, a idéia é utilizar a diferença entre $\hat{\boldsymbol{\theta}}_{(i)}$ e $\hat{\boldsymbol{\theta}}$. Então, se $\hat{\boldsymbol{\theta}}_{(i)}$ está distante de $\hat{\boldsymbol{\theta}}$ o i -ésimo caso pode ser considerado como uma observação influente.

Uma medida de influência global considerada por Xie e Wei (2007) é uma generalização da distância de Cook, que é definida como uma norma padronizada de $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$, e expressa por

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top [\ddot{\mathbf{L}}(\boldsymbol{\theta})] (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \quad (3.17)$$

em que $\ddot{\mathbf{L}}(\boldsymbol{\theta})$ é a matriz de informação observada.

Outra medida para avaliar a influência de um caso é apresentada por Cook e Weisberg (1982). Denominada de afastamento da verossimilhança, essa medida considera a diferença entre $\hat{\theta}_{(i)}$ e $\hat{\theta}$ e é expressa por

$$LD_i(\theta) = 2 \left[l(\hat{\theta}) - l(\hat{\theta}_{(i)}) \right], \quad (3.18)$$

em que $l(\hat{\theta})$ é o valor do logaritmo da função de verossimilhança da amostra completa e $l(\hat{\theta}_{(i)})$ é o valor do logaritmo da função de verossimilhança da amostra excluindo a i -ésima observação.

3.4.2 Influência local

As medidas de influência local foram propostas por Cook (1986) e são medidas obtidas a partir de alguns esquemas de perturbação no modelo. Assim, para obter as medidas de influência local, supõe-se que seja ω o vetor de perturbação de maneira que $\omega \in \Omega \subseteq \mathbf{R}^q$, em que ω é um conjunto aberto e $\hat{\theta}_\omega$ representa o estimador de máxima verossimilhança sob o modelo perturbado. O enfoque de influência local compara $\hat{\theta}_\omega$ com $\hat{\theta}$, em que estimativas similares indicam que as estimativas dos parâmetros são robustas com relação à perturbação no modelo. Por outro lado, diferenças relevantes sugerem que as estimativas produzidas são sensíveis a tais perturbações.

Cook (1986) sugere que as medidas de influência são obtidas depois que o modelo ou dados sejam perturbados por meio do vetor ω utilizando a função LD que representa o afastamento da verossimilhança definido por $LD(\omega) = 2[l(\hat{\theta}) - l(\hat{\theta}_\omega)]$, sobre a suposição que existe um $\omega_0 \in \Omega$ tal que $l(\theta \mid \omega_0) = l(\theta \mid \omega)$, em que ω_0 é o vetor de não perturbação. Com essa finalidade, considere em $\mathbf{R}^{q+1}$ a seguinte superfície

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

e o comportamento dessa função é analisada em uma vizinhança do ponto mínimo global ω_0 utilizando o conceito de curvatura normal de geometria diferencial.

Sob os conceitos de geometria diferencial, a idéia consiste em estudar a maior variação de LD em torno de ω_0 , o que equivale a selecionar uma direção unitária $\mathbf{d}$, $\|\mathbf{d}\| = 1$ e assim, calcular a curvatura normal para θ que é expressa como

$$C_d(\theta) = 2|\mathbf{d}^\top \Delta^\top \ddot{\mathbf{L}}(\theta)^{-1} \Delta \mathbf{d}|, \quad (3.19)$$

em que Δ é a matriz $p \times n$ que depende do esquema de perturbação e cujos elementos são iguais a $\Delta_{ji} = \partial^2 l(\theta \mid \omega) / \partial \theta_j \partial \omega_i$, $i = 1, \dots, n$ e $j = 1, \dots, p$ avaliados em $\hat{\theta}$ e ω_0 .

Utilizando a expressão (3.19), pode-se caracterizar a superfície α e, consequentemente, analisar a função $LD(\omega)$, assim é possível calcular a curvatura máxima $C_{\mathbf{d}_{max}}$ na direção correspondente, denotada por $\mathbf{d}_{max}$. A quantidade $C_{\mathbf{d}_{max}}$ corresponde ao maior autovalor da matriz $\mathbf{B} = \Delta^\top \ddot{\mathbf{L}}(\theta)^{-1} \Delta$ e $\mathbf{d}_{max}$ é o maior autovetor normalizado. Os elementos de $\mathbf{d}_{max}$ que são comparativamente discrepantes indicam que os respectivos elementos de ω podem

produzir mudanças substanciais nos estimadores de máxima verossimilhança e os respectivos pontos da amostra são possíveis pontos influentes. O gráfico dos elementos de $\mathbf{d}_{max}$ *versus* as observações são utilizados para indicar os elementos influentes.

Por outro lado, Lesaffre e Verbeke (1998) sugeriram calcular a estatística $C_{\mathbf{d}_i}(\boldsymbol{\theta})$, em que $\mathbf{d}_i$ é um vetor $n \times 1$ de zeros com valor um na i -ésima posição. Essa medida é denominada de influência local total e nesse caso, a curvatura na direção $\mathbf{d}_i$ assume a forma $C_i = 2|\boldsymbol{\Delta}_i^\top \ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1} \boldsymbol{\Delta}_i|$, em que $\boldsymbol{\Delta}_i^\top$ é a i -ésima linha de $\boldsymbol{\Delta}$. O gráfico dos C_i *versus* as observações também é utilizado para detectar as possíveis observações influentes.

Sabendo que as medidas de influência local são calculadas após um esquema de perturbação no modelo ou nos dados, então, diferentes esquemas de perturbação podem ser utilizados de acordo com o objetivo da análise. Dessa forma, considerando o modelo definido em (3.5) e o logaritmo da função de verossimilhança expresso em (3.12) e em (3.13), os seguintes esquemas de perturbação foram utilizados:

a. ***Perturbação de casos***

Considerando $0 \leq \omega_i \leq 1$ e $\boldsymbol{\omega}_0 = (1, \dots, 1)^\top$ como sendo o vetor de não perturbação. O logaritmo da função de verossimilhança perturbada, para cada modelo, incorpora diferentes pesos para cada elemento do conjunto de dados e é definido por:

i. *Log-gama-Weibull*

$$l(\boldsymbol{\theta}) = -\{r \log(\sigma) - r \log[\Gamma(\phi)]\} \sum_{i=1}^n \omega_i + \frac{\phi}{\sigma} \sum_{i=1}^n \omega_i \delta_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - \sum_{i=1}^n \left[\omega_i \delta_i \times \exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \right] + \sum_{i=1}^n \omega_i (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right), \phi \right] \right\}.$$

ii. *Log-gama-logística*

$$l(\boldsymbol{\theta}) = -\{r \log(\sigma) + r \log[\Gamma(\phi)]\} \sum_{i=1}^n \omega_i + \frac{1}{\sigma} \sum_{i=1}^n \omega_i \delta_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - 2 \sum_{i=1}^n \omega_i \delta_i \times \log \left[1 + \exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \right] + (\phi - 1) \sum_{i=1}^n \omega_i \delta_i \log \left\{ \log \left[1 + \exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \right] \right\} + \sum_{i=1}^n \omega_i (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\log \left[1 + \exp\left(\frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \right], \phi \right] \right\}.$$

b. ***Perturbação da variável resposta***

Para verificar a sensibilidade do modelo (3.5), considera-se que a variável resposta $y_i = \log(t_i)$ ($i = 1, \dots, n$) é submetida a um esquema de perturbação aditiva por meio

da adição de um vetor, denotado por ω , de pequenas perturbações, tal que $y_i^* = y_i + \omega_i \text{sd}(y_i)$, em que $\text{sd}(y_i)$ é um fator escala que pode ser o desvio padrão do logaritmo do tempos de falha e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Nesse caso, $\omega_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação e o logaritmo da função de verossimilhança perturbada, para os dois modelos, é expresso por:

i. Log-gama-Weibull

$$l(\theta) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \frac{\phi}{\sigma} \sum_{i=1}^n \delta_i (y_i^* - \mathbf{x}_i^\top \beta) - \sum_{i=1}^n \delta_i \exp\left(\frac{y_i^* - \mathbf{x}_i^\top \beta}{\sigma}\right) + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp\left(\frac{y_i^* - \mathbf{x}_i^\top \beta}{\sigma}\right), \phi \right] \right\}.$$

ii. Log-gama-logística

$$l(\theta) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \frac{1}{\sigma} \sum_{i=1}^n \delta_i (y_i^* - \mathbf{x}_i^\top \beta) - 2 \sum_{i=1}^n \delta_i \log \left[1 + \exp\left(\frac{y_i^* - \mathbf{x}_i^\top \beta}{\sigma}\right) \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \exp\left(\frac{y_i^* - \mathbf{x}_i^\top \beta}{\sigma}\right) \right] \right\} + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\log \left[1 + \exp\left(\frac{y_i^* - \mathbf{x}_i^\top \beta}{\sigma}\right) \right], \phi \right] \right\}.$$

c. **Perturbação de uma covariável**

Uma outra forma de avaliar a sensibilidade do modelo (3.5) é considerar pequenas perturbações em uma particular covariável contínua, denotada por X_j . Nesse caso, a covariável também é submetida a um esquema de perturbação aditiva, de tal forma que $x_{ij}^* = x_{ij} + \omega_i \text{sd}(x_{ij})$, em que $\text{sd}(x_{ij})$ é um fator de escala que pode ser o desvio padrão da covariável perturbada e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Dessa forma, considerando que $\mathbf{x}_i^{*\top} \beta = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j [x_{ij} + \omega_i \text{sd}(x_{ij})] + \dots + \beta_p x_{ip}$ e $\omega_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação, para cada um modelos, o logaritmo da função de verossimilhança perturbada é dada por:

i. Log-gama-Weibull

$$l(\theta) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \frac{\phi}{\sigma} \sum_{i=1}^n \delta_i (y_i - \mathbf{x}_i^{*\top} \beta) - \sum_{i=1}^n \delta_i \exp\left(\frac{y_i - \mathbf{x}_i^{*\top} \beta}{\sigma}\right) + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp\left(\frac{y_i - \mathbf{x}_i^{*\top} \beta}{\sigma}\right), \phi \right] \right\}.$$

ii. Log-gama-logística

$$l(\boldsymbol{\theta}) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \frac{1}{\sigma} \sum_{i=1}^n \delta_i (y_i - \mathbf{x}_i^{*\top} \boldsymbol{\beta}) - 2 \sum_{i=1}^n \delta_i \log \left[1 + \exp \left(\frac{y_i - \mathbf{x}_i^{*\top} \boldsymbol{\beta}}{\sigma} \right) \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \exp \left(\frac{y_i - \mathbf{x}_i^{*\top} \boldsymbol{\beta}}{\sigma} \right) \right] \right\} + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\log \left[1 + \exp \left(\frac{y_i - \mathbf{x}_i^{*\top} \boldsymbol{\beta}}{\sigma} \right) \right], \phi \right] \right\}.$$

Para esses três esquemas de perturbação calculou-se, numericamente, a matriz de curvaturas máximas definida por

$$\Delta = (\Delta_{vi})_{(p+2) \times n} = \left[\frac{\partial^2 l(\boldsymbol{\theta} | \boldsymbol{\omega})}{\partial \theta_v \partial \omega_i} \right]_{(p+2) \times n},$$

em que $v = 1, \dots, p+2$, $i = 1, \dots, n$ e $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top$ é o vetor de perturbação, isto é, os pesos que perturbam o modelo de regressão log-gama-Weibull ou o modelo de regressão log-gama-logística ou as observações.

3.5 Análise de resíduos

A fim de estudar as suposições dos erros bem como a presença de observações discrepantes, na literatura são propostas várias análises de resíduos, por exemplo, Collett (2003), Paula (2004), Weisberg (2005) e Colosimo e Giolo (2006).

De maneira geral, o resíduo referente a i -ésima observação é definido por meio de uma função, $r_i = r(y_i, \hat{\boldsymbol{\theta}})$, no qual depende das variáveis aleatórias e do vetor de parâmetros estimado. Dessa forma, os resíduos são utilizados para identificar a discrepância entre o modelo ajustado e o conjunto de dados.

No contexto de análise de sobrevivência, os resíduos martingale e o resíduo componente do desvio têm sido os mais utilizados, pois levam em conta a informação dos tempos censurados (SILVA et al., 2008)

Assim, os gráficos dos resíduos componente do desvio e martingale *versus* os tempos observados fornecem uma forma de verificar a adequação do modelo ajustado e auxiliam a encontrar observações atípicas. A seguir, esses resíduos são definidos para os modelos de regressão: log-gama-Weibull e log-gama-logística.

3.5.1 Resíduo martingale

O resíduo martingale apresentado por Fleming e Harrington (1991), foi proposto inicialmente no contexto de processos de contagens e é escrito da seguinte forma

$$r_{M_i} = \delta_i + \log[\hat{S}(y_i; \hat{\boldsymbol{\theta}})], \quad (3.20)$$

em que δ_i é o indicador de censura, $\hat{\theta}$ é o vetor de parâmetros estimados e $S(y_i; \hat{\theta})$ é a função de sobrevivência estimada. Assim, os resíduos martingale para os dois modelos de regressão obtidas da seguinte forma:

i. Log-gama-Weibull

Substituindo a função de sobrevivência (3.8) na expressão (3.20), os resíduos martingale do modelo de regressão log-gama-Weibull são obtidos por meio da expressão

$$r_{M_i} = \begin{cases} 1 + \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right\} & \text{se } \delta_i = 1 \\ \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right\} & \text{se } \delta_i = 0, \end{cases} \quad (3.21)$$

em que $\hat{z}_i = \frac{y_i - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$.

ii. Log-gama-logística

Analogamente, substituindo a função de sobrevivência (3.11) na expressão (3.20), os resíduos martingale do modelo de regressão log-gama-logística são dados por

$$r_{M_i} = \begin{cases} 1 + \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right\} & \text{se } \delta_i = 1 \\ \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right\} & \text{se } \delta_i = 0, \end{cases} \quad (3.22)$$

em que $\hat{z}_i = \frac{y_i - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$.

Segundo Collett (2003), esse resíduo não é simetricamente distribuído ao redor de zero, ou seja, tem o valor máximo em +1 e o valor mínimo $-\infty$.

3.5.2 Resíduo componente do desvio

Sabe-se que os resíduos martingale são apresentam forma assimétrica e por essa razão, os gráficos baseados nesse resíduo apresentem uma dificuldade na interpretação.

Diante disso, Therneau, Grambsch e Fleming (1990) introduziram um resíduo martingale modificado, denominado de resíduo componente do desvio, com a finalidade de tornar o resíduo martingale simetricamente distribuídos em torno do zero. Dessa forma, o resíduo é dado por

$$r_{D_i} = \text{sinal}(r_{M_i}) [-2 \{r_{M_i} + \delta_i \log(\delta_i - r_{M_i})\}], \quad (3.23)$$

em que δ_i é o indicador de censura, r_{M_i} é o resíduo martingale para o i -ésimo indivíduo e a função $\text{sinal}()$ é um função que conduz os valores +1 se o argumento é positivo e -1 se o argumento é negativo. Logo, o resíduo componente do desvio, para cada um dos modelos de regressão, é determinado da seguinte forma:

i. Log-gama-Weibull

Assim, substituindo o resíduo martingale (3.21) na expressão (3.23), o resíduo componente do desvio do modelo de regressão log-gama-Weibull é definido por

$$r_{D_i} = \begin{cases} \text{sin}(\hat{r}_{M_i}) \left\{ -2 \left[1 + \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right\} \right] + \right. \\ \left. \log \left\{ -\log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right\} \right\} \right\} & \text{se } \delta_i = 1, \\ \text{sin}(\hat{r}_{M_i}) \left\{ -2 \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right\} \right\}^{1/2} & \text{se } \delta_i = 0, \end{cases} \quad (3.24)$$

em que $\hat{r}_{M_i}$ é o resíduo martingale e $\hat{z}_i = \frac{y_i - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$.

ii. Log-gama-logística

Da mesma forma, substituindo o resíduo martingale (3.22) na expressão (3.23), o resíduo componente do desvio do modelo de regressão log-gama-logística é definido por

$$r_{D_i} = \begin{cases} \text{sin}(\hat{r}_{M_i}) \left\{ -2 \left[1 + \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right\} \right] + \right. \\ \left. \log \left\{ -\log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right\} \right\} \right\} & \text{se } \delta_i = 1, \\ \text{sin}(\hat{r}_{M_i}) \left\{ -2 \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right\} \right\}^{1/2} & \text{se } \delta_i = 0, \end{cases}$$

em que $\hat{r}_{M_i}$ é o resíduo martingale e $\hat{z}_i = \frac{y_i - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$.

3.6 Estudo de simulação

Um estudo de simulação foi realizado para pesquisar o comportamento da distribuição empírica dos resíduos martingale e componente do desvio para o modelo de regressão log-gama-Weibull e para o modelo de regressão log-gama-logística.

Assim, para o estudo de simulação, as variáveis $z_1, \dots, z_n$ da distribuição log-gama-Weibull (3.6) e da distribuição log-gama-logística (3.9) foram gerados a partir do método da aceitação-rejeição, maiores detalhes sobre o método podem ser encontrados em Ross (2006) e Bonat et al. (2012). Assim, para amostras de tamanho $n = 100$, $n = 300$ e $n = 500$, os valores dos parâmetros das distribuições de probabilidade foram fixados em $\phi = 0, 8$ e $1, 5$; $\sigma = 1, 0$; $\beta_0 = 2, 0$ e $\beta_1 = 4, 0$.

Logo, os tempos de sobrevivência foram gerados a partir do seguinte algoritmo (Zeviani, 2012) adaptado neste trabalho para dados censurados.

- i. Gerar $v \sim \text{uniforme}(a_1, b_1)$.

- ii. Gerar $u \sim \text{uniforme}(0, b_2)$.
- iii. Se $u \leq f(v) \Rightarrow z = v$, em que $f(\cdot)$ é a função densidade de probabilidade (3.6) ou (3.9).
- iv. Gerar $x_1 \sim \text{uniforme}(0, 1)$.
- v. Faça $y^* = \beta_0 + \beta_1 x_1 + \sigma z$.
- vi. Gerar $c \sim \text{uniforme}(0, \tau)$, em que τ foi ajustado até obter as porcentagens de censura à direita 10% e 30%.
- vii. Faça $y = \min(y^*, c)$ e $\delta = I(y^* \leq c)$ como sendo o indicador de censura.
- viii. Caso contrário, retorne ao passo i.

Dessa forma, para cada combinação de n , ϕ , σ e porcentagens de censuras, 1000 amostras aleatórias foram geradas por meio de simulações de Monte Carlo e para cada uma das amostras foi obtido as estimativas de máxima verossimilhança dos parâmetros dos modelos. Então, para cada modelo ajustado, os resíduos r_{M_i} (3.21-3.22) e r_{D_i} (3.24-3.25) foram calculados. Também, foram calculadas as médias das 1000 réplicas geradas e o erro quadrático médio de cada parâmetro para avaliar as estimativas de máxima verossimilhança.

Os resultados da simulação encontram-se nas Tabelas 3.1-3.2. A partir dessas tabelas é possível observar que as estimativas dos parâmetros estão próximas do verdadeiro valor, exceto quando o tamanho da amostra é pequena e para porcentagens de censura elevada. Em ambos os modelos de regressão, o erro quadrático médio dos parâmetros dos respectivos modelos aumenta quando a porcentagem de censura aumenta e diminui quando o tamanho da amostra aumenta.

Tabela 3.1 – Média e erro quadrático médio (em parênteses) das estimativas de máxima verossimilhança dos parâmetros do modelo de regressão log-gama-Weibull

n	Parâmetro	$\phi = 0,8$			$\phi = 1,5$		
		0%	10%	30%	0%	10%	30%
100	ϕ	1,030 (1,006)	1,048 (2,857)	1,185 (4,895)	4,132 (14,479)	3,957 (20,182)	5,098 (51,88)
	σ	1,056 (0,231)	1,041 (0,296)	1,080 (0,433)	0,609 (0,409)	0,597 (0,479)	0,633 (0,774)
	β_0	1,693 (1,594)	1,712 (2,758)	1,564 (4,328)	1,360 (5,391)	1,385 (6,846)	1,170 (12,678)
	β_1	4,014 (0,166)	3,980 (0,200)	3,969 (0,249)	4,003 (0,089)	4,000 (0,095)	4,006 (0,127)
300	ϕ	0,867 (0,089)	0,870 (0,084)	0,866 (0,112)	2,186 (0,760)	2,445 (0,532)	3,933 (1,131)
	σ	1,021 (0,044)	1,023 (0,043)	1,014 (0,061)	0,539 (0,067)	0,551 (0,058)	0,584 (0,097)
	β_0	1,895 (0,201)	1,894 (0,185)	1,897 (0,244)	1,783 (0,516)	1,717 (0,414)	1,506 (0,781)
	β_1	4,015 (0,054)	4,008 (0,059)	3,994 (0,074)	3,996 (0,028)	3,998 (0,030)	3,999 (0,036)
500	ϕ	0,849 (0,042)	0,855 (0,048)	0,191 (0,066)	1,637 (0,241)	1,654 (0,291)	1,713 (0,394)
	σ	1,019 (0,024)	1,023 (0,027)	0,467 (0,036)	0,512 (0,031)	0,513 (0,035)	0,517 (0,047)
	β_0	1,929 (0,098)	1,922 (0,114)	1,991 (0,143)	1,946 (0,202)	1,940 (0,230)	1,919 (0,316)
	β_1	4,003 (0,029)	3,989 (0,031)	3,998 (0,043)	3,999 (0,018)	3,999 (0,019)	3,998 (0,024)

Tabela 3.2 – Média e erro quadrático médio (em parênteses) das estimativas de máxima verossimilhança dos parâmetros do modelo de regressão log-gama-logística

n	Parâmetro	$\phi = 0,8$			$\phi = 1,5$		
		0%	10%	30%	0%	10%	30%
100	ϕ	0,912 (1,326)	0,888 (0,700)	1,081 (4,191)	2,349 (11,090)	2,183 (7,565)	2,626 (15,387)
	σ	0,969 (0,018)	0,964 (0,020)	0,962 (0,031)	0,920 (0,027)	0,920 (0,025)	0,899 (0,036)
	β_0	1,861 (0,803)	1,900 (0,703)	1,732 (1,520)	1,635 (2,370)	1,700 (2,057)	1,500 (3,063)
	β_1	4,024 (0,446)	3,990 (0,435)	3,988 (0,558)	4,014 (0,317)	3,986 (0,363)	4,990 (0,430)
300	ϕ	0,820 (0,020)	0,822 (0,025)	0,827 (0,028)	1,573 (0,285)	1,628 (0,787)	1,706 (1,795)
	σ	0,993 (0,006)	0,992 (0,006)	0,991 (0,009)	0,983 (0,003)	0,979 (0,005)	0,978 (0,007)
	β_0	1,949 (0,160)	1,959 (0,180)	1,944 (0,211)	1,943 (0,305)	1,922 (0,563)	1,881 (0,661)
	β_1	4,027 (0,138)	3,999 (0,145)	3,999 (0,180)	4,005 (0,112)	4,001 (0,117)	4,003 (0,139)
500	ϕ	0,817 (0,012)	0,812 (0,012)	0,818 (0,017)	1,532 (0,071)	1,528 (0,074)	1,543 (0,170)
	σ	0,996 (0,003)	0,995 (0,004)	0,993 (0,005)	0,989 (0,002)	0,990 (0,002)	0,987 (0,003)
	β_0	1,963 (0,092)	1,968 (0,093)	1,954 (0,118)	1,973 (0,168)	1,989 (0,169)	1,974 (0,215)
	β_1	4,003 (0,081)	4,019 (0,083)	4,009 (0,100)	4,008 (0,067)	3,992 (0,069)	4,000 (0,083)

Por outro lado, as Figuras 3.3-3.6 apresentam o gráfico dos resíduos *versus* os valores esperados das estatísticas de ordem da distribuição normal padrão. Esse gráfico é conhecido como gráfico normal de probabilidades que serve para avaliar o afastamento da suposição de normalidade dos resíduos (WEISBERG, 2005). Dessa forma, desses gráficos as seguintes interpretações são obtidas:

- i. Observou-se que a distribuição empírica do resíduo componente do desvio concorda com a distribuição normal padrão.
- ii. A distribuição empírica do resíduo martingale se apresenta, em geral, com uma acentuada assimetria.
- iii. Quando a censura decresce, a distribuição empírica do resíduo componente do desvio parece chegar próxima da distribuição normal padrão mais rapidamente do que, a do resíduo martingale considerados na análise.
- iv. Quando o tamanho da amostra aumenta, a distribuição empírica do resíduo componente do desvio parece apresentar melhor acordo com a distribuição normal.

Figura 3.3 – Gráficos normais de probabilidade para o resíduo componente do desvio (r_{D_i}) do modelo de regressão log-gama-Weibull com $\phi = 0,8$

Figura 3.4 – Gráficos normais de probabilidade para o resíduo componente do desvio (r_{D_i}) do modelo de regressão log-gama-Weibull com $\phi = 1, 5$

Figura 3.5 – Gráficos normais de probabilidade para o resíduo componente do desvio (r_{D_i}) do modelo de regressão log-gama-logística com $\phi = 0,8$

Figura 3.6 – Gráficos normais de probabilidade para o resíduo componente do desvio (r_{D_i}) do modelo de regressão log-gama-logística com $\phi = 1,5$

3.7 Aplicação

Nesse estudo, foram utilizadas informações provenientes de 49 estudantes de doutorado que entram em um determinado programa de pós-graduação entre os períodos de 1999 a 2012. O interesse de estudo é verificar se o tempo mediano de titulação se encontra dentro do prazo máximo (quatro anos), estipulado pela CAPES. Além disso, deseja-se verificar se o sexo e a idade no ano da primeira matrícula exercem alguma influência no tempo de conclusão dos alunos, bem como se existe interação entre essas covariáveis.

Dessa forma, a variável resposta foi determinada como sendo o tempo (em meses)

da primeira matrícula até a titulação (data da defesa). No entanto, os alunos que desistiram, ou que trancaram a matrícula e não retornaram ou que não concluíram o curso durante esse período, foram considerados como tempos censurados. Assim, as seguintes variáveis foram observadas:

- i. y : Logaritmo do tempo de titulação (em meses);
- ii. x_1 : Idade (em anos);
- iii. x_2 : Sexo (0= feminino, 1=masculino).

Figura 3.7 – TTT-plot (a), curva de sobrevivência estimada pelo método de Kaplan-Meier para o tempo de titulação (b) e para a covariável sexo (c)

Primeiramente, para verificar o comportamento dos dados de titulação foram construídos as curvas de Kaplan-Meier e a curva TTT, representadas na Figura 3.7. A partir dessas figuras, observa-se que o TTT-plot indica que os tempos de titulação dos alunos apresenta uma taxa de falha de crescente (Figura 3.7a). Além disso, os tempos de titulação não apresentam um patamar acima de zero, como ilustra a Figura 3.7b e existe uma evidência de que sexo dos alunos não influenciam no tempo de titulação. Dessa forma, de acordo com que foi observado na Figura 3.7, pode-se considerar os seguintes modelos:

- i. Modelo log-gama-Weibull

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \sigma z_i, \quad i = 1, \dots, 49,$$

em que $y = \log(t)$ denota o logaritmo do tempo de titulação.

- ii. Modelo log-gama-logística

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \sigma z_i, \quad i = 1, \dots, 49,$$

em que $y = \log(t)$ denota o logaritmo do tempo de titulação.

Para maximizar as funções (3.12) e (3.13) e assim encontrar as estimativas de máxima verossimilhança dos parâmetros dos modelos propostos, foi utilizada a sub-rotina `MaxBFGS` da linguagem matricial de programação `Opt` versão 6.20 com os valores iniciais $\phi = 1,000$; $\sigma = 0,760$; $\beta_0 = 4,880$; $\beta_1 = -0,153$ e $\beta_2 = 0,026$, obtidos a partir do ajuste do modelo de regressão Weibull no pacote estatístico R versão 2.15.1. Dessa forma, na Tabela 3.3 encontra-se as estimativas dos parâmetros dos modelos, os erros padrões e as significâncias dos parâmetros.

Tabela 3.3 – Estimativas de máxima verossimilhança dos parâmetros dos modelos para os dados de tempos de titulação

Parâmetro	Estimativa	Erro padrão	<i>p</i> -valor	Intervalo de confiança(95%)
Log-gama-Weibull				
ϕ	0,9520	1,0377	–	(0,0000; 2,9858)
σ	0,0630	0,0445	–	(0,0000; 0,1502)
β_0	4,1351	0,1590	0,0000	(3,8234; 4,4467)
β_1	-0,0045	0,0026	0,0884	(-0,0096; 0,0007)
β_2	-0,4237	0,1112	0,0001	(-0,6417; -0,2058)
β_3	0,0101	0,0030	0,0006	(0,0043; 0,0159)
Log-gama-logística				
ϕ	0,3020	0,1673	–	(0,0000; 0,6299)
σ	0,0236	0,0096	–	(0,0048; 0,0423)
β_0	4,1255	0,0907	0,0000	(3,9478; 4,3032)
β_1	-0,0035	0,0021	0,0953	(-0,0077; 0,0006)
β_2	-0,3555	0,0989	0,0003	(-0,5494; -0,1616)
β_3	0,0088	0,0025	0,0005	(0,0038; 0,0137)

Por meio da Tabela 3.3, observa-se que a covariável x_1 não influencia na variabilidade do logaritmo do tempo de titulação, ao nível de 5% de significância, em ambos os modelos de regressão.

Para dar continuidade a análise dos dados, considerou-se a estatística AIC_{SUR} definida na Seção 3.3 para a escolha do modelo mais adequado para descrever o comportamento do tempo de titulação. A partir dos valores das estatísticas, observa-se que o modelo de regressão log-gama-Weibull ($AIC_{SUR} = -61,0802$) é mais adequado do que o modelo de regressão log-gama-logística ($AIC_{SUR} = -61,3444$).

Por outro lado, quando considerou-se o teste de hipótese para modelos não encaixados da Seção 3.3, ao nível de 5% de significância, verifica-se que existem evidência de que o modelo de regressão log-gama-logística é equivalente ao modelo de regressão log-gama-Weibull, pois não rejeita-se a hipótese nula ($T = -0,0003$).

Portanto, levando-se em consideração os dois critérios de escolha de modelos, o restante da análise dos dados de titulação foi realizado, utilizando o modelo de regressão log-gama-Weibull.

Então, prosseguindo com a análise. O próximo passo é detectar possíveis pontos influentes no modelo de regressão log-gama-Weibull para os dados de tempo de titulação. As medidas de influência global e local foram calculadas utilizando a linguagem matricial de programação Ox versão 6.20.

Assim, a distância de Cook generalizada (3.17) e afastamento da verossimilhança (3.18) são apresentadas na Figura 3.8. A partir dessa figura, verifica-se que os pontos #18 e #21 são possíveis observações influentes.

Em relação aos gráficos de influência local, considerando as perturbações de casos ($C_{d_{max}} = 2,1805$), perturbação do logaritmo do tempo de titulação ($C_{d_{max}} = 0,3925$) e perturbação da covariável x_1 ($C_{d_{max}} = 4,4124$), observa-se que os pontos #18, #21 e #29, podem ser considerados como possíveis pontos influentes, como ilustram as Figuras 3.9-3.10.

Por outro lado, o gráfico do resíduo martingale (3.21) e do resíduo componente do desvio (3.24) *versus* os índices das observações, bem como, o gráfico do resíduo componente do desvio *versus* os valores ajustados são apresentados na Figura 3.11. Observa-se que há evidências de os pontos #111 e #136 sejam discrepantes (Figura 3.11a).

Figura 3.8 – Gráficos de medidas de influência global para o modelo de regressão log-gama-Weibull ajustado aos dados de tempo de titulação. (a) Distância de Cook generalizada e (b) Afastamento da verossimilhança

Figura 3.9 – Gráficos de influência local d_{max} do modelo de regressão log-gama-Weibull ajustado aos dados. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_1

Figura 3.10 – Gráficos de influência local total C_i do modelo de regressão log-gama-Weibull ajustado aos dados. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_1

Figura 3.11 – Análise de resíduo do modelo de regressão log-gama-Weibull ajustado aos dados. (a e b) Resíduo componente do desvio e (c) Resíduo martingale

Logo, a análise de sensibilidade (influência global e influência local) e a análise de resíduo detectaram como possíveis pontos influentes as observações #18 e #21, pois apareceram com maior frequência. Essas observações que foram apontadas como possíveis pontos influentes, são alunos que tem as seguintes descrições:

- i. A observação 18 corresponde a um aluno do sexo masculino com idade de 31 anos e com o maior tempo observado, no qual não havia titulado.
- ii. A observação 21 corresponde a um aluno do sexo feminino, também com idade de 31 anos e com o menor tempo de titulação, 35 meses.

Dessa forma, para analisar o impacto dessas observações nas estimativas dos parâmetros, reajustou-se o modelo eliminando-se, individualmente, cada observação, em seguida, eliminando as duas observações. Assim, na Tabela 3.4, são apresentadas as mudanças relativas (em porcentagens) de cada parâmetro estimado, definido por $\mathbf{RC}_{\theta_j} = \left[(\hat{\theta}_j - \hat{\theta}_{j(i)}) / \hat{\theta}_j \right] \times 100$, em que $\theta_{j(i)}$ é o estimador de máxima verossimilhança sem a i -ésima observação.

Tabela 3.4 – Mudanças relativas [-RC-em %], estimativas dos parâmetros e os correspondentes p -valores entre parênteses, excluindo as observações #18 e #21

Exclusão	$\hat{\phi}$	$\hat{\sigma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Nenhuma	-	-	-	-	-	-
	0,9520	0,0630	4,1351	-0,0045	-0,4237	0,0101
	(-)	(-)	(0,0000)	(0,0884)	(0,0001)	(0,0006)
#18	[4]	[15]	[3]	[53]	[27]	[25]
	0,9257	0,0536	4,0282	-0,0021	-0,3096	0,0076
	(-)	(-)	(0,0000)	(0,2810)	(0,0007)	(0,0019)
#21	[-351]	[-120]	[6]	[22]	[6]	[6]
	4,3450	0,1388	3,8747	-0,0035	-0,3981	0,0095
	(-)	(-)	(0,0000)	(0,1625)	(0,0004)	(0,0016)
#18 e #21	[49]	[51]	[2]	[51]	[31]	[29]
	0,4919	0,0310	4,0603	-0,0022	-0,2913	0,0072
	(-)	(-)	(0,0000)	(0,1892)	(0,0003)	(0,0009)

Por meio dos resultados da Tabela 3.4, observa-se que as estimativas de máxima verossimilhança dos parâmetros do modelo de regressão log-gama-logística são robustas sobre a deleção das observações influentes. Além disso, significância dos parâmetros estimados não se alterou (ao nível de 5% de significância) após a remoção dos casos, ou seja, não houve mudanças inferenciais após a remoção das observações consideradas influentes nos gráficos de diagnósticos. Assim sendo, as observações foram mantidas no conjunto de dados.

Por fim, para verificar a qualidade do ajuste do modelo de regressão log-gama-logística, foi construído o gráfico de probabilidade normal para o resíduo componente do desvio com envelope simulado (ATKINSON, 1985). A partir dessa figura, observa-se que há indícios de um bom ajuste do modelo de regressão log-gama-Weibull para os dados de tempo de titulação, como ilustra a Figura 3.12.

Figura 3.12 – Gráfico normal de probabilidade do modelo de regressão log-gama-Weibull para os dados de tempos de titulação

Portanto, retirando a covariável não significativa, as estimativas de máxima verossimilhança dos parâmetros do modelo de regressão log-gama-Weibull final são apresentadas na Tabela 3.5.

Tabela 3.5 – Estimativa de máxima verossimilhança dos parâmetros do modelo final para os dados de tempos de titulação

Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)	
ϕ	1,9210	2,1187	—	(0,0000; 6,0735)
σ	0,1004	0,0642	—	(0,0000; 0,2263)
β_0	3,8979	0,1678	0,0000	(3,5690; 4,2268)
β_2	-0,2850	0,0844	0,0007	(-0,4503;-0,1197)
β_3	0,0061	0,0022	0,0046	(0,0019; 0,0104)

Logo, o modelo final ajustado aos dados de tempo de titulação é expresso por

$$\hat{y}_i = 3,8979 - 0,2850 x_{i2} + 0,0061 x_{i1} x_{i2}, \quad i = 1, \dots, 49.$$

Assim, de acordo com esse modelo, como $\hat{\beta}_2$ é negativo, indica que a probabilidade de titulação reduz em 75% ($\exp(-0,2850) = 0,7520$) quando um estudante do sexo masculino é observado. Além disso, o tempo de titulação para estudantes do sexo feminino não se altera com a idade, ou seja, o tempo de titulação das mulheres é aproximadamente de 4%. Por outro lado, o tempo de titulação os estudantes do sexo masculino varia de acordo com a idade, nesse caso, estudantes mais velhos apresentam o tempo de titulação maior do que os estudantes mais novos.

4 MODELO DE REGRESSÃO PARA DADOS COM CENSURA INTERVALAR

O que distingue a análise de sobrevivência de outras áreas da estatística é a presença de censura, que é uma característica de dados de tempo de vida. Por censura, entende-se que o tempo de falha de um indivíduo não pode ser observado por algum motivo, até o final do estudo. Ao invés disso, somente uma informação parcial sobre o tempo de falha é observado. Há vários tipos de censura incluindo censura à direita, à esquerda e intervalar.

Neste trabalho, o interesse são os dados com censura intervalar, que ocorrem naturalmente em estudos em que a unidade amostral seja avaliada periodicamente. Por exemplo, supondo um estudo de AIDS, o tempo de falha associado a uma variável aleatória T é definido como sendo o tempo de infecção por HIV. Então, uma vez que a infecção por HIV somente é determinada por exames de sangue durante visitas periódicas, o tempo de ocorrência exato de T não é observado, nesse caso, o tempo de infecção é conhecido somente dentro de um intervalo, denotado por $(U, V]$, em que U representa a última avaliação com o teste negativo e V representa a primeira avaliação com teste positivo. Assim tem-se que, a variável T representa um tempo de censura intervalar.

De acordo com Lindsey e Ryan (1998), os tempos exatos de falha, bem como censuras à direita e à esquerda são casos particulares de censura intervalar. Quando $U = V$ tem-se tempos exatos de falha observados, quando $V = \infty$ tem-se censura à direita e por último quando $U = 0$ tem-se censura à esquerda, sendo este último pouco abordado na literatura.

Várias abordagens e soluções para problemas envolvendo dados com censura intervalar foram propostos na literatura, principalmente com relação à análise paramétrica. Por exemplo, Odell, Anderson e D'Agostino (1992) utilizaram a distribuição Weibull para modelos de regressão com censura intervalar, Lindsey (1998) comparou alguns modelos de regressão paramétrica para analisar dados com censura intervalar e Sparling, Younes e Lachin (2006) consideraram uma família paramétrica que acomoda efeitos fixos e covariáveis tempo. Hashimoto et al. (2010) apresentaram medidas de diagnóstico para dados com censura intervalar e Hashimoto et al. (2012) consideram um modelo de regressão log-Burr XII para dados de sobrevivência grupados.

A abordagem paramétrica é baseada na especificação de uma distribuição para os tempos de vida. Assim, utilizar distribuições na qual a sua função densidade de probabilidade seja flexível pode auxiliar na análise de dados com censura intervalar, pois os tempos exatos de falha não são conhecidos, e assim torna-se difícil utilizar técnicas descritivas para propor uma distribuição de probabilidade a dados com censura intervalar.

Por essa razão, para analisar os dados de longevidade de gado leiteiro foi considerado dois modelos de regressão baseados em duas novas distribuições de probabilidade propostas neste trabalho, uma vez, que essas distribuições apresentam função de taxa de falha com diferentes formas.

4.1 Especificação do modelo log-gama-G

Nesse caso, também foi utilizado um modelo log-liner. Assumindo que $\sigma > 0$ é o parâmetro de escala, $\beta = (\beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos, $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})$ é o vetor de covariáveis associado a i -ésima observação. Então, o modelo de regressão que relaciona a variável resposta Y a um vetor de covariáveis $\mathbf{x}$ é um modelo escrito da seguinte forma

$$y_i = \mathbf{x}_i^\top \beta + \sigma z_i, \quad i = 1, \dots, n, \quad (4.1)$$

em que y_i é o logaritmo dos tempos de sobrevivência da i -ésima observação e z_i é uma variável aleatória associada a i -ésima observação.

Entretanto, na presença de censura intervalar, os dados observados consistem de um intervalo $(\log(u_i), \log(v_i))$ para cada indivíduo, no qual esses intervalos são conhecidos e inclui y_i com probabilidade um, ou seja, $P(\log(u_i) \leq y_i \leq \log(v_i)) = 1$, e se $\log(u_i) = \log(v_i)$, então, tem-se o logaritmo do tempo exato.

Como para algumas observações, o tempo exato de falha não é conhecido, não é possível construir um histograma ou a curva TTT-plot para verificar qual o comportamento dos dados e assim associar uma possível distribuição de probabilidade.

Por essa razão, para dados na presença de censura intervalar, os modelos propostos no Capítulo 3 são uma escolha razoável para analisar dados com essa característica, pois são modelos de regressão baseado em distribuições que possuem um parâmetro de forma a mais e que possuem como casos particulares alguns modelos de regressão conhecidos.

Nesse contexto, o modelo de regressão para dados com censura intervalar, sob o enfoque paramétrico, podem assumir as seguintes distribuições de probabilidade:

a. Log-gama-Weibull

A função densidade de probabilidade e a função de sobrevivência de $Y|\mathbf{x}$ assumem as formas

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

e

$$S(y|\mathbf{x}) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left[\exp \left(\frac{y - \mathbf{x}^\top \beta}{\sigma} \right), \phi \right], \quad (4.2)$$

respectivamente.

Os casos particulares do modelo de regressão log-gama-Weibull são:

i. Modelo de regressão Weibull

O modelo é obtido substituindo $\phi = 1$ na expressão (4.2). Dessa forma, tem-se o modelo regressão Weibull, com função de sobrevivência dada por

$$S(y|\mathbf{x}_i) = \exp \left\{ - \exp \left(\frac{y - \mathbf{x}_i^\top \beta}{\sigma} \right) \right\}.$$

ii. Modelo de regressão exponencial

Quando $\phi = 1$ e $\sigma = 1$ são substituídos na expressão (4.2), tem-se o modelo de regressão exponencial, cuja função de sobrevivência é dada por

$$S(y|\mathbf{x}_i) = \exp \left\{ - \exp (y - \mathbf{x}_i^\top \boldsymbol{\beta}) \right\}.$$

iii. Modelo de regressão Rayleigh

Substituindo $\phi = 1$ e $\sigma = 1/2$ na expressão (4.2), tem-se o modelo de regressão Rayleigh, com função de sobrevivência escrita como

$$S(y|\mathbf{x}_i) = \exp \left\{ - \exp[2(y - \mathbf{x}_i^\top \boldsymbol{\beta})] \right\}.$$

b. Log-gama-logística

A função densidade de probabilidade e a função de sobrevivência de $Y|\mathbf{x}$ assumem as formas

$$f(y|\mathbf{x}) = \frac{1}{\sigma \Gamma(\phi)} \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \left[1 + \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \right]^{-2} \times \\ \left\{ \log \left[1 + \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \right] \right\}^{\phi-1}, \quad -\infty < y < \infty$$

e

$$S(y|\mathbf{x}) = 1 - \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \exp \left(\frac{y - \mathbf{x}^\top \boldsymbol{\beta}}{\sigma} \right) \right], \phi \right\}, \quad (4.3)$$

respectivamente.

Os casos particulares do modelo de regressão log-gama-logístico são:

i. Modelo de regressão log-logístico

Quando $\phi = 1$ na expressão (4.3), tem-se o modelo regressão log-logístico, com função de sobrevivência dada por

$$S(y|\mathbf{x}_i) = \left[1 + \exp \left(\frac{y - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right]^{-1}.$$

ii. Modelo de regressão Burr XII

Substituindo $\phi = 1$ e considerando uma reparametrização $k = 1(k > 0)$ na expressão (4.3), tem-se o modelo de regressão Burr XII (SILVA et al., 2008), com função de sobrevivência escrita como

$$S(y|\mathbf{x}_i) = \left[1 + \exp \left(\frac{y - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma} \right) \right]^{-k}.$$

4.2 Estimação de parâmetros por máxima verossimilhança

De acordo com Colosimo e Giolo (2006) o indivíduo que apresenta um tempo exato de falha, contribui para a função de verossimilhança com a função densidade de probabilidade, $f(t)$. Por outro lado, a contribuição de um indivíduo censurado à direita é dada pela função de sobrevivência, $S(t)$ e a contribuição de um indivíduo censurado à esquerda é dada pela função de distribuição acumulada, $F(t)$. Por fim, a contribuição de um indivíduo que apresenta um intervalo de tempo, (u, v) é dada pela probabilidade de que o tempo de ocorrência do evento pertença a este intervalo.

Dessa forma, sob a suposição de censuras independentes, a função de verossimilhança descrita por Odell et al. (1992) para um amostra de n observações independentes é dada por

$$L(\theta) = \prod_{i \in E} f(t_i) \prod_{i \in L} F(t_i) \prod_{i \in R} S(t_i) \prod_{i \in I} [S(u_i) - S(v_i)], \quad i = 1, \dots, n, \quad (4.4)$$

em que E é o conjunto de indivíduos que foram falha, L é o conjunto de indivíduos censurados à esquerda, R é o conjunto de indivíduos censurados à direita e I é o conjunto de indivíduos que apresentaram censura intervalar.

Um caso particular dessa função de verossimilhança, ocorre quando observa-se somente a presença de falhas e intervalos de tempo. Assim, o logaritmo da função de verossimilhança definida em (4.4) pode ser reescrita da seguinte forma,

$$l(\theta) = \sum_{i \in E} \log[f(t_i)] + \sum_{i \in I} \log[S(u_i) - S(v_i)], \quad i = 1, \dots, n. \quad (4.5)$$

Portanto, para um conjunto de dados que consiste de n observações, $(\log(u_1), \log(v_1), \mathbf{x}_1), \dots, (\log(u_n), \log(v_n), \mathbf{x}_n)$, em que $(\log(u_i), \log(v_i))$ representa o intervalo de tempo da ocorrência do evento de interesse e se $\log(u_i) = \log(v_i)$ tem-se o tempo de falha exato e $\mathbf{x}_i$ é o vetor de covariáveis associados a i -ésima observação ($i = 1, \dots, n$), o logaritmo da função de verossimilhança (4.5) para os dois modelos propostos são:

i. Log-gama-Weibull

$$l(\theta) = \sum_{i \in E} l_1(\lambda, z_i) + \sum_{i \in I} l_2(\lambda, zu_i, zv_i), \quad (4.6)$$

$$\text{em que } l_1(\lambda, z_i) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \phi \sum_{i \in E} z_i - \sum_{i \in E} \exp(z_i),$$

$$l_2(\lambda, zu_i, zv_i) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\exp(zv_i), \phi] - \gamma[\exp(zu_i), \phi] \},$$

$z_i = (\log(t_i) - \mathbf{x}_i^\top)/\sigma$, $zu_i = (\log(u_i) - \mathbf{x}_i^\top)/\sigma$, $zv_i = (\log(v_i) - \mathbf{x}_i^\top)/\sigma$, r é o número de falhas e r_I é o número de censuras intervalares.

A função escore é obtida, calculando-se a derivada do logaritmo da função de verossimilhança (4.6) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Dessa forma

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -n\psi(\phi) + \sum_{i \in E} z_i + \sum_{i \in I} [\dot{a}]_\phi, \quad \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = -\frac{r}{\sigma} - \frac{\phi}{\sigma} \sum_{i \in E} z_i + \frac{1}{\sigma} \sum_{i \in E} z_i \exp(z_i) + \sum_{i \in I} [\dot{a}]_\sigma$$

e

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = -\frac{\phi}{\sigma} \sum_{i \in E} x_{ij} + \frac{1}{\sigma} \sum_{i \in E} x_{ij} \exp(z_i) + \sum_{i \in I} [\dot{a}]_{\beta_j},$$

em que $\psi(\cdot)$ é a função digama, $[\dot{a}]_\phi = \frac{\partial a}{\partial \phi}$, $[\dot{a}]_\sigma = \frac{\partial a}{\partial \sigma}$, $[\dot{a}]_{\beta_j} = \frac{\partial a}{\partial \beta_j}$ e $a = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\exp(zv_i), \phi] - \gamma[\exp(zu_i), \phi] \}$.

ii. Log-gama-logística

$$l(\boldsymbol{\theta}) = \sum_{i \in E} l_1(\lambda, z_i) + \sum_{i \in I} l_2(\lambda, zu_i, zv_i), \quad (4.7)$$

em que $l_1(\lambda, z_i) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \sum_{i \in E} z_i - 2 \sum_{i \in E} lz_i + (\phi - 1) \sum_{i \in F} \log(lz_i)$

$$l_2(\lambda, zu_i, zv_i) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\log(lzv_i), \phi] - \gamma[\log(lzu_i), \phi] \},$$

$z_i = (\log(t_i) - \mathbf{x}_i^\top)/\sigma$, $zu_i = (\log(u_i) - \mathbf{x}_i^\top)/\sigma$, $zv_i = (\log(v_i) - \mathbf{x}_i^\top)/\sigma$, $lz_i = \log[1 + \exp(z_i)]$, $lzu_i = \log[1 + \exp(zu_i)]$, $lzv_i = \log[1 + \exp(zv_i)]$, r é o número de falhas e r_I é o número de censuras intervalares. A função escore é obtida, calculando-se a derivada do logaritmo da função de verossimilhança (4.7) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Dessa forma

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -n\psi(\phi) + \sum_{i \in E} \log(lz_i) + \sum_{i \in I} [\dot{a}]_\phi,$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = -\frac{r}{\sigma} - \frac{1}{\sigma} \sum_{i \in E} z_i + \frac{2}{\sigma} \sum_{i \in E} \frac{z_i \exp(z_i)}{1 + \exp(z_i)} - \frac{(\phi - 1)}{\sigma} \sum_{i \in E} \frac{z_i \exp(z_i)}{lz_i [1 + \exp(z_i)]} + \sum_{i \in I} [\dot{a}]_\sigma$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = -\frac{1}{\sigma} \sum_{i \in E} x_{ij} + \frac{2}{\sigma} \sum_{i \in E} \frac{x_{ij} \exp(z_i)}{1 + \exp(z_i)} - \frac{(\phi - 1)}{\sigma} \sum_{i \in E} \frac{x_{ij} \exp(z_i)}{lz_i [1 + \exp(z_i)]} + \sum_{i \in I} [\dot{a}]_{\beta_j},$$

em que $\psi(\cdot)$ é a função digama, $[\dot{a}]_\phi = \frac{\partial a}{\partial \phi}$, $[\dot{a}]_\sigma = \frac{\partial a}{\partial \sigma}$, $[\dot{a}]_{\beta_j} = \frac{\partial a}{\partial \beta_j}$ e $a = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\log(lzv_i), \phi] - \gamma[\log(lzu_i), \phi] \}$.

Porém, as equações $\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = 0$, $\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = 0$ e $\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = 0$ são não lineares e por isso, não possuem soluções exatas. Nesse caso, utilizando a mesma abordagem descrita na Seção 3.2, as estimativas dos parâmetros dos modelos são obtidas maximizando os logaritmos das funções de verossimilhança (4.6) e (4.7).

4.3 Análise de sensibilidade

Neste capítulo, as medidas de influência global e local são utilizadas para verificar possíveis observações influentes para o modelo de regressão log-gama-Weibull e log-gama-logística com censura intervalar. Essa etapa da análise verifica o quanto as estimativas dos modelos propostos para dados com censura intervalar são resistentes a pequenas perturbações nos dados ou nos modelos, como pode ser visto em Hashimoto et al. (2012).

4.3.1 Influência global

As medidas para análise de sensibilidade dos modelos de regressão log-gama-Weibull e log-gama-logística são obtidas sob a idéia de deleção de casos proposta por Cook (1977), na qual o enfoque é estudar o efeito de retirar o i -ésimo indivíduo de um conjunto de dados. Essa metodologia é conhecida na literatura como influência global.

Dessa forma, o modelo (4.1) com deleção de casos é denotado por

$$y_{(i)} = \mathbf{x}_{(i)}^\top \boldsymbol{\beta} + \sigma z_{(i)}, \quad i = 1, \dots, n, \quad (4.8)$$

em que o subscrito “(i)” significa que o i -ésimo caso foi retirado da amostra. Para o modelo (4.8), o logaritmo da função de verossimilhança é denotado por $l_{(i)}(\boldsymbol{\theta})$.

Logo, para $Y = \log(T)$ uma variável aleatória contínua com distribuição log-gama-Weibull (3.7) ou log-gama-logística (3.10), o logaritmo da função de verossimilhança com parâmetro $\boldsymbol{\theta}$ é denotado por $l(\boldsymbol{\theta})$.

Por outro lado, supondo que $\hat{\boldsymbol{\theta}}_{(i)} = (\hat{\phi}_{(i)}, \hat{\sigma}_{(i)}, \hat{\boldsymbol{\beta}}_{(i)}^\top)^\top$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ para $l_{(i)}(\boldsymbol{\theta})$, a influência do i -ésimo caso nas estimativas de máxima verossimilhança de $\hat{\boldsymbol{\theta}} = (\hat{\phi}, \hat{\sigma}, \hat{\boldsymbol{\beta}}^\top)^\top$ é avaliada por meio da diferença entre $\hat{\boldsymbol{\theta}}_{(i)}$ e $\hat{\boldsymbol{\theta}}$.

Se a deleção de um caso provoca variações nas estimativas, maiores atenções devem ser dadas para aquele caso. Consequentemente, se $\hat{\boldsymbol{\theta}}_{(i)}$ está distante de $\hat{\boldsymbol{\theta}}$, então, o caso é observado como uma observação influente. A primeira medida de influência global é definida como a norma padronizada de $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$, conhecida como distância de Cook generalizada, expressa por

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top \left[-\ddot{\mathbf{L}}(\boldsymbol{\theta}) \right] (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}). \quad (4.9)$$

Uma outra medida para verificar a existência de pontos influentes no modelo é definida em termos da diferença entre $l_{(i)}(\boldsymbol{\theta})$ e $l(\hat{\boldsymbol{\theta}})$. Essa medida é denominada de afastamento da verossimilhança e denotada por

$$LD_i(\boldsymbol{\theta}) = 2 \left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)}) \right]. \quad (4.10)$$

4.3.2 Influência local

O método de influência local introduzida por Cook (1986), objetiva estudar a parte de diagnóstico de um modelo por meio de pequenas perturbações no modelo ou nos dados,

ao invés de remover uma observação. Segundo Paula (2004) a idéia do método consiste em verificar a existência de pontos que sob modestas modificações no modelo causam variações desproporcionais nos resultados.

Dessa forma, seja $l(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança correspondente ao modelo postulado, em que $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos $p \times 1$ e $l(\boldsymbol{\theta}|\boldsymbol{\omega})$ o logaritmo da função de verossimilhança do modelo perturbado, em que $\boldsymbol{\omega}$ é o vetor de perturbação $q \times 1$. Suponha que existe um vetor $\boldsymbol{\omega}_0 \in \Omega$ tal que $l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\boldsymbol{\omega}_0)$ para todo o $\boldsymbol{\theta}$. Então, a medida sugerida por Cook(1986) é definida por

$$LD(\boldsymbol{\omega}) = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})], \quad (4.11)$$

em que $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denota a estimativa de máxima verossimilhança sob o modelo perturbado e $\boldsymbol{\omega} = (\omega_1, \dots, \omega_q)^\top$ é o vetor de perturbações aplicados ao modelo proposto. A expressão (4.11) é conhecida como afastamento da verossimilhança.

Assim, a proposta de influência local é estudar o comportamento da expressão (4.11) em torno do vetor de não perturbação $\boldsymbol{\omega}_0$ e selecionar uma direção unitária $\mathbf{d}$ que determina o plano normal de Ω e contém o eixo da função $LD(\boldsymbol{\omega})$. Logo, desde que $LD(\boldsymbol{\omega}) \geq 0$, $\boldsymbol{\omega}_0$ é um ponto de mínimo da expressão (4.11). A sugestão de Cook (1986) é investigar a curvatura normal da linha projetada $LD(\boldsymbol{\omega}_0 + a\mathbf{d})$, em que $a \in \mathbb{R}$, em torno de $a = 0$ para alguma direção arbitrária $\mathbf{d}$. A curvatura normal de $\boldsymbol{\theta}$ na direção de $\mathbf{d}$ para o modelo proposto é dado por

$$C_{\mathbf{d}}(\boldsymbol{\theta}) = 2|\mathbf{d}^\top \Delta [\ddot{\mathbf{L}}(\boldsymbol{\theta})]^{-1} \Delta \mathbf{d}|,$$

em que Δ é uma matriz $(p + 2 \times n)$ dependendo do esquema de perturbação e seus elementos são dados por $\Delta_{ij} = \partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega}) / \partial \theta_j \partial \omega_i$, $i = 1, \dots, n$ e $j = 1, \dots, p + 2$, avaliada em $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ e $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0$.

Cook (1986) sugere tomar a direção correspondente a maior curvatura, denotada por $\mathbf{d}_{\max}$, que é o maior autovetor e o seu autovalor $C_{\mathbf{d}_{\max}}$ da matriz $\ddot{F} = \Delta^\top [\ddot{\mathbf{L}}(\boldsymbol{\theta})]^{-1} \Delta$.

Por outro lado, Lesaffre e Verbeke (1998) sugerem considerar a direção do i -ésimo indivíduo que corresponde ao vetor $\mathbf{d}_i = (0, \dots, 1, \dots, 0)^\top$, tal que o i -ésimo elemento recebe o valor um. Nesse caso, a curvatura normal chamada de influência local total para o i -ésimo indivíduo é dada por

$$C_i = 2|\mathbf{d}_i^\top \Delta [\ddot{\mathbf{L}}(\boldsymbol{\theta})]^{-1} \Delta \mathbf{d}_i|,$$

em que Δ_i denota a i -ésima linha da matriz Δ .

Assim, os gráficos dos elementos de $\mathbf{d}_{\max}$ *versus* o índice das observações bem como os gráficos dos elementos C_i *versus* o índice das observações são utilizados para detectar possíveis pontos influentes.

Como as medidas de influência local são calculadas assumindo um esquema de perturbação no modelo ou nos dados, diferentes esquemas de perturbação podem ser considerados, de acordo com o objetivo da análise. Dessa forma, para o modelo definido em (4.1), os

seguintes esquemas de perturbação foram utilizados no modelo de regressão log-gama-Weibull e no modelo de regressão log-gama-logística:

a. *Perturbação de casos*

Assumindo que $0 \leq \omega_i \leq 1$ e $\omega_0 = (1, \dots, 1)^\top$ é o vetor de não perturbação. O logaritmo da função de verossimilhança perturbada, para cada modelo, incorpora diferentes pesos para cada elemento do conjunto de dados e é definido por:

i. Log-gama-Weibull

$$l(\theta) = \sum_{i \in E} \omega_i l_1(\lambda, z_i) + \sum_{i \in I} \omega_i l_2(\lambda, zu_i, zv_i),$$

$$\text{em que } l_1(\lambda, z_i) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \phi \sum_{i \in E} z_i - \sum_{i \in E} \exp(z_i),$$

$$l_2(\lambda, zu_i, zv_i) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\exp(zv_i), \phi] - \gamma[\exp(zu_i), \phi] \},$$

ii. Log-gama-logística

$$l(\theta) = \sum_{i \in E} \omega_i l_1(\lambda, z_i) + \sum_{i \in I} \omega_i l_2(\lambda, zu_i, zv_i),$$

$$\text{em que } l_1(\lambda, z_i) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \sum_{i \in E} z_i - 2 \sum_{i \in E} lz_i + (\phi - 1) \sum_{i \in F} \log(lz_i)$$

$$l_2(\lambda, zu_i, zv_i) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\log(lzv_i), \phi] - \gamma[\log(lzu_i), \phi] \},$$

b. *Perturbação da variável resposta*

Para verificar a sensibilidade do modelo (4.1), considera-se que a variável resposta $y_i = \log(t_i)$ ($i = 1, \dots, n$) é submetida a um esquema de perturbação aditiva por meio da adição de um vetor, denotado por ω , de pequenas perturbações, tal que $y_i^* = y_i + \omega_i \text{sd}(y_i)$, em que $\text{sd}(y_i)$ é um fator escala que pode ser a estimativa do desvio padrão do logaritmo do tempos de falha e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Nesse caso, para $\omega_0 = (0, \dots, 0)^\top$ sendo o vetor de não perturbação, o logaritmo da função de verossimilhança perturbada para cada um modelos é expresso por:

i. Log-gama-Weibull

$$l(\theta) = \sum_{i \in E} l_1(\lambda, z_i^*) + \sum_{i \in I} l_2(\lambda, zu_i, zv_i),$$

em que $l_1(\lambda, z_i^*) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \phi \sum_{i \in E} z_i^* - \sum_{i \in E} \exp(z_i^*)$,

$$l_2(\lambda, zu_i, zv_i) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\exp(zv_i), \phi] - \gamma[\exp(zu_i), \phi] \},$$

$$z_i^* = (y_i^* - \mathbf{x}_i^\top) / \sigma.$$

ii. Log-gama-logística

$$l(\boldsymbol{\theta}) = \sum_{i \in E} l_1(\lambda, z_i^*) + \sum_{i \in I} l_2(\lambda, zu_i, zv_i),$$

em que $l_1(\lambda, z_i^*) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \sum_{i \in E} z_i^* - 2 \sum_{i \in E} lz_i^* + (\phi - 1) \sum_{i \in E} \times$

$$\log(lz_i^*), \quad l_2(\lambda, zu_i, zv_i) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\log(lzv_i), \phi] - \gamma[\log(lzu_i), \phi] \},$$

$$z_i^* = (y_i^* - \mathbf{x}_i^\top) / \sigma \text{ e } lz_i^* = \log[1 + \exp(z_i^*)].$$

c. **Perturbação de uma covariável**

Uma outra forma de avaliar a sensibilidade do modelo (4.1) é considerar pequenas perturbações em uma determinada covariável contínua, denotada por X_j . Nesse caso, a covariável também é submetida a um esquema de perturbação aditiva, de tal forma que $x_{ij}^* = x_{ij} + \omega_i \text{sd}(x_{ij})$, em que $\text{sd}(x_{ij})$ é um fator de escala que pode ser a estimativa do desvio padrão da covariável perturbada e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Dessa forma, considerando que $\mathbf{x}_i^{*\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j [x_{ij} + \omega_i \text{sd}(x_{ij})] + \dots + \beta_p x_{ip}$ e $\boldsymbol{\omega}_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação, o logaritmo da função de verossimilhança perturbada para cada um modelos é dada por:

i. Log-gama-Weibull

$$l(\boldsymbol{\theta}) = \sum_{i \in E} l_1(\lambda, z_i^*) + \sum_{i \in I} l_2(\lambda, zu_i^*, zv_i^*),$$

em que $l_1(\lambda, z_i^*) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \phi \sum_{i \in E} z_i^* - \sum_{i \in E} \exp(z_i^*)$,

$$l_2(\lambda, zu_i^*, zv_i^*) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\exp(zv_i^*), \phi] - \gamma[\exp(zu_i^*), \phi] \},$$

$$z_i^* = (y_i - \mathbf{x}_i^\top) / \sigma, zu_i^* = (\log(u_i) - \mathbf{x}_i^\top) / \sigma, zv_i^* = (\log(v_i) - \mathbf{x}_i^\top) / \sigma \text{ e } \mathbf{x}_i^{*\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_t x_{itw} + \dots + \beta_p x_{ip}.$$

ii. Log-gama-logística

$$l(\boldsymbol{\theta}) = \sum_{i \in E} l_1(\lambda, z_i^*) + \sum_{i \in I} l_2(\lambda, zu_i^*, zv_i^*),$$

em que $l_1(\lambda, z_i^*) = -r \log(\sigma) - r \log[\Gamma(\phi)] + \sum_{i \in E} z_i^* - 2 \sum_{i \in E} l z_i^* + (\phi - 1) \sum_{i \in E} \times$
 $\log(l z_i^*), \quad l_2(\lambda, z u_i^*, z v_i^*) = -r_I \log[\Gamma(\phi)] + \sum_{i \in I} \log \{ \gamma[\log(l z v_i^*), \phi] - \gamma[\log(l z u_i^*), \phi] \},$
 $z_i^* = (y_i - \mathbf{x}_i^\top) / \sigma, z u_i^* = (\log(u_i) - \mathbf{x}_i^\top) / \sigma, z v_i^* = (\log(v_i) - \mathbf{x}_i^\top) / \sigma, l z_i^* = \log[1 +$
 $\exp(z_i^*)], l z u_i^* = \log[1 + \exp(z_i^*)], l z v_i^* = \log[1 + \exp(z_i^*)] \text{ e } \mathbf{x}_i^{*\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} +$
 $\dots + \beta_t x_{it\omega} + \dots + \beta_p x_{ip}.$

Para esses três esquemas de perturbação calculou-se, numericamente, a matriz de curvaturas máximas definida por

$$\Delta = (\Delta_{vi})_{(p+2) \times n} = \left[\frac{\partial^2 l(\boldsymbol{\theta} | \boldsymbol{\omega})}{\partial \theta_v \partial \omega_i} \right]_{(p+2) \times n},$$

em que $v = 1, \dots, p+2, i = 1, \dots, n$ e $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top$ é o vetor de perturbação, isto é, os pesos que perturbam o modelo (4.1) com distribuição log-gama-Weibull ou com distribuição log-gama-logística ou as observações.

4.4 Análise de resíduos

A fim de estudar as suposições dos erros bem como a presença de observações discrepantes na presença de censura intervalar, Farrington (2000) propôs resíduos nesse contexto, mas para modelos semi-paramétricos e Hashimoto et al. (2010) propuseram resíduos na presença de censura intervalar sob o enfoque paramétrico. Os resíduos apresentados neste capítulo têm a mesma abordagem adotada por Hashimoto et al. (2010), porém, são definidos na presença de observações exatas e para censura intervalar.

4.4.1 Resíduo de martingale

Uma adaptação do resíduo de Cox-Snell (COX; SNELL, 1968) para os dados de censura intervalar é uma versão do resíduo discutido por Lagakos (1980), também chamado de resíduo martingale (BARLOW; PRENTICE, 1988). Os resíduos de martingale são assimétricos e assumem valores máximos em $+1$ e valores mínimos em $-\infty$. Assim, o resíduo martingale para dados com censura intervalar é expresso da seguinte forma

$$r_{M_i} = \frac{\hat{S}[\log(u_i)] \log \hat{S}[\log(u_i)] - \hat{S}[\log(v_i)] \log \hat{S}[\log(v_i)]}{\hat{S}[\log(u_i)] - \hat{S}[\log(v_i)]},$$

em que $S(\cdot)$ é a função de sobrevivência.

Dessa forma, o resíduo martingale para os modelos de regressão log-gama-Weibull e log-gama-logística são dados, respectivamente, por:

i. Log-gama-Weibull

$$r_{M_i} = \begin{cases} 1 + \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma[\exp(\hat{z}_i), \hat{\phi}] \right\} & \text{se } i \in E, \\ \frac{\hat{S}[\log(u_i)] \log \hat{S}[\log(u_i)] - \hat{S}[\log(v_i)] \log \hat{S}[\log(v_i)]}{\hat{S}[\log(u_i)] - \hat{S}[\log(v_i)]} & \text{se } i \in I, \end{cases} \quad (4.12)$$

em que $\hat{S}(\cdot)$ é a função de sobrevivência definida em (4.2), $\hat{z}_i = \frac{\log(t_i) - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$, $\hat{z}u_i = \frac{\log(u_i) - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$ e $\hat{z}v_i = \frac{\log(v_i) - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$.

ii. Log-gama-logística

$$r_{M_i} = \begin{cases} 1 + \log \left\{ 1 - \frac{1}{\Gamma(\hat{\phi})} \gamma[\log[1 + \exp(\hat{z}_i)], \hat{\phi}] \right\} & \text{se } i \in E, \\ \frac{\hat{S}[\log(u_i)] \log \hat{S}[\log(u_i)] - \hat{S}[\log(v_i)] \log \hat{S}[\log(v_i)]}{\hat{S}[\log(u_i)] - \hat{S}[\log(v_i)]} & \text{se } i \in I, \end{cases} \quad (4.13)$$

em que $\hat{S}(\cdot)$ é a função de sobrevivência definida em (4.3), $\hat{z}_i = \frac{\log(t_i) - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$, $\hat{z}u_i = \frac{\log(u_i) - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$ e $\hat{z}v_i = \frac{\log(v_i) - \mathbf{x}_i^\top \hat{\beta}}{\hat{\sigma}}$.

4.4.2 Resíduo componente do desvio modificado

O resíduo componente do desvio, apresentado por McCullagh e Nelder (1989), é muito utilizado no contexto de modelos lineares generalizados para verificar a adequação do modelo utilizado. Devido a esse apreço pelo resíduo, Therneau, Grambsch e Fleming (1990) baseados no resíduo componente do desvio, estenderam esse resíduo para os modelos de riscos proporcionais.

Dessa forma, o resíduo componente do desvio utilizado no contexto de análise de sobrevivência é definido como

$$r_{D_i} = \text{sinal}(r_{M_i}) [-2 \{r_{M_i} + \log(1 - r_{M_i})\}]^{1/2}.$$

em que r_{M_i} é o resíduo martingale. O resíduo componente do desvio pode ser considerado como uma modificação feita no resíduo martingale com a finalidade de torná-lo mais simétrico ao redor do zero (COLLETT, 2003).

Entretanto, esse resíduo não é ideal para situações censuras intervalares são observadas. Por essa razão, Hashimoto et al. (2010) propuseram um resíduo componente do desvio modificado para levar em consideração a informação dos tempos intervalares.

Assim, o resíduo componente do desvio modificado para os modelos de regressão log-gama-Weibull e log-gama-logística são expressos, respectivamente, por:

i. Log-gama-Weibull

$$r_{D_i} = \begin{cases} \text{sinal}(\hat{r}_{M_i}) \left\{ -2 \left\{ 1 + \log \left[1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right] + \right. \right. \\ \left. \left. \log \left\{ -\log \left[1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\exp(\hat{z}_i), \hat{\phi} \right] \right] \right\} \right\} \right\}^{1/2} & \text{se } i \in E, \\ \text{sinal}(\hat{r}_{M_i}) \left\{ -2 [\hat{r}_{M_i} + \log(1 - \hat{r}_{M_i})] \right\}^{1/2} & \text{se } i \in I, \end{cases} \quad (4.14)$$

$$\text{em que } \hat{r}_{M_i} = \frac{\{\Gamma(\hat{\phi}) - \gamma[\exp(\hat{z}u_i), \hat{\phi}]\} \log\{1 - \frac{1}{\Gamma(\hat{\phi})} \gamma[\exp(\hat{z}u_i), \hat{\phi}]\} - \{\Gamma(\hat{\phi}) - \gamma[\exp(\hat{z}v_i), \hat{\phi}]\} \log\{1 - \frac{1}{\Gamma(\hat{\phi})} \gamma[\exp(\hat{z}v_i), \hat{\phi}]\}}{\gamma[\exp(\hat{z}v_i), \hat{\phi}] - \gamma[\exp(\hat{z}u_i), \hat{\phi}]},$$

$$\hat{z}_i = \frac{\log(t_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}}, \hat{z}u_i = \frac{\log(u_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \text{ e } \hat{z}v_i = \frac{\log(v_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}}.$$

ii. Log-gama-logística

$$r_{D_i} = \begin{cases} \text{sinal}(\hat{r}_{M_i}) \left\{ -2 \left\{ 1 + \log \left[1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right] + \right. \right. \\ \left. \left. \log \left\{ -\log \left[1 - \frac{1}{\Gamma(\hat{\phi})} \gamma \left[\log[1 + \exp(\hat{z}_i)], \hat{\phi} \right] \right] \right\} \right\} \right\}^{1/2} & \text{se } i \in E, \\ \text{sinal}(\hat{r}_{M_i}) \left\{ -2 [\hat{r}_{M_i} + \log(1 - \hat{r}_{M_i})] \right\}^{1/2} & \text{se } i \in I, \end{cases} \quad (4.15)$$

$$\text{em que } \hat{r}_{M_i} = \frac{\{\Gamma(\hat{\phi}) - \gamma[\exp(\hat{z}u_i), \hat{\phi}]\} \log\{1 - \frac{1}{\Gamma(\hat{\phi})} \gamma[\exp(\hat{z}u_i), \hat{\phi}]\} - \{\Gamma(\hat{\phi}) - \gamma[\exp(\hat{z}v_i), \hat{\phi}]\} \log\{1 - \frac{1}{\Gamma(\hat{\phi})} \gamma[\exp(\hat{z}v_i), \hat{\phi}]\}}{\gamma[\exp(\hat{z}v_i), \hat{\phi}] - \gamma[\exp(\hat{z}u_i), \hat{\phi}]},$$

$$\hat{z}_i = \frac{\log(t_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}}, \hat{z}u_i = \frac{\log(u_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}} \text{ e } \hat{z}v_i = \frac{\log(v_i) - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{\hat{\sigma}}.$$

O gráfico do resíduo deviance modificado bem como o do resíduo martingale *versus* os tempos observados fornecem uma forma de verificar a adequação do modelo ajustado e auxiliam a encontrar observações atípicas.

4.4.3 Estudo de simulação

Um estudo de simulação foi realizado para investigar o comportamento da distribuição empírica do resíduo componente do desvio (r_{D_i}) dos modelos de regressão log-gama-Weibull e log-gama-logística para dados com censura intervalar. Assim, para amostras de tamanho $n = 100$, $n = 300$ e $n = 500$, considerando os valores dos parâmetros fixados em $\phi = 0,6$; $\sigma = 1,0$; $\beta_0 = 3,8$ e $\beta_1 = 2$, os tempos de sobrevivência foram gerados a partir do algoritmo apresentado na Seção 3.5 do Capítulo 3, utilizando o método da aceitação-rejeição.

Entretanto, para os tempos com censura intervalar, os intervalos de tempo (u_i, v_i) foram gerados seguindo o mesmo enfoque adotado por Hashimoto et al. (2010), de tal forma que $\exp(y_i) \in (u_i, v_i)$ com probabilidade um.

Dessa forma, para cada combinação de n , ϕ , σ e porcentagens de censuras, 1000 amostras aleatórias foram geradas por meio de simulações de Monte Carlo e para cada uma das amostras foi obtido as estimativas de máxima verossimilhança dos parâmetros dos modelos. Então, para cada modelo ajustado, os resíduos r_{D_i} foram calculados e também foram obtidas as médias das 1000 réplicas geradas e o erro quadrático médio de cada parâmetro para avaliar as estimativas de máxima verossimilhança.

Os resultados da simulação encontram-se nas Tabelas 4.1-4.2. Essas figuras representam o gráfico dos resíduos *versus* os valores esperados das estatísticas de ordem da distribuição normal padrão para cada modelo de regressão. Esse gráfico é conhecido como gráfico normal de probabilidades e serve para avaliar o afastamento da suposição de normalidade dos resíduos (WEISBERG, 2005).

Desses gráficos as seguintes interpretações são obtidas:

- i. Observou-se claramente que a distribuição empírica do resíduo deviance modificado concorda com a distribuição normal padrão;
- ii. Quando a censura decresce, a distribuição empírica do resíduo deviance modificado parece chegar próximo a distribuição normal padrão mais rapidamente do que os outros dois resíduos considerados na análise;
- iii. Quando o tamanho da amostra aumenta, a distribuição empírica do resíduo deviance modificado parece apresentar melhor acordo com a distribuição normal.

Figura 4.1 – Gráficos normais de probabilidade para o resíduo componente do desvio modificado (r_{D_i}) do modelo de regressão log-gama-Weibull.

Figura 4.2 – Gráficos normais de probabilidade para o resíduo componente do desvio modificado (r_{D_i}) do modelo de regressão log-gama-logística

4.5 Aplicação

O conjunto de dados utilizado para exemplificar a metodologia proposta neste Capítulo foram obtidos a partir de um estudo de rebanhos de vacas leiteiras, desenvolvido pelo Departamento de Medicina Veterinária da Universidade Federal de Lavras. O objetivo dessa pesquisa era verificar se a suplementação oferecida ao rebanho estava influenciando no tempo de ovulação dos animais. O experimento foi realizado com 50 vacas leiteiras divididas em dois grupos, na qual o grupo controle corresponde aos animais sem suplementação e o grupo tratamento corresponde aos animais tratados à base de uma suplementação para induzir a ovulação.

Nesse estudo, a variável resposta é o tempo (em dias) após o parto até a primeira ovulação, porém, somente para alguns animais foi possível observar o tempo exato da primeira ovulação. Para os outros animais, a única informação que se tem é que o tempo exato de falha ocorreu em um intervalo de tempo, o que caracteriza a presença de censura intervalar. Dessa forma, as seguintes variáveis foram observadas:

- i. $[\log(u_i), \log(v_i)]$: Logaritmo do intervalo de tempo de ovulação (em dias);
- ii. x_1 : Tratamento (0= controle, 1=suplementação);
- iii. x_2 : Ovário (0= direito, 1=esquerdo).
- iv. x_3 : Categoria (duas crias, três crias, quatro crias, cinco crias).

Nesse caso, considera-se a utilização de variáveis *dummys*, ou seja, tem-se as seguintes variáveis *dummys* para a covariável x_3 :

$$x_3(\text{categoria}) = \begin{cases} d_1 = 1 & \text{se três crias, 0 se duas crias} \\ d_2 = 1 & \text{se quatro crias, 0 se duas crias} \\ d_3 = 1 & \text{se cinco crias, 0 se duas crias} \end{cases}$$

Observa-se que conjunto de dados é composto de covariáveis dicotômicas, exceto pela covariável x_3 que é politômica. Então, foram construídas curvas de Kaplan-Meier para verificar o comportamento dessas covariáveis em relação ao tempo de ovulação. Dessa forma, a Figura 4.3 representa as curvas de sobrevivência estimadas pelo método da Kaplan-Meier e estratificadas pelas respectivas covariáveis.

Figura 4.3 – Curva de sobrevivência estimada pelo método de Kaplan-Meier para a covariável x_1 (a), para a covariável x_3 (b) e para a covariável x_2 (c)

Por outro lado, Figura 4.4 mostra a curva de sobrevivência estimada pelo método de Kaplan-Meier e pelos dois modelos propostos. A partir desses gráficos, observa-se que existe

evidências de que as covariáveis não exercem influência no tempo de ovulação dos animais e que os modelos seriam razoavelmente aceitáveis para analisar esse conjunto de dados.

Figura 4.4 – Curvas de sobrevivência estimadas pelos os dois modelos *versus* a curva estimada pelo método de Kaplan-Meier para dados de ovulação

Logo, de acordo com que foi observado na Figura 4.3, pode-se considerar os seguintes modelos:

i. Modelo log-gama-Weibull

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 d_{1i} + \beta_4 d_{2i} + \beta_5 d_{3i} + \sigma z_i, \quad i = 1, \dots, 50,$$

em que $y_i = \log(t_i)$ denota o logaritmo do tempo de ovulação que pertence ao intervalo $(\log(u_i), \log(v_i)]$.

ii. Modelo log-gama-logística

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 d_{1i} + \beta_4 d_{2i} + \beta_5 d_{3i} + \sigma z_i, \quad i = 1, \dots, 50,$$

em que $y_i = \log(t_i)$ denota o logaritmo do tempo de ovulação que pertence ao intervalo $(\log(u_i), \log(v_i)]$.

Para maximizar as funções (4.6) e (4.7) e assim encontrar as estimativas de máxima verossimilhança dos parâmetros dos modelos propostos, foi utilizada a sub-rotina `MaxBGFS` da linguagem matricial de programação `Opt` versão 6.20 com os valores iniciais $\phi = 1,000$;

$\sigma = 0,760$; $\beta_0 = 4,880$; $\beta_1 = -0,153$ e $\beta_2 = 0,026$, obtidos a partir do ajuste do modelo de regressão Weibull no pacote estatístico R versão 2.14.0.

Dessa forma, na Tabela 4.1 encontra-se as estimativas de máxima verossimilhança, os erros padrões e os intervalos de confiança dos parâmetros dos dois modelos.

Tabela 4.1 – Estimativa de máxima verossimilhança dos parâmetros dos modelos para os dados de tempos de ovulação

Parâmetro	Estimativa	Erro padrão	p-valor	Intervalo de confiança(95%)
Log-gama-Weibull				
ϕ	0,5912	0,6790	–	(0,0000; 1,9221)
σ	0,3571	0,2655	–	(0,0000; 0,8775)
β_0	3,8911	0,4375	0,0000	(3,0337; 4,7485)
β_1	0,0212	0,1434	0,8824	(-0,2598; 0,3022)
β_2	0,1832	0,1776	0,3024	(-0,1649; 0,5312)
β_3	-0,1422	0,1863	0,4452	(-0,5073; 0,2228)
β_4	-0,0909	0,0829	0,2731	(-0,2533; 0,0716)
β_5	-0,3615	0,2655	0,1734	(-0,8819; 0,1589)
Log-gama-logística				
ϕ	0,3811	0,2233	–	(0,0000; 0,8187)
σ	0,2145	0,0675	–	(0,0821; 0,3469)
β_0	3,8539	0,2197	0,0000	(3,4233; 4,2844)
β_1	0,1150	0,1572	0,4642	(-0,1930; 0,4230)
β_2	0,0920	0,1838	0,6168	(-0,2683; 0,4523)
β_3	-0,0619	0,2044	0,7621	(-0,4624; 0,3387)
β_4	-0,1634	0,1971	0,4072	(-0,5498; 0,2230)
β_5	-0,2219	0,2504	0,3757	(-0,7125; 0,2689)

Considerando um nível de significância de 5%, nenhuma variável explicativa influência na variabilidade do tempo de ovulação do rebanho de vacas leiteiras, em ambos os modelos ajustados, como observado inicialmente pelas curvas de Kaplan-Meier.

Por outro lado, como os tempos exatos de falha não são conhecidos para algumas observações, o critério de seleção de modelos adotado nesse caso, é a estatística AIC_{SUR} definida em (3.14). Observa-se que o modelo de regressão log-gama-Weibull ($AIC_{SUR}=176,7208$) é mais adequado do que o modelo de regressão log-gama-logística ($AIC_{SUR}=177,0734$), pois apresenta o menor valor da estatística. Portanto, para a análise de diagnóstico será utilizado o modelo de regressão log-gama-Weibull.

Assim, para detectar possíveis pontos influentes no modelo de regressão log-gama-Weibull para os dados de tempo de ovulação foram calculadas as medidas de influência global e local utilizando a linguagem matricial de programação O× versão 6.20.

As medidas da distância de Cook generalizada (4.9) e afastamento da verossimilhança (4.10) são apresentadas na Figura 4.5. A partir dessa figura, verifica-se que os pontos #25 e #31 são possíveis observações influentes.

Nas Figuras 4.6-4.7, têm-se os gráficos de influência local, considerando as perturbações de caso ($C_{d_{max}} = 3, 3117$) e perturbação do logaritmo do tempo de ovulação ($C_{d_{max}} = 0, 0020$). A partir desses gráficos é possível identificar os pontos #25, #31 e #40 como possíveis observações influentes.

Por outro lado, o gráfico do resíduo martingale (4.12) e do resíduo componente do desvio (4.14) *versus* os índices das observações são apresentados na Figura 4.8. Observa-se que há evidências de os pontos #25, #27, #28, #38 e #41 sejam discrepantes (Figura 4.8).

Figura 4.5 – Gráficos de medidas de influência global para o modelo de regressão log-gama-Weibull ajustado aos dados de tempo de ovulação. (a) Distância de Cook generalizada e (b) Afastamento da verossimilhança

Figura 4.6 – Gráficos de influência local d_{max} do modelo de regressão log-gama-Weibull ajustado. (a) Perturbação de casos e (b) Perturbação da resposta

Figura 4.7 – Gráficos de influência local total C_i do modelo de regressão log-gama-Weibull ajustado. (a) Perturbação de casos e (b) Perturbação da resposta

Figura 4.8 – Análise de resíduo do modelo de regressão log-gama-Weibull ajustado. (a) Resíduo componente do desvio e (b) Resíduo martingale

Logo, a análise de sensibilidade (influência global e influência local) e a análise de resíduo detectaram como possíveis pontos influentes as observações #25 e #31, pois apareceram com maior frequência. Essas observações que foram apontadas como possíveis pontos influentes, são animais que tem as seguintes descrições:

- i. A observação 25 corresponde ao tempo de ovulação pertence ao intervalo de $(91, 93]$, medido no ovário direito e de um animal que teve duas crias.
- ii. A observação 31 corresponde ao tempo de ovulação exato 56, medido no ovário esquerdo e de um animal que já teve três crias.

Assim, para analisar o impacto dessas observações nas estimativas dos parâmetros, reajustou-se o modelo eliminando-se, individualmente, cada observação, em seguida, eliminando as duas observações.

Desse modo, na Tabela 4.2, são apresentadas as mudanças relativas (em porcentagens) de cada parâmetro estimado, definido por $\mathbf{RC}_{\theta_j} = \left[(\hat{\theta}_j - \hat{\theta}_{j(i)}) / \hat{\theta}_j \right] \times 100$, em que $\theta_{j(i)}$ é o estimador de máxima verossimilhança sem a i -ésima observação.

Tabela 4.2 – Mudanças relativas [-RC-em %], estimativas dos parâmetros e os correspondentes p -valores entre parênteses, excluindo as observações #25 e #31

Exclusão	$\hat{\phi}$	$\hat{\sigma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Nenhuma	-	-	-	-				
	0,5912	0,3571	3,8911	0,0212	0,1832	-0,1422	-0,0909	-0,3615
	(-)	(-)	(0,0000)	(0,8824)	(0,3024)	(0,4452)	(0,2731)	(0,1734)
#25	[18]	[17]	[2]	[-621]	[29]	[89]	[92]	[40]
	0,4865	0,2952	3,7989	0,1528	0,1298	-0,0155	-0,0074	-0,2177
	(-)	(-)	(0,0000)	(0,2764)	(0,4562)	(0,9293)	(0,9358)	(0,3553)
#31	[1]	[0]	[0]	[62]	[-1]	[-10]	[2]	[0]
	0,5853	0,3561	3,8797	0,0081	0,1855	-0,1566	-0,0891	-0,3612
	(-)	(-)	(0,0000)	(0,9646)	(0,4371)	(0,5912)	(0,2736)	(0,2179)
#25 e #31	[27]	[24]	[1]	[-253]	[-1]	[-10]	[72]	[23]
	0,4328	0,2722	3,8519	0,0749	0,1967	-0,1568	-0,0252	-0,2781
	(-)	(-)	(0,0000)	(0,6291)	(0,2848)	(0,4613)	(0,7679)	(0,2809)

Por meio dos resultados da Tabela 4.2, observa-se que as estimativas de máxima verossimilhança dos parâmetros do modelo de regressão log-gama-logística são robustas sobre a deleção das observações influentes. Além disso, significância dos parâmetros estimados não se alterou (ao nível de 5% de significância) após a remoção dos casos, ou seja, não houve mudanças inferenciais após a remoção das observações consideradas influentes nos gráficos de diagnósticos. Assim sendo, as observações foram mantidas no conjunto de dados.

Por fim, para verificar a qualidade do ajuste do modelo de regressão log-gama-Weibull, foi construído o gráfico de probabilidade normal para o resíduo componente do desvio com envelope simulado (ATKINSON, 1985), como ilustra a Figura 4.9. A partir dessa figura, observa-se que há indícios de um bom ajuste do modelo de regressão log-gama-Weibull para os dados de tempo de ovulação.

Portanto, retirando a covariável não significativa, as estimativas de máxima verossimilhança dos parâmetros do modelo de regressão log-gama-Weibull final são apresentadas na Tabela 4.3.

Tabela 4.3 – Estimativa de máxima verossimilhança dos parâmetros do modelo final para os dados de tempos de ovulação

Parâmetro	Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)
ϕ	1,6532	1,8172	–	(0,0000; 5,2148)
σ	0,6855	0,4076	–	(0,0000; 1,4844)
β_0	3,2409	1,0654	0,0024	(1,1528; 5,3290)

Desse modo, o modelo final ajustado aos dados de tempo de ovulação de vacas leiteiras é expresso apenas em função do intercepto,

$$\hat{y}_i = 3,2409; \quad i = 1, \dots, 50.$$

Assim, a partir do modelo ajustado, é possível obter as seguintes informações sobre o conjunto de dados de tempo de ovulação:

- i. Ao nível de 5% de significância, observa-se que a suplementação administrada aos animais não tem efeito sobre o tempo de ovulação, bem como, o número de crias que o animal teve.
- ii. A chance de um animal não ovular depois de 35 dias após o parto é estimada por meio da função de sobrevivência em (2.5). Assim,

$$\hat{S}(35) = 1 - \frac{1}{\Gamma(1,6532)} \gamma \left[\left(\frac{35}{\exp(3,2409)} \right)^{1/0,6855}; 1,6532 \right] = 0,42.$$

Portanto, a chance de um animal não ovular depois de 35 dias após o parto é de 42%.

- iii. Analogamente, a chance de um animal não ovular depois de 70 dias é estimada por

$$\hat{S}(70) = 1 - \frac{1}{\Gamma(1,6532)} \gamma \left[\left(\frac{70}{\exp(3,2409)} \right)^{1/0,6855}; 1,6532 \right] = 0,04.$$

Logo, a chance do animal ovular aumenta, quando o tempo aumenta.

Figura 4.9 – Gráfico normal de probabilidade do modelo de regressão log-gama-Weibull para os dados de ovulação

5 MODELO DE LONGA-DURAÇÃO

No contexto de análise de sobrevivência, determinados estudos caracterizam-se por apresentar uma fração significativa de sobreviventes, ou seja, unidades experimentais que não apresentam o evento de interesse, mesmo após um longo período de acompanhamento. Por exemplo, em estudos de tratamento de câncer, pode ocorrer de uma grande parte da amostra não apresentar a reincidência da doença. Nesse caso, é plausível admitir que entre as observações censuradas, há a possibilidade de existir uma proporção de pacientes curados.

Por essa razão, considerar os modelos de sobrevivência usuais, os quais supõem que a função de sobrevivência converge para zero quando a variável tempo tende a infinito (função de sobrevivência própria) pode não ser adequado.

Então, para construir um modelo que leva em consideração a informação de indivíduos não susceptíveis ao evento de interesse, Berkson e Gage (1952) propuseram um modelo de longa-duração, utilizado para estimar essa fração de curados. Em geral, os modelos com fração de cura são utilizados para analisar dados de sobrevivência provenientes de estudos de câncer, por exemplo, câncer de mama, leucemia, câncer da próstata e melanoma (CHEN; HARRINGTON; IBRAHIM, 2002).

Entretanto, esses modelos têm sido abordados na literatura em outros contextos, por exemplo, Yamaguchi (1992) aplicou o modelo com fração de cura para o tempo do emprego permanente no Japão. Chen, Ibrahim e Sinha (1999) estudaram os modelos de fração de cura sob o enfoque bayesiano. Hoggart e Griffin (2001) analisaram o tempo para um cliente mudar de banco.

Em Maller e Zhou (1996), Ibrahim, Chen e Sinha (2001) e Tsodikov, Ibrahim e Yakovlev (2003) podem ser encontrados uma boa revisão sobre os modelos de longa-duração. A seguir, dois modelos de longa-duração mais utilizados na literatura são apresentados.

i. *Modelo de mistura*

Berkson e Gage (1952) introduziram o modelo de fração de cura em forma de mistura, na qual considera uma função de sobrevivência imprópria para a população total (curados e não curados) em conjunto com uma função de sobrevivência própria para a parte da população composta por pacientes não curados. O modelo de mistura considera que a amostra censurada pertence a dois grupos, os indivíduos susceptíveis ao evento de interesse e os indivíduos não-susceptíveis (FAREWELL, 1982).

Em geral, assume-se que a proporção de indivíduos não-susceptíveis ao evento de interesse é denotado por π_0 . Portanto, a função de sobrevivência imprópria, apresentada por Berkson e Gage (1952), é dada por

$$S_{pop}(t) = \pi_0 + (1 - \pi_0)S(t),$$

em que $S(t)$ é a função de sobrevivência própria.

ii. Modelo tempo-promoção

Alternativamente, Yakovlev e Tsodikov (1996) propuseram uma nova classe de modelos com fração de cura utilizando uma estrutura de riscos competitivos. O modelo de longa-duração proposto por esses autores assume que o número de causas competindo para a ocorrência do evento de interesse é uma variável aleatória não observável, denotada por M , que segue uma distribuição de probabilidade discreta. Além disso, considera-se $R_l (l = 1, \dots, M)$ como sendo variáveis aleatórias, representando o tempo até a ocorrência do evento de interesse devido à causa l . Essas variáveis também são independentes e identicamente distribuídas com função de distribuição acumulada $F(\cdot)$ e independentes de M . Cancho et al. (2011) sugerem que para a inclusão de indivíduos não suscetíveis ao evento de interesse na análise, o tempo para a ocorrência do evento seja definido como $T = \min(R_l, 0 \leq l \leq M)$, em que $P(R_0 = \infty) = 1$. Dessa forma, é possível estimar a proporção de curados, denotada por p_0 , da população que não está sujeita à ocorrência do evento de interesse.

Nessas condições, Rodrigues et al. (2009) consideraram que dada uma função de sobrevivência $S(t)$, a função de sobrevivência populacional da variável aleatória T pode ser escrita como

$$S_{pop}(t) = A_p[S(t)] = \sum_{m=0}^{\infty} P(M = m)[S(t)]^m,$$

em que $A_p(\cdot)$ é a função geradora de probabilidades e $0 \leq S(t) \leq 1$.

Neste trabalho, o modelo de tempo-promoção foram considerados, pois levam em consideração a informação que a ocorrência do evento de interesse sofre a influência de outras causas para a sua ocorrência e esse número de causas foram considerados como sendo Poisson e geométrica. Os modelos são descritos a seguir.

5.1 Especificação dos modelos tempo-promoção

5.1.1 Modelo de tempo-promoção Poisson

Se a variável aleatória M segue uma distribuição Poisson com média k , então, função de sobrevivência populacional apresentada por Chen, Ibrahim e Sinha (1999) é dada por

$$S_{pop}(t) = \exp[-k F(t)], \quad (5.1)$$

em que $F(t)$ é a função de distribuição acumulada.

Rodrigues et al. (2009) e Cancho et al. (2011) mostraram que a proporção de indivíduos curados ou a fração de cura de uma população pode ser estimada, calculando-se o

limite de $t \rightarrow \infty$ da função de sobrevivência populacional. Assim, a estimativa da proporção de curados obtida por meio do modelo de tempo-promoção Poisson é dada por

$$p_0 = \lim_{t \rightarrow \infty} S_{pop}(t) = \exp(-k). \quad (5.2)$$

Observa-se que a fração de cura $p_0 = \exp(-k)$ tende a zero quando k tende a infinito e tende a um quando k tende a zero, o que é intuitivamente esperado, uma vez que se k cresce, quando M aumenta, diminui-se a probabilidade de cura.

Então, para a função de sobrevivência populacional dada em (5.1), considerando as distribuições gama-Weibull gama-log-logística, tem-se:

i. Distribuição Poisson gama-Weibull

Assumindo que $F(t)$ é a função de distribuição acumulada dada em (2.5), a função de sobrevivência populacional é expressa por

$$S_{pop}(t) = \exp \left\{ -\frac{k}{\Gamma(\phi)} \gamma \left[\left(\frac{t}{\lambda} \right)^\alpha, \phi \right] \right\}, \quad (5.3)$$

em que $\phi, \lambda, \alpha, k > 0$ são os parâmetros desconhecidos. Assim, a função densidade de probabilidade populacional é dada por

$$f_{pop}(t) = -\frac{\partial[S_{pop}(t)]}{\partial t} = \frac{k\alpha}{\lambda^\alpha \Gamma(\phi)} t^{\alpha\phi-1} \exp \left\{ -\left(\frac{t}{\lambda} \right)^\alpha - \frac{k}{\Gamma(\phi)} \gamma \left[\left(\frac{t}{\lambda} \right)^\alpha, \phi \right] \right\}, \quad (5.4)$$

em que $t > 0$ e as expressões definidas em (5.3) e (5.4) são funções do modelo denominado de Poisson gama-Weibull.

ii. Distribuição Poisson gama-log-logística

Assumindo que $F(t)$ é a função de distribuição acumulada dada em (2.8), a função de sobrevivência populacional é expressa por

$$S_{pop}(t) = \exp \left\{ -\frac{k}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{t}{\lambda} \right)^\alpha \right], \phi \right\} \right\}, \quad (5.5)$$

em que $\phi, \lambda, \alpha, k > 0$ são os parâmetros desconhecidos. Consequentemente, a função densidade de probabilidade populacional é

$$f_{pop}(t) = -\frac{\partial[S_{pop}(t)]}{\partial t} = \frac{k\alpha}{\lambda^\alpha \Gamma(\phi)} t^{\alpha-1} \exp[-kF(t)] \left[1 + \left(\frac{t}{\lambda} \right)^\alpha \right]^{-2} \times \left\{ \log \left[1 + \left(\frac{t}{\lambda} \right)^\alpha \right] \right\}^{\phi-1}, \quad t > 0, \quad (5.6)$$

em que $F(t)$ é a função de distribuição acumulada da gama-log-logística e as expressões definidas em (5.5) e (5.6) são funções do modelo denominado de Poisson gama-log-logística.

5.1.2 Modelo de tempo-promoção geométrica

Se a variável aleatória M segue uma distribuição geométrica com probabilidade de sucesso p , então, função de sobrevivência populacional definida por Cancho, Louzada-Neto e Barriga (2012) é dada por

$$S_{pop}(t) = \frac{p}{p + (1 - p)F(t)}, \quad (5.7)$$

em que $F(t)$ é a função de distribuição acumulada.

A estimativa da proporção de curados obtida por meio do modelo tempo-promoção é dada da seguinte forma

$$p_0 = \lim_{t \rightarrow \infty} S_{pop}(t) = p.$$

Então, a função de sobrevivência populacional dada em (5.7) é expressa de duas formas:

i. Distribuição geométrica gama-Weibull

Assumindo que $F(t)$ é a função de distribuição acumulada dada em (2.5), a função de sobrevivência populacional é expressa por

$$S_{pop}(t) = \frac{p\Gamma(\phi)}{p\Gamma(\phi) + (1 - p)\gamma\left[\left(\frac{t}{\lambda}\right)^\alpha, \phi\right]}, \quad (5.8)$$

em que $\phi, \lambda, \alpha, k > 0$ são os parâmetros desconhecidos. Dessa forma, a função densidade de probabilidade é dada por

$$f_{pop}(t) = -\frac{\partial[S_{pop}(t)]}{\partial t} = \frac{p(1 - p)\alpha}{\lambda^\alpha \Gamma(\phi)} t^{\alpha\phi-1} \exp\left[-\left(\frac{t}{\lambda}\right)^\alpha\right] [p + (1 - p)F(t)]^{-2}, \quad (5.9)$$

em que $t > 0$ e $F(t)$ é a função de distribuição acumulada da distribuição gama generalizada e as expressões definidas em (5.8) e (5.9) são funções do modelo denominado de geométrica gama-Weibull.

ii. Distribuição geométrica gama-log-logística

Assumindo que $F(t)$ é a função de distribuição acumulada dada em (2.8), a função de sobrevivência populacional é expressa por

$$S_{pop}(t) = \frac{p\Gamma(\phi)}{p\Gamma(\phi) + (1 - p)\gamma\left\{\log\left[1 + \left(\frac{t}{\lambda}\right)^\alpha\right], \phi\right\}}, \quad (5.10)$$

em que $\phi, \lambda, \alpha, k > 0$ são os parâmetros desconhecidos. Assim, a função de probabilidade condicional é denotada por

$$f_{pop}(t) = -\frac{\partial[S_{pop}(t)]}{\partial t} = \frac{p(1 - p)\alpha}{\lambda^\alpha \Gamma(\phi)} t^{\alpha-1} \left[1 + \left(\frac{t}{\lambda}\right)^\alpha\right]^{-2} \times \left\{\log\left[1 + \left(\frac{t}{\lambda}\right)^\alpha\right]\right\}^{\phi-1} [p + (1 - p)F(t)]^{-2}, t > 0, \quad (5.11)$$

em que $F(t)$ é a função de distribuição acumulada da distribuição gama log-logística e as expressões definidas em (5.10) e (5.11) são funções do modelo denominado de geométrica gama-log-logística.

5.2 Estimação de parâmetros por máxima verossimilhança

Para definição do modelo de regressão no contexto de modelos de tempo-promoção, considera-se que os tempos de vida $t_1, \dots, t_n$ são variáveis aleatórias independentes e que cada $t_i (i = 1, \dots, n)$ segue uma distribuição de probabilidade definida na Seção 5.1.1 ou na Seção 5.1.2. Então, para associar covariáveis ao modelo, Chen, Ibrahim e Sinha (1999) assumem que:

i. Modelo de tempo-promoção Poisson

Assumindo um modelo de tempo-promoção Poisson, uma função de ligação logarítmica do parâmetro $k_i (i = 1, \dots, n)$ pode ser igualada a um preditor linear η_i , tal que

$$\log(k_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad i = 1, \dots, n, \quad (5.12)$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})^\top$ representa os valores das covariáveis correspondentes à i -ésima observação.

ii. Modelo de tempo-promoção geométrica

Para o modelo de tempo-promoção geométrica, uma função de ligação probito do parâmetro $p_i (i = 1, \dots, n)$ pode ser igualada a um preditor linear η_i , tal que

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad i = 1, \dots, n, \quad (5.13)$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})^\top$ representa os valores das covariáveis correspondentes à i -ésima observação.

Dessa forma, para definir a função de verossimilhança para o modelo de tempo-promoção, considera-se que os valores da amostra sejam constituídos por $(t_1, \delta_1, \mathbf{x}_1), \dots, (t_n, \delta_n, \mathbf{x}_n)$, em que t_i é a variável resposta, $\delta_i = I(T_i \leq C_i)$ é o indicador de censura e $\mathbf{x}_i$ é o vetor de covariáveis associados a i -ésima observação, para $i = 1, \dots, n$. Assim, de acordo com de Castro, Cancho e Rodrigues (2009) a função de verossimilhança, assumindo que a censura é não-informativa, pode ser representada da seguinte forma

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n [f_{\text{pop}}(t_i)]^{\delta_i} [S_{\text{pop}}(t_i)]^{1-\delta_i}, \quad (5.14)$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos, δ_i é o indicador de censura, $S_{\text{pop}}(t_i)$ é a função de sobrevivência populacional e $f_{\text{pop}}(t_i)$ é a função densidade de probabilidade populacional.

Portanto, o logaritmo da função de verossimilhança do modelo de tempo-promoção Poisson e do modelo de tempo-promoção geométrica podem ser escritos das seguintes formas:

i. Poisson gama-Weibull

Substituindo as funções definidas em (5.4) e (5.3) na expressão (5.14), o logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha\phi \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha\phi - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^n \delta_i \left(\frac{t_i}{\lambda} \right)^\alpha - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \gamma \left[\left(\frac{t_i}{\lambda} \right)^\alpha, \phi \right], \quad (5.15)$$

em que $\boldsymbol{\theta} = (\phi, \lambda, \alpha, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, r é o número que observações que não foram censuradas e $\gamma(x, \phi)$ é a função gama incompleta.

ii. Poisson gama-log-logística

Substituindo as funções definidas em (5.6) e (5.5) na expressão (5.14), o logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\} - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right], \phi \right\}, \quad (5.16)$$

em que $\boldsymbol{\theta} = (\phi, \lambda, \alpha, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, r é o número que observações que não foram censuradas e $\gamma(x, \phi)$ é a função gama incompleta.

iii. Geométrica gama-Weibull

Substituindo as funções definidas em (5.9) e (5.8) na expressão (5.14), o logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha\phi \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha\phi - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^n \delta_i \left(\frac{t_i}{\lambda} \right)^\alpha - \sum_{i=1}^n (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left[\left(\frac{t_i}{\lambda} \right)^\alpha, \phi \right] \right\}, \quad (5.17)$$

em que $\boldsymbol{\theta} = (\phi, \lambda, \alpha, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, r é o número que observações que não foram censuradas e $\gamma(x, \phi)$ é a função gama incompleta.

iv. Geométrica gama-log-logística

Substituindo as funções definidas em (5.11) e (5.10) na expressão (5.14), o logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\} - \sum_{i=1}^n (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right], \phi \right\} \right\}, \quad (5.18)$$

em que $\boldsymbol{\theta} = (\phi, \lambda, \alpha, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, r é o número que observações que não foram censuradas e $\gamma(x, \phi)$ é a função gama incompleta.

Portanto, para maximizar os logaritmos das funções de verossimilhança (5.15) - (5.18) e assim, encontrar as estimativas dos parâmetros dos modelos de tempo-promoção, foi utilizada a sub-rotina MaxBFGS da linguagem matricial de programação Ox versão 6.20 (DOORNIK, 2007), como descrito na Seção 3.2.

5.3 Análise de sensibilidade

Para verificar a presença de possíveis pontos influentes, os métodos de análise de sensibilidade são ferramentas usuais para detectar esses pontos, que podem causar anomalias do modelo ajustado.

Para encontrar os pontos influentes, neste trabalho, considerou-se as medidas de influência global, distância de Cook definida em (3.17) e afastamento da verossimilhança definida em (3.18).

Por outro lado, as medidas de influência local também são calculadas, porém, nesse caso, um esquema de perturbação no modelo ou nos dados deve ser considerado. Dessa forma, os seguintes esquemas de perturbação foram utilizados no modelo de regressão de longa-duração:

a. Perturbação de casos

Considerando $0 \leq \omega_i \leq 1$ e $\boldsymbol{\omega}_0 = (1, \dots, 1)^\top$ é o vetor de não perturbação. O logaritmo da função de verossimilhança perturbada incorpora diferentes pesos para cada elemento do conjunto de dados e é definido por:

i. Poisson gama-Weibull

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = \{r \log(\alpha) - r\alpha \log(\lambda) - r \log[\Gamma(\phi)]\} \sum_{i=1}^n \omega_i + \sum_{i=1}^n \omega_i \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} + (\alpha\phi - 1) \times \sum_{i=1}^n \omega_i \delta_i \log(t_i) - \sum_{i=1}^n \omega_i \delta_i \left(\frac{t_i}{\lambda} \right)^\alpha - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \omega_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \gamma \left[\left(\frac{t_i}{\lambda} \right)^\alpha, \phi \right].$$

ii. Poisson gama-log-logística

$$l(\boldsymbol{\theta}) = \{r \log(\alpha) - r\alpha \log(\lambda) - r \log[\Gamma(\phi)]\} \sum_{i=1}^n \omega_i + (\alpha - 1) \sum_{i=1}^n \omega_i \delta_i \log(t_i) + \\ \sum_{i=1}^n \omega_i \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \sum_{i=1}^n \omega_i \delta_i \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \omega_i \delta_i \times \\ \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\} - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \omega_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right], \phi \right\}.$$

iii. Geométrica gama-Weibull

$$l(\boldsymbol{\theta}) = \{r \log(\alpha) - r\alpha\phi \log(\lambda) - r \log[\Gamma(\phi)]\} \sum_{i=1}^n \omega_i + (\alpha\phi - 1) \sum_{i=1}^n \omega_i \delta_i \log(t_i) + \\ \sum_{i=1}^n \omega_i \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^n \omega_i \delta_i \left(\frac{t_i}{\lambda} \right)^\alpha - \sum_{i=1}^n \omega_i (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + \right. \\ \left. \frac{1}{\Gamma(\phi)} \gamma \left[\left(\frac{t_i}{\lambda} \right)^\alpha, \phi \right] \right\}.$$

iv. Geométrica gama-log-logística

$$l(\boldsymbol{\theta}) = \{r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda)\} \sum_{i=1}^n \omega_i + (\alpha - 1) \sum_{i=1}^n \omega_i \delta_i \log(t_i) + \\ \sum_{i=1}^n \omega_i \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \sum_{i=1}^n \omega_i \delta_i \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \omega_i \delta_i \log \left\{ \log \left[1 + \right. \right. \\ \left. \left. \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\} - \sum_{i=1}^n \omega_i (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right], \phi \right\} \right\}.$$

b. ***Perturbação da variável resposta***

Para verificar a sensibilidade do modelo (5.12) ou do modelo (5.13), considera-se que a variável resposta $t_i (i = 1, \dots, n)$ é submetida a um esquema de perturbação aditiva por meio da adição de um vetor, denotado por $\boldsymbol{\omega}$, de pequenas perturbações, tal que $t_i^* = t_i + \omega_i \text{sd}(t_i)$, em que $\text{sd}(t_i)$ é um fator escala que pode ser a estimativa do desvio padrão de t_i e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Nesse caso, o logaritmo da função de verossimilhança perturbada, assumindo que $\boldsymbol{\omega}_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação, é expresso das seguintes formas:

i. Poisson gama-Weibull

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = r \log(\alpha) - r\alpha\phi \log(\lambda) - r \log [\Gamma(\phi)] + \sum_{i=1}^n \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} + (\alpha\phi - 1) \sum_{i=1}^n \delta_i \log(t_i^*) - \sum_{i=1}^n \delta_i \left(\frac{t_i^*}{\lambda} \right)^\alpha - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \gamma \left[\left(\frac{t_i^*}{\lambda} \right)^\alpha, \phi \right].$$

ii. Poisson gama-log-logística

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i^*) + \sum_{i=1}^n \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i^*}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i^*}{\lambda} \right)^\alpha \right] \right\} - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \gamma \left\{ \log \left[1 + \left(\frac{t_i^*}{\lambda} \right)^\alpha \right], \phi \right\}.$$

iii. Geométrica gama-Weibull

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha\phi \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha\phi - 1) \sum_{i=1}^n \delta_i \log(t_i^*) + \sum_{i=1}^n \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^n \delta_i \left(\frac{t_i^*}{\lambda} \right)^\alpha - \sum_{i=1}^n (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left[\left(\frac{t_i^*}{\lambda} \right)^\alpha, \phi \right] \right\}.$$

iv. Geométrica gama-log-logística

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i^*) + \sum_{i=1}^n \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i^*}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i^*}{\lambda} \right)^\alpha \right] \right\} - \sum_{i=1}^n (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{t_i^*}{\lambda} \right)^\alpha \right], \phi \right\} \right\}.$$

c. Perturbação de uma covariável

Uma outra forma de avaliar a sensibilidade do modelo (5.12) ou do modelo (5.13) é considerar pequenas perturbações em uma particular covariável contínua, denotada por X_j . Nesse caso, a covariável também é submetida a um esquema de perturbação aditiva, de tal forma que $x_{ij}^* = x_{ij} + \omega_i \text{sd}(x_{ij})$, em que $\text{sd}(x_{ij})$ é um fator de escala que pode ser a estimativa do desvio padrão da covariável perturbada e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA;

CORDEIRO, 2010). Dessa forma, considerando $\mathbf{x}_i^{*\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j [x_{ij} + \omega_i \text{sd}(x_{ij})] + \dots + \beta_p x_{ip}$ e $\boldsymbol{\omega}_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação. O logaritmo da função de verossimilhança perturbada é dada por:

i. Poisson gama-Weibull

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = r \log(\alpha) - r\alpha \log(\lambda) - r \log [\Gamma(\phi)] + \sum_{i=1}^n \delta_i \mathbf{x}_i^{*\top} \boldsymbol{\beta} + (\alpha\phi - 1) \times \\ \sum_{i=1}^n \delta_i \log(t_i) - \sum_{i=1}^n \delta_i \left(\frac{t_i}{\lambda}\right)^\alpha - \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta}) \gamma \left[\left(\frac{t_i}{\lambda}\right)^\alpha, \phi \right].$$

ii. Poisson gama-log-logística

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \delta_i \mathbf{x}_i^{*\top} \boldsymbol{\beta} - \\ 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i}{\lambda}\right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda}\right)^\alpha \right] \right\} - \\ \frac{1}{\Gamma(\phi)} \sum_{i=1}^n \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta}) \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda}\right)^\alpha \right], \phi \right\}.$$

iii. Geométrica gama-Weibull

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r\alpha \log(\lambda) - r \log [\Gamma(\phi)] + (\alpha\phi - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \mathbf{x}_i^{*\top} \boldsymbol{\beta} - \\ \sum_{i=1}^n \delta_i \left(\frac{t_i}{\lambda}\right)^\alpha - \sum_{i=1}^n (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left[\left(\frac{t_i}{\lambda}\right)^\alpha, \phi \right] \right\}.$$

iv. Geométrica gama-log-logística

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda) + (\alpha - 1) \sum_{i=1}^n \delta_i \log(t_i) + \sum_{i=1}^n \mathbf{x}_i^{*\top} \boldsymbol{\beta} - \\ 2 \sum_{i=1}^n \delta_i \log \left[1 + \left(\frac{t_i}{\lambda}\right)^\alpha \right] + (\phi - 1) \sum_{i=1}^n \delta_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda}\right)^\alpha \right] \right\} - \\ \sum_{i=1}^n (\delta_i + 1) \log \left\{ \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta}) + \frac{1}{\Gamma(\phi)} \gamma \left\{ \log \left[1 + \left(\frac{t_i}{\lambda}\right)^\alpha \right], \phi \right\} \right\}.$$

Para esses três esquemas de perturbação calculou-se, numericamente, a matriz de curvaturas máximas definida por

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{vi})_{(p+3) \times n} = \left[\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \theta_v \partial \omega_i} \right]_{(p+3) \times n},$$

em que $v = 1, \dots, p + 3$, $i = 1, \dots, n$ e $\omega = (\omega_1, \dots, \omega_n)^\top$ é o vetor de perturbação, isto é, os pesos que perturbam o modelo ou as observações.

5.4 Aplicação

Para ilustrar a metodologia proposta, foi utilizado um conjunto de dados proveniente do estudo piloto de Grana, Chinol e Robertson (2002), na qual eles tinham o interesse em para avaliar o efeito da radioimunoterapia (RIT) no tratamento de pacientes com glioma. A RIT é um tratamento que permite aos médicos direcionar a radiação diretamente para as células cancerígenas do paciente de dentro do corpo ao invés de fontes externas.

Os paciente com glioma foram coletados no período de dezembro de 1994 a agosto de 1997 e estão disponíveis em Everitt e Hothorn (2010). Assim, para um total de 37 pacientes foi observado para cada um o tempo (em meses) a partir do tratamento até a morte do paciente causada pela doença.

Dessa forma, as seguintes variáveis foram observadas na análise:

- i. y : Logaritmo do tempo de vida (em meses);
- ii. x_1 : Idade (em anos);
- iii. x_2 : Sexo (0= feminino, 1=masculino);
- iv. x_3 : Tratamento (0= controle, 1=RIT).

Para estudar o comportamento dos dados, curvas de sobrevivência estimadas pelo método de Kaplan-Meier e o gráfico TTT-plot foram construídos. Esses gráficos são representados pelas Figuras 5.1-5.2.

A partir da Figura 5.1b, verifica-se que a função de taxa de falha dos tempos de vida dos pacientes com glioma se caracteriza pela forma unimodal segundo a interpretação de Aarset (1987), pois o gráfico TTT-plot apresenta a curva côncava e, então, convexa. Observa-se, também, que a curva de sobrevivência estimada pelo método de Kaplan-Meier apresenta um patamar em torno de 0,35; indicando a existência de uma proporção de curados, como ilustra a Figura 5.1a.

As curvas de sobrevivência estimadas pelo método de Kaplan-Meier estratificado para a covariável x_2 (Figura 5.2a) e para a covariável x_3 (Figura 5.2b) indicam que o tratamento de radioimunoterapia influencia no tempo de vida de pacientes com glioma e por outro lado, o tempo de sobrevivência de pacientes do sexo feminino é igual ao tempo de sobrevivência de pacientes do sexo masculino.

De acordo com o comportamento dos tempos de vida dos pacientes com glioma observado na Figuras 5.1-5.2, considerou-se para esta análise, os modelos de tempo-promoção propostos na Seção 5.1.

Figura 5.1 – TTT-plot (a) e estimativas de Kaplan-Meier da função de sobrevivência (b) para os dados de glioma

Figura 5.2 – Curva de sobrevivência estimada pelo método de Kaplan-Meier para a covariável x_3 (a) e para a covariável x_2 (b) dos dados de glioma

Dessa forma, para maximizar as funções de verossimilhança (5.15-5.18) e assim, encontrar as estimativas de máxima verossimilhança dos parâmetros, dos seus respectivos modelos, foi utilizada a sub-rotina `MaxBGFs` da linguagem matricial de programação `Ox` versão 6.20 com os valores iniciais $\phi = 1,000$; $\lambda = 1,000$; $\sigma = 1,000$; $\beta_0 = 6,014$; $\beta_1 = -0,034$; $\beta_2 = 0,006$ e $\beta_3 = -1,441$ obtidos a partir do ajuste do modelo de regressão Weibull no pacote estatístico R versão 2.15.1.

Os resultados da estimação de parâmetros modelos de tempo-promoção Poisson e de tempo-promoção geométrica, bem como os erros padrões e as significâncias dos parâmetros encontram-se na Tabela 5.1.

Tabela 5.1 – Estimativa de máxima verossimilhança dos parâmetros dos modelos de longa-duração ajustados aos dados de glioma

Parâmetro	Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)
Poisson gama-Weibull				
ϕ	21,4000	3,7820	–	(13,9875;28,8126)
λ	0,0100	0,0018	–	(0,0065; 0,0135)
α	0,3800	0,0321	–	(0,3171; 0,4429)
β_0	-4,2513	1,0174	0,0000	(-6,2454;-2,2572)
β_1	0,0782	0,0196	0,0001	(0,0399; 0,1166)
β_2	-0,3133	0,4667	0,5021	(-1,2280; 0,6014)
β_3	2,6600	0,5545	0,0000	(1,5733; 3,7467)
Poisson gama-log-logística				
ϕ	10,7919	20,4658	–	(0,0000;50,9041)
λ	1,6892	3,6815	–	(0,0000; 8,9048)
α	2,9825	4,8333	–	(0,0000;12,4556)
β_0	-3,5216	1,2792	0,0059	(-6,0287;-1,0145)
β_1	0,0758	0,0193	0,0001	(0,0379; 0,1136)
β_2	-0,3719	0,4743	0,4329	(-1,3015; 0,5577)
β_3	2,5877	0,5525	0,0000	(1,5047; 3,6707)
Geométrica gama-Weibull				
ϕ	24,9343	4,4308	–	(16,2501;33,6185)
λ	0,0100	0,0019	–	(0,0062; 0,0137)
α	0,3941	0,0325	–	(0,3303; 0,4578)
β_0	4,0792	1,4277	0,0043	(1,2811; 6,8774)
β_1	-0,0892	0,0267	0,0008	(-0,1415;-0,0369)
β_2	0,8273	0,6638	0,2126	(-0,4736; 2,1283)
β_3	-3,3812	0,7919	0,0000	(-4,9332;-1,8291)
Geométrica gama-log-logística				
ϕ	12,8248	27,6383	–	(0,0000; 66,9949)
λ	1,7332	4,0818	–	(0,0000; 9,7334)
α	3,2908	6,6660	–	(0,0000;16,3559)
β_0	3,2231	2,0993	0,1247	(-0,8914;7,3375)
β_1	-0,0892	0,0268	0,0009	(-0,1417;-0,0367)
β_2	0,8854	0,6670	0,1844	(-0,4219; 2,1926)
β_3	-3,3089	0,7864	0,0000	(-4,8503;-1,7675)

Por meio dos resultados da Tabela 5.1, observa-se que a um nível de significância de 5%, que a covariável x_2 não foi significativa para explicar a variabilidade presente no logaritmo

do tempo de vida dos pacientes com glioma.

Para dar continuidade na análise, o próximo passo foi escolher o modelo mais adequado entre os quatro modelos propostos, utilizando os critérios de seleção de modelos apresentado na Seção 3.3.

Desse modo, os resultados das estatísticas BIC e AIC_{SUR} são apresentados na Tabela 5.2, enquanto que os resultados dos testes de hipóteses são apresentados na Tabela 5.3. Por meio desses resultados, observa-se que o modelo mais adequado para analisar os dados de tempo de vida de pacientes com glioma é o modelo Poisson gama-Weibull, pois apresenta os menores valores das estatísticas AIC_{SUR} e além disso, os testes de hipóteses indicam que os modelos Poisson gama-Weibull e geométrica gama-Weibull são equivalentes.

Tabela 5.2 – Estatísticas BIC e AIC_{SUR} dos modelos de longa-duração

Modelo	BIC	AIC_{SUR}
Poisson gama-Weibull (PGW)	206,3566	197,8802
Poisson gama-log-logística (PGLL)	205,4992	197,0228
Geométrica gama-Weibull (GGW)	208,3718	199,8954
Geométrica gama-log-logística (GGLL)	207,4012	198,9248

Tabela 5.3 – Teste de hipóteses para modelos não encaixados com H_0 : Modelos equivalentes

	Hipóteses	Estatística T	p -valor
PGW <i>versus</i> PGLL	H_1 : PGW	-11,6433	<0,0001
GGW <i>versus</i> GGLL	H_1 : GGW	-0,0002	0,5000
PGW <i>versus</i> GGW	H_1 : PGW	0,0268	0,4893

Após ter definido o modelo e estimado os seus parâmetros, foi realizado uma análise de sensibilidade para detectar possíveis pontos influentes no modelo Poisson gama-Weibull. As medidas de influência global e local foram calculadas utilizando a linguagem matricial de programação Ox versão 6.20.

A distância de Cook generalizada (3.17) e afastamento da verossimilhança (3.18) são apresentadas na Figura 5.3. A partir dessa figura, verifica-se que os pontos #1, #6, #19 são possíveis observações influentes. Nas Figuras 5.4-5.5, têm-se os gráficos de influência local, considerando as perturbações de caso ($C_{d_{max}} = 1,5346$), perturbação do tempo de vida ($C_{d_{max}} = 0,0554$) e perturbação da covariável x_1 ($C_{d_{max}} = 3,8194$). A partir desses gráficos, é possível identificar os pontos #1, #14, #18, #24, #28, #31 e #35 como possíveis observações influentes.

Figura 5.3 – Gráficos de medidas de influência global para o modelo Poisson gama-Weibull ajustado. (a) Distância de Cook generalizada e (b) Afastamento da verossimilhança

Figura 5.4 – Gráficos de d_{max} do modelo Poisson gama-Weibull ajustado. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_1

Figura 5.5 – Gráficos de C_i do modelo Poisson gama-Weibull ajustado. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_1

Considerando os pontos que apareceram com maior frequência nos gráficos, a análise de sensibilidade (influência global e influência local) detectou como possíveis pontos influentes as observações #1, #24, #28, #31 e #35. O passo seguinte foi verificar o impacto dessas observações nas estimativas dos parâmetros. Reajustou-se o modelo, eliminando-se, individualmente, cada observação, em seguida, eliminando as duas observações.

Na Tabela 5.4, são apresentadas as mudanças relativas (em porcentagens) de cada parâmetro estimado, definido por $\mathbf{RC}_{\theta_j} = \left[(\hat{\theta}_j - \hat{\theta}_{j(i)}) / \hat{\theta}_j \right] \times 100$, em que $\theta_{j(i)}$ é o estimador de máxima verossimilhança sem a i -ésima observação.

Tabela 5.4 – Mudanças relativas [-**RC**-em %], estimativas dos parâmetros e os correspondentes p -valores entre parênteses, excluindo as observações #1, #24, #28, #31 e #35
(continua)

Exclusão	$\hat{\phi}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
A	-	-	-	-	-	-	-
	21,4000	0,0100	0,3800	-4,2513	0,0782	-0,3133	2,6600
	(-)	(-)	(-)	(0,0000)	(0,0001)	(0,5021)	(0,0000)
A_1	[-10]	[0]	[-6]	[-6]	[0]	[9]	[0]
	23,5012	0,0100	0,4012	-4,4930	0,0775	-0,2841	2,6665
	(-)	(-)	(-)	(0,0000)	(0,0001)	(0,5573)	(0,0000)
A_2	[0]	[0]	[2]	[9]	[2]	[-77]	[-12]
	21,4290	0,0100	0,3742	-3,8877	0,0764	-0,5559	2,9917
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,2170)	(0,0000)
A_3	[2]	[0]	[1]	[-5]	[-5]	[20]	[-3]
	20,9031	0,0100	0,3760	-4,4828	0,0822	-0,2493	2,7414
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,5896)	(0,0000)
A_4	[-3]	[0]	[-1]	[5]	[8]	[50]	[2]
	21,9536	0,0100	0,3846	-4,0436	0,0717	-0,1558	2,6114
	(-)	(-)	(-)	(0,0001)	(0,0004)	(0,7492)	(0,0000)
A_5	[-1]	[0]	[0]	[-5]	[-5]	[29]	[-5]
	21,5653	0,0100	0,3815	-4,5410	0,0824	-0,2228	2,7854
	(-)	(-)	(-)	(0,0000)	(0,0000)	(0,6302)	(0,0000)
A_6	[-9]	[0]	[-4]	[2]	[4]	[-64]	[-10]
	23,3893	0,0100	0,3961	-4,1538	0,0752	-0,5150	2,9253
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,2697)	(0,0000)
A_7	[-7]	[0]	[-5]	[-12]	[-5]	[31]	[-4]
	22,9422	0,0100	0,3973	-4,7819	0,0823	-0,2169	2,7609
	(-)	(-)	(-)	(0,0000)	(0,0001)	(0,6508)	(0,0000)
A_8	[-14]	[0]	[-7]	[-1]	[9]	[68]	[1]
	24,5466	0,0100	0,4082	-4,2883	0,0708	-0,1017	2,6244
	(-)	(-)	(-)	(0,0001)	(0,0004)	(0,8410)	(0,0000)
A_9	[-11]	[0]	[-6]	[-14]	[-6]	[41]	[-6]
	23,8173	0,0100	0,4033	-4,8553	0,0829	-0,1842	2,8218
	(-)	(-)	(-)	(0,0000)	(0,0000)	(0,7008)	(0,0000)
A_{10}	[2]	[0]	[3]	[4]	[-2]	[-57]	[-15]
	20,9319	0,0100	0,3702	-4,0663	0,0796	-0,4904	3,0493
	(-)	(-)	(-)	(0,0002)	(0,0001)	(0,2742)	(0,0000)

Tabela 5.4 – Mudanças relativas [-RC-em %], estimativas dos parâmetros e os correspondentes p -valores entre parênteses, excluindo as observações #1, #24, #28, #31 e #35 (conclusão)

A_{11}	[-3]	[0]	[0]	[13]	[9]	[-36]	[-10]
	21,9490	0,0100	0,3791	-3,7186	0,0708	-0,4267	2,9266
	(-)	(-)	(-)	(0,0003)	(0,0003)	(0,3629)	(0,0000)
A_{12}	[-1]	[0]	[1]	[2]	[-3]	[-48]	[-17]
	21,6153	0,0100	0,3760	-4,1815	0,0807	-0,4632	3,1028
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,3005)	(0,0000)
A_{13}	[1]	[0]	[0]	[1]	[4]	[60]	[0]
	21,1274	0,0100	0,3784	-4,2131	0,0751	-0,1249	2,6701
	(-)	(-)	(-)	(0,0002)	(0,0005)	(0,7956)	(0,0000)
A_{14}	[-2]	[0]	[-1]	[-25]	[-22]	[77]	[-16]
	21,9130	0,0100	0,3828	-5,3268	0,0954	-0,0711	3,0787
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,8779)	(0,0000)
A_{15}	[-2]	[0]	[-1]	[-2]	[3]	[71]	[-2]
	21,8703	0,0100	0,3842	-4,3196	0,07611	-0,0924	2,7263
	(-)	(-)	(-)	(0,0001)	(0,0003)	(0,8480)	(0,0000)
A_{16}	[-7]	[0]	[-3]	[-1]	[-1]	[-42]	[-13]
	22,8351	0,0100	0,3923	-4,2888	0,0791	-0,4454	2,9952
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,3379)	(0,0000)
A_{17}	[-14]	[0]	[-6]	[6]	[11]	[-16]	[-8]
	24,3952	0,0100	0,4035	-3,9772	0,0693	-0,3634	2,8642
	(-)	(-)	(-)	(0,0001)	(0,0004)	(0,4563)	(0,0000)
A_{18}	[-11]	[0]	[-4]	[-6]	[-3]	[-32]	[-15]
	23,7083	0,0100	0,3985	-4,5129	0,0806	-0,4125	3,0625
	(-)	(-)	(-)	(0,0000)	(0,0001)	(0,3734)	(0,0000)
A_{19}	[1]	[0]	[2]	[10]	[6]	[-24]	[-12]
	21,1442	0,0100	0,3728	-3,8393	0,0733	-0,3893	2,9692
	(-)	(-)	(-)	(0,0005)	(0,0005)	(0,4030)	(0,0000)
A_{20}	[-2]	[0]	[1]	[-14]	[-18]	[-1]	[-25]
	21,8348	0,0100	0,3766	-4,8484	0,0919	-0,3160	3,3358
	(-)	(-)	(-)	(0,0003)	(0,0001)	(0,4843)	(0,0000)
A_{21}	[-1]	[0]	[0]	[-17]	[-13]	[99]	[-12]
	21,6060	0,0100	0,3812	-4,9861	0,0880	-0,0022	2,9716
	(-)	(-)	(-)	(0,0003)	(0,0004)	(0,9964)	(0,0000)
A_{22}	[-10]	[0]	[-5]	[3]	[8]	[-4]	[-10]
	23,4910	0,0100	0,3976	-4,1432	0,0723	-0,3245	2,9143
	(-)	(-)	(-)	(0,0002)	(0,0006)	(0,5031)	(0,0000)
A_{23}	[-13]	[0]	[-5]	[-27]	[-21]	[24]	[-26]
	24,1010	0,0100	0,3998	-5,3906	0,0950	-0,2382	3,3598
	(-)	(-)	(-)	(0,0001)	(0,0001)	(0,6119)	(0,0000)
A_{24}	[-13]	[0]	[-6]	[5]	[-12]	[45]	[-22]
	24,1584	0,0100	0,4011	-4,0555	0,0875	-0,1720	3,2459
	(-)	(-)	(-)	(0,0003)	(0,0005)	(0,7217)	(0,0000)

Os resultados da Tabela 5.4, mostraram que a significância dos parâmetros não foi alterada, considerando um nível de 5% de significância e logo, não houve mudanças inferenciais

após a remoção das observações consideradas influentes nos gráficos de influência global e local. Essas observações que foram apontadas como possíveis pontos influentes, são pacientes que tem as seguintes descrições:

- i. A observação 1 corresponde ao paciente do sexo feminino com tempo de vida mediano e com uma idade jovem, pertencente ao grupo de controle.
- ii. A observação 24 corresponde ao paciente do sexo feminino curado depois de 50 meses e com idade jovem, pertencente ao grupo de tratamento.
- iii. A observação 28 corresponde ao paciente do sexo masculino com pouco tempo de vida e com idade avançada, pertencente ao grupo de tratamento.
- iv. A observação 31 corresponde ao paciente do sexo feminino com menor tempo de vida e com uma idade avançada, pertencente ao grupo de tratamento.
- v. A observação 35 corresponde ao paciente do sexo masculino com tempo de vida mediano e com uma idade mediana, pertencente ao grupo de tratamento.

Diante da descrição dos pacientes em tratamento de glioma, observa-se que os pacientes não apresentam nenhum padrão incomum, por isso, as observações foram mantidas no conjunto de dados.

Portanto, o modelo Poisson gama-Weibull final é dado por um preditor linear que é composto pela informação de uma constante e das covariáveis x_1 e x_3 , pois ao nível de 5% de significância, a covariável x_2 foi não significativa.

Assim, novamente utilizando a linguagem matricial de programação OX versão 6.20, as estimativas dos parâmetros do modelo final foram calculadas. Na Tabela 5.5, são apresentadas as estimativas dos parâmetros do modelo Poisson gama-Weibull final, bem como os erros padrões e a significância dos parâmetros.

Tabela 5.5 – Estimativas de máxima verossimilhança do modelo Poisson gama-Weibull final ajustado aos dados de glioma

Parâmetro	Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)
ϕ	20,9310	3,7178	–	(13,6443;28,2177)
λ	0,0100	0,0018	–	(0,0064; 0,0136)
α	0,3759	0,0327	–	(0,3118; 0,4401)
β_0	-4,2356	1,0136	0,0000	(-6,2221;-2,2490)
β_1	0,0742	0,0183	0,0001	(0,0383; 0,1101)
β_3	2,6754	0,5591	0,0000	(1,5796; 3,7712)

A partir do modelo final, é possível estimar a proporção de curados, utilizando a expressão definida em (5.2) e considerando a propriedade de invariância dos estimadores de

máxima verossimilhança. Assim, tem-se que

$$\hat{p}_{0i} = \exp(-\exp(-4,2356 + 0,0742\text{idade}_i + 2,6754\text{tratamento}_i)), \quad \text{e} \quad \hat{p}_0 = \frac{\sum_{i=1}^{37} \hat{p}_{0i}}{37} = 0,3328.$$

Portanto, por meio do modelo é possível obter as seguintes informações

- i. Aproximadamente 33% dos pacientes diagnosticados com glioma foram curados quando receberam o tratamento com radioimunoterapia.
- ii. A cada aumento em uma unidade da covariável idade, a proporção de pacientes curados fica aumentado em aproximadamente 8% ($\exp(0,0742)$).
- iii. Existem indicativos que indivíduos tratados com radioimunoterapia aumentam em 15 vezes a proporção de curados em relação aos indivíduos do grupo controle.

A Figura 5.6 apresenta a curva de sobrevivência estimada pelo método de Kaplan-Meier *versus* a curva de sobrevivência estimada pelo modelo de longa-duração Poisson gama-Weibull na ausência de covariáveis. Observa-se pela figura que a estimativa da fração de curados é um pouco elevada quando as covariáveis são inseridas no modelo. Além disso, o modelo mostra um ajuste razoável aos dados de tempo de vida de pacientes com glioma.

Figura 5.6 – Curvas de sobrevivência estimadas pelo modelo Poisson gama-Weibull *versus* a curva de sobrevivência estimada pelo método de Kaplan-Meier para os dados de glioma

6 MODELO DE REGRESSÃO COM EFEITO ALEATÓRIO

Em algumas aplicações, às vezes, não é possível incluir todos os fatores de risco relevantes no modelo de sobrevivência, devido a falta de conhecimento sobre quais fatores são importantes ou a falta dos dados necessários. Esse fato pode ocasionar uma heterogeneidade que deve ser considerada na interpretação dos resultados, pois pode implicar em resultados não esperados. Por exemplo, considerando estudos de incidência de uma doença, pode ocorrer a pré-disposição de alguns indivíduos para uma doença específica.

Em uma outra situação prática, para um mesmo indivíduo, pode-se observar um único evento várias vezes (medidas repetidas ou eventos recorrentes) e isso faz com que os dados apresentem correlação. Além disso, quando os indivíduos compartilham características comuns, observadas dentro de um mesmo grupo, os tempos de sobrevivência podem estar correlacionados. Por exemplo, dados observados em pessoas da mesma família ou em animais da mesma ninhada.

Nos casos citados, os dados apresentam uma estrutura de grupos, portanto, os modelos tradicionais para análise de sobrevivência não podem ser utilizados, pois não levam em consideração essa informação. Pode-se, então, supor que todos os indivíduos dentro de um grupo apresentem respostas correlacionadas, porque estas são dependentes de variáveis causais. Por exemplo, a correlação pode ser explicada por omissões de variáveis explicativas importantes, tais como fatores genéticos ou ambientais.

Portanto, a inclusão dessas variáveis como variáveis explicativas no modelo de regressão poderia proporcionar independência (condicional) entre as respostas do mesmo grupo. Assim, um modo de representar essa correlação dos dados é incluir ao modelo uma variável latente para cada indivíduo. Essa variável não é observada e é denominada de efeito aleatório. Dessa forma, espera-se que um modelo que considera as variáveis aleatórias não observadas como um fator aleatório desconhecido, produza resultados mais consistentes do que o modelo que não considera essa correlação entre as respostas.

Então, ao utilizar um modelo com efeito aleatório para realizar uma análise estatística, é necessário atribuir uma distribuição para os efeitos aleatórios para estimar os parâmetros do modelo. Em geral, no contexto de análise de sobrevivência, para a distribuição do efeito aleatório tem-se adotado a distribuição gama, devido a interpretação que pode ser feita sobre o parâmetro da distribuição (COLOSIMO; GIOLO, 2006). Mas, também, é possível encontrar trabalhos na qual a distribuição do efeito aleatório é a distribuição normal, por exemplo, Hedeker, Siddiqui e Hu (2000) apresentaram um modelo de sobrevivência para dados agrupados com efeito aleatório e Vaida e Xu (2000) apresentaram um modelo de riscos proporcionais com efeito aleatório.

Por outro lado, outros trabalhos têm focado o estudo na escolha da distribuição do efeito aleatório, como por exemplo, Verbeke e Lesaffre (1997), Lee e Thompson (2008), Zhang

e Davidian (2001) e McCulloch e Neuhaus (2011). Esses trabalhos discutem a má especificação da distribuição do efeito aleatório e suas principais consequências. Tais estudos, apontam que a estimativa da variância da distribuição do efeito aleatório pode ser afetada pela má especificação do modelo.

Por essa razão, outras distribuições têm sido atribuídas aos efeitos aleatórios, como alternativa à distribuição normal e a distribuição gama. Como pode ser observado no estudo de Fabio, Paula e de Castro (2012) que apresentaram um modelo Poisson com efeito aleatório que segue uma distribuição log-gama-generalizada e Cabral, Lachos e Madruga (2012) que propuseram um modelo com efeito aleatório que segue uma distribuição skew-normal.

Assim, seguindo a mesma idéia desses estudos, este capítulo será voltado para apresentar uma distribuição alternativa para os efeitos aleatórios.

6.1 Especificação do modelo de regressão com efeito aleatório log-gama-G

Para definir um modelo de sobrevivência com efeito aleatório, admite-se que os tempos de sobrevivência estão sujeitos a censura à direita, na qual o mecanismo de censura é aleatório e uma amostra dividida em m grupos. Então, assumindo que T_{ji} é o tempo de sobrevivência do indivíduo i no grupo j , com $i = 1, \dots, n_j$ e $j = 1, \dots, N$ e C_{ji} é uma variável aleatória, independente de T , que representa o tempo de censura associado a esse indivíduo, os dados observados são denotados por $t_{ji} = \min(T_{ji}, C_{ji})$ e

$$\delta_{ji} = \begin{cases} 1 & \text{se } T_{ji} \leq C_{ji} \\ 0 & \text{se } T_{ji} > C_{ji} \end{cases}$$

em que δ_{ji} é o indicador de censura do i -ésimo indivíduo no j -ésimo grupo.

Por outro lado, sabendo que os tempos de sobrevivência podem estar relacionados a outras variáveis, conhecidas como covariáveis e assumindo que todos os indivíduos do mesmo grupo têm em comum o efeito aleatório, denotado por B_j . Então, supondo que os efeitos aleatórios são variáveis aleatórias não observáveis, o modelo de regressão nessa situação é expresso da seguinte forma

$$Y_{ji} = \mathbf{x}_{ji}^\top \boldsymbol{\beta} + B_j + \sigma Z_{ji}, \quad (6.1)$$

em que $Y_{ji} = \log(T_{ji})$ é o logaritmo do tempo de sobrevivência do i -ésimo indivíduo associado ao j -ésimo grupo, $\mathbf{x}_{ji}$ é um vetor $(p \times 1)$ e Z_{ji} 's são os erros aleatórios independentes e identicamente distribuídos. O vetor $\boldsymbol{\beta}$ e σ são parâmetros desconhecidos e para cada grupo tem-se um efeito aleatório B_j , representado por variáveis aleatórias independentes e identicamente distribuídas com função densidade denotada por $g(\cdot)$.

Além disso, assume-se que os efeitos aleatórios são independentes dos tempos de censura e que $\text{Cov}(B_j, Z_{ji}) = 0$, na qual condicionado ao efeito aleatório B_j , as respostas dentro do grupo j são independentes.

Dessa forma, supondo que a variável aleatório Y segue uma distribuição de valor extremo, então, o modelo definido em (6.1) é um modelo de regressão Weibull com efeito aleatório. Assim, para esse modelo, as seguintes suposições são feitas:

- i. $y_{ji}|b_j \sim \text{valor extremo}(\sigma, \mu_{ji})$, com função densidade de probabilidade marginal e função de sobrevivência dadas, respectivamente, por

$$f(y_{ji}|b_j) = \frac{1}{\sigma} \exp \left[\left(\frac{y_{ji} - \mu_{ji}}{\sigma} \right) - \exp \left(\frac{y_{ji} - \mu_{ji}}{\sigma} \right) \right] \quad e \quad (6.2)$$

$$S(y_{ji}|b_j) = \exp \left[- \exp \left(\frac{y_{ji} - \mu_{ji}}{\sigma} \right) \right]. \quad (6.3)$$

- ii. $\mu_{ji} = \mathbf{x}_{ji}^\top \boldsymbol{\beta} + b_j$, em que $\mathbf{x}_{ji}^\top = (x_{ji1}, \dots, x_{jip})^\top$ é o vetor de covariáveis, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos associados as covariáveis e b_j é o efeito aleatório associado ao j -ésimo grupo.

- iii. Têm-se duas alternativas para a distribuição do efeito aleatório:

- I. $b_j \sim \text{log-gama-Weibull}(\phi, \sigma_{lgw}, 0)$, com função densidade de probabilidade definida por

$$g(b_j) = \frac{1}{\sigma \Gamma(\phi)} \exp \left[\frac{\phi(b_j - \mu)}{\sigma} - \exp \left(\frac{b_j - \mu}{\sigma} \right) \right], \quad -\infty < b_j < \infty,$$

em que $\phi > 0$, $\sigma_{lgw} > 0$ e $\mu = 0$.

- II. $b_j \sim \text{log-gama-logística}(\phi, \sigma_{lgl}, 0)$, com função densidade de probabilidade definida por

$$g(b_j) = \frac{1}{\sigma \Gamma(\phi)} \exp \left(\frac{b_j - \mu}{\sigma} \right) \left[1 + \exp \left(\frac{b_j - \mu}{\sigma} \right) \right]^{-2} \left\{ \log \left[1 + \exp \left(\frac{b_j - \mu}{\sigma} \right) \right] \right\}^{\phi-1},$$

em que $-\infty < b_j < \infty$, $\phi > 0$, $\sigma_{lgl} > 0$ e $\mu = 0$.

Portanto, se para o modelo em (6.1) supõe que o efeito aleatório segue uma distribuição log-gama-Weibull, então, tem-se um modelo de regressão Weibull com efeito aleatório log-gama-Weibull. Por outro lado, se o efeito aleatório segue uma distribuição log-gama-logística, tem-se um modelo de regressão Weibull com efeito aleatório log-gama-logística. Além disso, qualquer um desses modelos pode ser utilizado para analisar um conjunto de dados, na qual é verificada a existência de um efeito aleatório.

6.2 Estimação por máxima verossimilhança

Para encontrar as estimativas dos parâmetros do modelo de regressão Weibull com efeito aleatório foi utilizado o método da máxima verossimilhança. Nesse caso, o método consiste em maximizar o logaritmo função de verossimilhança marginal, obtida por meio da integração da função de verossimilhança em relação aos efeitos aleatórios b_j .

Assim, sabendo que as observações são censuradas, a função de verossimilhança, descrita por Villegas, Paula e Leiva (2011), dentro de cada grupo é representada por

$$L_j(y_{ji}|b_j) = \prod_{i=1}^{n_j} [f(y_{ji}|b_j)^{\delta_i} S(y_{ji}|b_j)^{1-\delta_i}], \quad (6.4)$$

em que $f(y_{ji}|b_j)$ é a função densidade de probabilidade definida em (6.2) e $S(y_{ji}|b_j)$ é a função de sobrevivência definida em (6.3).

Dessa forma, supondo que os B_j e os Y_{ij} são variáveis aleatórias independentes, a contribuição do j -ésimo grupo para a função de verossimilhança marginal é

$$\int L_j(y_{ij}|b_j) g(b_j) db_j,$$

em que $g(\cdot)$ é a função densidade de probabilidade do efeito aleatório e $L_j(y_{ji}|b_j)$ é definida em (6.4). Portanto, sob a suposição de independência entre os grupos, a função de verossimilhança marginal apresentada por Villegas, Paula e Leiva (2011) é dada por

$$L(\theta) = \prod_{j=1}^N \int \prod_{i=1}^{n_j} [f(y_{ij}|b_j)^{\delta_i} S(y_{ij}|b_j)^{1-\delta_i}] g(b_j) db_j, \quad (6.5)$$

Então, para um conjunto de dados que consiste de n observações, $(y_{11}, \delta_{11}, \mathbf{x}_{11}), \dots, (y_{1n_j}, \delta_{1n_j}, \mathbf{x}_{1n_j}), \dots, (y_{N1}, \delta_{N1}, \mathbf{x}_{N1}), \dots, (y_{Nn_j}, \delta_{Nn_j}, \mathbf{x}_{Nn_j})$, em que $n = n_1 + \dots + n_j$, $y_{ji} = \log(t_{ji})$, δ_{ji} é o indicador de censura e $\mathbf{x}_{ji}$ é o vetor de covariáveis associados a i -ésima observação do j -ésimo grupo, o logaritmo da função de verossimilhança marginal em (6.5), supondo diferentes distribuições para o efeito aleatório, é expresso da seguinte forma:

i. Efeito aleatório log-gama-Weibull

Assumindo que o efeito aleatório segue uma distribuição log-gama-Weibull.

$$l(\theta) = \frac{1}{\sigma\Gamma(\phi)} \sum_{j=1}^N \int_{-\infty}^{+\infty} \prod_{i=1}^{n_j} [f(y_{ji}|b_j)^{\delta_i} S(y_{ji}|b_j)^{1-\delta_i}] \exp \left[\frac{\phi b_j}{\sigma_{lgw}} - \exp \left(\frac{b_j}{\sigma_{lgw}} \right) \right] db_j \quad (6.6)$$

em que $\theta = (\phi, \sigma, \sigma_{lgw}, \beta^\top)^\top$ é o vetor de parâmetros desconhecidos, $f(y_{ji}|b_j)$ é a função densidade de probabilidade definida em (6.2) e $S(y_{ji}|b_j)$ é a função de sobrevivência definida em (6.3).

ii. Efeito aleatório log-gama-logística

Assumindo que o efeito aleatório segue uma distribuição log-gama-logística.

$$l(\theta) = \frac{1}{\sigma\Gamma(\phi)} \sum_{j=1}^N \int_{-\infty}^{+\infty} \prod_{i=1}^{n_j} [f(y_{ji}|b_j)^{\delta_i} S(y_{ji}|b_j)^{1-\delta_i}] \exp \left(\frac{b_j}{\sigma_{lgl}} \right) \left[1 + \exp \left(\frac{b_j}{\sigma_{lgl}} \right) \right]^{-2} \times \left\{ \log \left[1 + \exp \left(\frac{b_j}{\sigma_{lgl}} \right) \right] \right\}^{\phi-1} db_j, \quad (6.7)$$

em que $\boldsymbol{\theta} = (\phi, \sigma, \sigma_{lgw}, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos, $f(y_{ji}|b_j)$ é a função densidade de probabilidade definida em (6.2) e $S(y_{ji}|b_j)$ é a função de sobrevivência definida em (6.3).

O problema em maximizar a função de verossimilhança definida em (6.6) e em (6.7) é a presença de N integrais sobre os efeitos aleatórios b_j e por essa razão, em muitas situações, essas integrais não têm solução analítica (BONAT et al. 2012).

Dessa forma, para estimar os parâmetros dos modelos com efeito aleatório é necessário a utilização de métodos de integração numérica ou métodos de estimação que possam de alguma forma resolver as integrais envolvidas, como por exemplo, métodos bayesianos e métodos que utilizam o algoritmo EM (VAIDA; XU, 2000).

Para este trabalho, considerou-se o método integração numérica de Gauss-Hermite, assim como é utilizada por Villegas, Paula e Leiva (2011) e Bonat et al. (2012). Esse método é uma extensão do método de Quadratura Gaussiana para resolver integrais da forma

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x) dx.$$

Neste caso, essa integral é aproximada por uma soma ponderada da função avaliada nos pontos de Gauss e pesos de integração, ou seja,

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x) dx \approx \sum_{v=1}^m w_v f(x_v),$$

em que m é o número de pontos utilizados para a aproximação. Os valores dos x_v são as raízes do polinômio de Hermite $H_m(x)$ ($v = 1, \dots, m$) e os pesos w_v associados são dados por

$$w_v = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 [H_{m-1}(x_v)]^2}. \quad (6.8)$$

Para a aproximação de integrais via o método de Gauss-Hermite é necessário obter os pesos de integração w_v e dos pontos de Gauss x_v . A função `gauss.quad()` do pacote `statmod` do pacote estatística R calcula os pesos e os pontos de Gauss-Hermite (BONAT et al, 2012).

Assim, utilizando o método de integração numérica de Gauss-Hermite, o logaritmo da função de verossimilhança da expressão (6.6) e (6.7) podem ser reescritas da seguinte forma:

i. Efeito aleatório log-gama-Weibull

$$l(\boldsymbol{\theta}) = \sum_{j=1}^N \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp \left[\frac{\phi x_v}{\sigma_{lgw}} - \exp \left(\frac{x_v}{\sigma_{lgw}} \right) \right] \times \prod_{i=1}^{n_j} \left[f(y_{ji}|x_v)^{\delta_i} S(y_{ji}|x_v)^{1-\delta_i} \right] \right\}, \quad (6.9)$$

em que w_v são os pesos definido em (6.8), $f(y_{ji}|x_v)$ é a função densidade de probabilidade definida em (6.2) e $S(y_{ji}|x_v)$ é a função de sobrevivência definida em (6.3).

ii. Efeito aleatório log-gama-logística

$$l(\boldsymbol{\theta}) = \sum_{j=1}^N \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp\left(\frac{x_v}{\sigma_{lgl}}\right) \left[1 + \exp\left(\frac{x_v}{\sigma_{lgl}}\right)\right]^{-2} \times \right. \\ \left. \left\{ \log \left[1 + \exp\left(\frac{x_v}{\sigma_{lgl}}\right)\right] \right\}^{\phi-1} \prod_{i=1}^{n_j} \left[f(y_{ji}|x_v)^{\delta_i} S(y_{ji}|x_v)^{1-\delta_i} \right] \right\}, \quad (6.10)$$

em que w_v são os pesos definido em (6.8), $f(y_{ji}|x_v)$ é a função densidade de probabilidade definida em (6.2) e $S(y_{ji}|x_v)$ é a função de sobrevivência definida em (6.3).

Portanto, utilizando o algoritmo de *Quase-Newton*, implementado na linguagem matricial de programação Ox versão 6.10, é possível maximizar as funções de verossimilhanças marginais (6.9) e (6.10); e assim, obter as estimativas dos parâmetros dos modelos.

Os intervalos de confiança e os testes de hipóteses são determinados utilizando a distribuição assintótica dos estimadores de máxima verossimilhança, na qual é uma distribuição normal com a matriz de covariância obtida como a inversa da matriz de informação de Fisher sob certas condições de regularidade. Mais especificamente, a matriz de covariância assintótica é dada por $\mathbf{I}^{-1}(\boldsymbol{\theta})$ em que $\mathbf{I}(\boldsymbol{\theta}) = E[\ddot{\mathbf{L}}(\boldsymbol{\theta})]$ tal que $\ddot{\mathbf{L}}(\boldsymbol{\theta}) = -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]$.

Como não é possível determinar a matriz de informação de Fisher $\mathbf{I}(\boldsymbol{\theta})$ devido a presença de observações censuradas, uma forma alternativa é utilizar a matriz de informação observada $\ddot{\mathbf{L}}(\boldsymbol{\theta})$, avaliada no estimador de máxima verossimilhança $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. A aproximação normal assintótica de $\hat{\boldsymbol{\theta}}$ é expressa como

$$\hat{\boldsymbol{\theta}}^\top \sim N_{(p+3)}[\boldsymbol{\theta}^\top; \ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1}],$$

em que $\ddot{\mathbf{L}}(\boldsymbol{\theta})$ é a matriz de informação observada $(p+3)(p+3)$.

Assim, o intervalo de confiança assintótico para β_j , com $j = 1, \dots, p$, considerando um nível de confiança de $(1 - \alpha)100\%$, é expresso por

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)},$$

em que $z_{\alpha/2}$ é o valor que excede com probabilidade $\alpha/2$ a distribuição normal padrão.

Para prever o efeito aleatório, Villegas, Paula e Leiva (2011) e Fabio, Paula e de Castro (2012) utilizaram o estimador de Bayes, denotado por

$$\tilde{b}_j = E(b_j | \mathbf{y}_j; \boldsymbol{\theta}) = \frac{\int_{-\infty}^{\infty} b_j f(\mathbf{y}_j | b_j; \boldsymbol{\beta}) f(b_j) db_j}{\int_{-\infty}^{\infty} f(\mathbf{y}_j | b_j; \boldsymbol{\beta}) f(b_j) db_j}.$$

6.3 Análise de sensibilidade

Da mesma forma que nos modelos de regressão sem a adição do efeito aleatório, a próxima etapa é considerar uma análise de sensibilidade para verificar se existem observações influentes que podem causar importantes distorções nos resultados da análise.

As técnicas de diagnóstico de modelos com efeito aleatório, no contexto de análise de sobrevivência, são recentes e por isso são poucos os trabalhos encontrados na literatura. Dentre, os trabalhos encontrados, têm-se Xiang et al. (2007) e Villegas, Paula e Leiva (2011) que apresentam medidas de influência global e local para dados censurados.

6.3.1 Influência global

Sabendo que a amostra observada é formada por $(Y_{jl}, \delta_{jl}, \mathbf{x}_{jl})$, em que Y_{jl} é o logaritmo da variável resposta, δ_{jl} é o indicador de censura e $\mathbf{x}_{jl}$ é o vetor de covariáveis para a l -ésima observação do j -ésimo grupo. Então, a eliminação de um indivíduo (ou caso) em modelo de regressão é especificada por

$$Y_{jl} = \mathbf{x}_{jl}^\top \boldsymbol{\beta} + b_j + \sigma Z_{jl}, \quad l = 1, \dots, n_j, \quad l \neq i, \quad (6.11)$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos, σ é o parâmetro escala e Z_{jl} é uma variável aleatória.

Agora, considere que uma quantidade com subscrito i , significa a quantidade original com o i -ésimo caso eliminado. Assim, para o modelo (6.11), o logaritmo da função de verossimilhança de $\boldsymbol{\theta}$ excluindo o i -ésimo caso é denotado por $l_{(i)}(\boldsymbol{\theta})$, em que $\hat{\boldsymbol{\theta}}_{(i)}$ é o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ calculada maximizando-se $l_{(i)}(\boldsymbol{\theta})$. Para avaliar a influência do i -ésimo caso no estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$, a idéia é utilizar a diferença entre $\hat{\boldsymbol{\theta}}_{(i)}$ e $\hat{\boldsymbol{\theta}}$. Então, se $\hat{\boldsymbol{\theta}}_{(i)}$ está distante de $\hat{\boldsymbol{\theta}}$ o i -ésimo caso pode ser considerado como uma observação influente.

Uma medida de influência global, considerada por Xie e Wei (2007), é uma generalização da distância de Cook. Essa medida é definida como uma norma padronizada de $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ e expressa por

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^\top [\ddot{\mathbf{L}}(\boldsymbol{\theta})] (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \quad (6.12)$$

em que $\ddot{\mathbf{L}}(\boldsymbol{\theta})$ é a matriz de informação observada.

Outra medida para avaliar a influência de um caso é apresentada por Cook e Weisberg (1982). Denominada de afastamento da verossimilhança, a medida considera a diferença entre $\hat{\boldsymbol{\theta}}_{(i)}$ e $\hat{\boldsymbol{\theta}}$ e é definida por

$$LD_i(\boldsymbol{\theta}) = 2 \left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)}) \right], \quad (6.13)$$

em que $l(\hat{\boldsymbol{\theta}})$ é o valor do logaritmo da função de verossimilhança da amostra completa e $l(\hat{\boldsymbol{\theta}}_{(i)})$ é o valor do logaritmo da função de verossimilhança da amostra excluindo a i -ésima observação.

6.3.2 Influência local

Utilizando a mesma abordagem adotada nos capítulos anteriores, as medidas de influência local são calculadas assumindo um esquema de perturbação no modelo ou nos dados.

Considerando o modelo definido em (6.1) e o logaritmo da função de verossimilhança expresso em (6.9) e em (6.10), os seguintes esquemas de perturbação foram utilizados no modelo de regressão Weibull com efeito aleatório:

a. ***Perturbação de grupos***

O logaritmo da função de verossimilhança perturbada, considerando $0 \leq \omega_j \leq 1$ e $\omega_0 = (1, \dots, 1)^\top$ é o vetor de não perturbação, incorpora diferentes pesos para cada grupo do conjunto de dados e é definido por:

i. *Log-gama-Weibull*

$$l(\theta) = \sum_{j=1}^N \omega_j \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp \left[\frac{\phi x_v}{\sigma_{lgw}} - \exp \left(\frac{x_v}{\sigma_{lgw}} \right) \right] \times \prod_{i=1}^{n_j} \left[f(y_{ji}|x_v)^{\delta_i} S(y_{ji}|x_v)^{1-\delta_i} \right] \right\},$$

ii. *Log-gama-logística*

$$l(\theta) = \sum_{j=1}^N \omega_j \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp \left(\frac{x_v}{\sigma_{lgl}} \right) \left[1 + \exp \left(\frac{x_v}{\sigma_{lgl}} \right) \right]^{-2} \times \left\{ \log \left[1 + \exp \left(\frac{x_v}{\sigma_{lgl}} \right) \right] \right\}^{\phi-1} \prod_{i=1}^{n_j} \left[f(y_{ji}|x_v)^{\delta_i} S(y_{ji}|x_v)^{1-\delta_i} \right] \right\},$$

b. ***Perturbação da variável resposta***

Para verificar a sensibilidade do modelo (6.1), considera-se que a variável resposta $y_{ji} = \log(t_{ji})$ é submetida a um esquema de perturbação aditiva por meio da adição de um vetor, denotado por ω , de pequenas perturbações, tal que $y_{ji}^* = y_{ji} + \omega_{ji} \text{sd}(y_{ji})$, em que $\text{sd}(y_{ji})$ é um fator escala que pode ser a estimativa do desvio padrão do logaritmo do tempos de falha e $\omega_{ji} \in \mathbf{R}$ (Silva et al., 2010). Nesse caso, para $\omega_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação, o logaritmo da função de verossimilhança perturbada é expresso por:

i. *Log-gama-Weibull*

$$l(\theta) = \sum_{j=1}^N \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp \left[\frac{\phi x_v}{\sigma_{lgw}} - \exp \left(\frac{x_v}{\sigma_{lgw}} \right) \right] \times \prod_{i=1}^{n_j} \left[f(y_{ji}^*|x_v)^{\delta_i} S(y_{ji}^*|x_v)^{1-\delta_i} \right] \right\},$$

ii. Log-gama-logística

$$l(\boldsymbol{\theta}) = \sum_{j=1}^N \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp\left(\frac{x_v}{\sigma_{lgl}}\right) \left[1 + \exp\left(\frac{x_v}{\sigma_{lgl}}\right)\right]^{-2} \times \right. \\ \left. \left\{ \log \left[1 + \exp\left(\frac{x_v}{\sigma_{lgl}}\right)\right] \right\}^{\phi-1} \prod_{i=1}^{n_j} \left[f(y_{ji}^* | x_v)^{\delta_i} S(y_{ji}^* | x_v)^{1-\delta_i} \right] \right\},$$

c. **Perturbação de uma covariável**

Uma outra forma de avaliar a sensibilidade do modelo (6.1) é considerar pequenas perturbações em uma particular covariável contínua, denotada por X_l . Nesse caso, a covariável também é submetida a um esquema de perturbação aditiva, de tal forma que $x_{jil}^* = x_{jil} + \omega_i \text{sd}(x_{jil})$, em que $\text{sd}(x_{jil})$ é um fator de escala que pode ser a estimativa do desvio padrão da covariável perturbada e $\omega_{ji} \in \mathbf{R}$. Dessa forma, considerando que $\mathbf{x}_i^{*\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{ji1} + \dots + \beta_l [x_{jil} + \omega_{ji} \text{sd}(x_{jil})] + \dots + \beta_p x_{jip}$ e $\boldsymbol{\omega}_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação., o logaritmo da função de verossimilhança perturbada é dada por:

i. Log-gama-Weibull

$$l(\boldsymbol{\theta}) = \sum_{j=1}^N \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp \left[\frac{\phi x_v}{\sigma_{lgw}} - \exp\left(\frac{x_v}{\sigma_{lgw}}\right) \right] \times \right. \\ \left. \prod_{i=1}^{n_j} \left[f^*(y_{ji} | x_v)^{\delta_i} S^*(y_{ji} | x_v)^{1-\delta_i} \right] \right\},$$

ii. Log-gama-logística

$$l(\boldsymbol{\theta}) = \sum_{j=1}^N \log \left\{ \frac{1}{\sigma \Gamma(\phi)} \sum_{v=1}^m w_v \exp(x_v^2) \exp\left(\frac{x_v}{\sigma_{lgl}}\right) \left[1 + \exp\left(\frac{x_v}{\sigma_{lgl}}\right)\right]^{-2} \times \right. \\ \left. \left\{ \log \left[1 + \exp\left(\frac{x_v}{\sigma_{lgl}}\right)\right] \right\}^{\phi-1} \prod_{i=1}^{n_j} \left[f^*(y_{ji} | x_v)^{\delta_i} S^*(y_{ji} | x_v)^{1-\delta_i} \right] \right\},$$

$$\text{em que } f^*(y_{ji} | x_v) = \frac{1}{\sigma} \exp \left[\left(\frac{y_{ji} - \mathbf{x}_{ji}^{*\top} \boldsymbol{\beta} - x_v}{\sigma} \right) \right] - \exp \left(\frac{y_{ji} - \mathbf{x}_{ji}^{*\top} \boldsymbol{\beta} - x_v}{\sigma} \right) \Bigg]$$

$$\text{e } S^*(y_{ji} | x_v) = \exp \left[- \exp \left(\frac{y_{ji} - \mathbf{x}_{ji}^{*\top} \boldsymbol{\beta} - x_v}{\sigma} \right) \right].$$

Para esses três esquemas de perturbação calculou-se, numericamente, a matriz de curvaturas máximas definida por

$$\Delta = (\Delta_{vji})_{(p+3) \times n} = \left[\frac{\partial^2 l(\theta|\omega)}{\partial \theta_v \partial \omega_{ji}} \right]_{(p+3) \times n},$$

em que $n = n_1 + \dots + n_j$, $v = 1, \dots, p+3$ e $\omega = (\omega_1, \dots, \omega_n)^\top$ é o vetor de perturbação, isto é, os pesos que perturbam o modelo de regressão Weibull com efeito aleatório log-gama-Weibull ou com efeito aleatório log-gama-logística ou as observações.

6.4 Aplicação

O conjunto de dados utilizado nessa aplicação foi obtido de Colosimo e Giolo (2006). O banco de dados é composto por quatro touros da raça Nelore que produziram um total de 155 animais. Todos os animais produzidos por esses touros nasceram na primavera entre os anos de 1993 e 1998, sendo 68% fêmeas. O número de animais por touro variou de 31 a 56. O tempo, em dias, que um animal levou para ganhar 160kg no período de nascimento até a desmama foi utilizado como variável resposta. Assim, as seguintes variáveis foram observadas:

- i. y : Logaritmo do tempo de ganho de peso (em dias);
- ii. x_1 : Sexo (0= fêmea, 1= macho);
- iii. x_2 : Idade da vaca no parto (em anos).

Para verificar o comportamento dos dados de tempo de ganho de peso foram construídos as curvas de Kaplan-Meier e a curva TTT, representadas na Figura 6.1.

Figura 6.1 – TTT-plot (a), curva de sobrevivência estimada pelo método de Kaplan-Meier para o tempo de ganho de peso (b) e para a covariável x_2 (c)

A partir desses gráficos, observa-se que o TTT-plot indica que o tempo de ganho de peso dos animais apresenta uma taxa de falha crescente, podendo ser considerada a distribuição

Weibull (Figura 6.1a). Apesar dos tempos de ganho de peso apresentarem um patamar acima de zero, é razoável supor que os animais a longo prazo atingiram o peso de 160kg (Figura 6.1b). Também, observa-se que existe uma evidência de que sexo dos animais influenciam no tempo de ganho de peso, como ilustra a Figura 6.1c.

De acordo com que foi observado na Figura 6.1, pode-se considerar o modelo de regressão Weibull dado por

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + b_j + \sigma z_{ij}, \quad j = 1, \dots, 4, \quad i = 1, \dots, n_j,$$

em que $y = \log(t)$ denota o logaritmo do tempo de ganho de peso, Z é uma variável aleatório com distribuição valor extremo e b_j é o efeito aleatório referentes aos touros com as seguintes distribuições de probabilidade:

i. Log-gama-Weibull

$$g(b_j) = \frac{1}{\sigma \Gamma(\phi)} \exp \left[\left(\frac{b_j - \mu}{\sigma} \right) - \exp \left(\frac{b_j - \mu}{\sigma} \right) \right],$$

em que $j = 1, \dots, 4$ e $\mu = 0$.

ii. Log-gama-logística

$$g(b_j) = \frac{1}{\sigma \Gamma(\phi)} \exp \left(\frac{b_j - \mu}{\sigma} \right) \left[1 + \exp \left(\frac{b_j - \mu}{\sigma} \right) \right]^{-2} \left\{ \log \left[1 + \exp \left(\frac{b_j - \mu}{\sigma} \right) \right] \right\},$$

em que $j = 1, \dots, 4$ e $\mu = 0$.

Para maximizar as funções de verossimilhanças marginais (6.9) e (6.10); e assim encontrar as estimativas de máxima verossimilhança dos modelos, foi utilizada a sub-rotina `MaxBGFS` da linguagem matricial de programação `Opt` versão 6.20 com os valores iniciais $\phi = 1,000$; $\sigma = 0,760$; $\beta_0 = 4,880$; $\beta_1 = -0,153$ e $\beta_2 = 0,026$, obtidos a partir do ajuste do modelo de regressão Weibull no pacote estatístico R versão 2.15.1.

Dessa forma, na Tabela 6.1 encontra-se as estimativas dos parâmetros dos modelos, os erros padrões e as significâncias dos parâmetros. Logo, observa-se que a idade da vaca não exerce influência na variabilidade do tempo de ganho de peso dos animais, ao nível de 5% de significância.

Para o passo seguinte, foi considerado a estatística BIC definida em (2.21) e a estatística AIC_{SUR} definida em (3.14) para escolher o modelo mais adequado. Os valores dessas estatísticas são apresentados na Tabela 6.2. Por meio desses resultados, observa-se que o modelo de regressão Weibull com efeito aleatório log-gama-logística é mais adequado para analisar esse conjunto de dados, pois apresentou o menor valor das estatísticas.

Tabela 6.1 – Estimativa de máxima verossimilhança dos parâmetros dos modelos para os dados de tempos de ganho de peso

Parâmetro	Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)
Efeito aleatório log-gama-Weibull				
ϕ	2,9934	0,9574	–	(1,1170; 4,8698)
σ_w	0,1668	0,0167	–	(0,1340; 0,1995)
σ_{lgw}	0,4003	0,0920	–	(0,2200; 0,5806)
β_0	5,2214	0,0608	0,0000	(5,1023; 5,3406)
β_1	-0,0127	0,0413	0,0000	(-0,2563;-0,0946)
β_2	-0,0117	0,0066	0,0742	(-0,0246; 0,0012)
Efeito aleatório log-gama-logística				
ϕ	4,6569	2,4917	–	(0,0000; 9,5405)
σ_w	0,1668	0,0167	–	(0,1340; 0,1995)
σ_{lgl}	0,0783	0,0444	–	(0,0000; 0,1654)
β_0	5,2214	0,0608	0,0000	(5,1023; 5,3406)
β_1	-0,1755	0,0413	0,0000	(-0,2563; -0,0946)
β_2	-0,0117	0,0066	0,0742	(-0,0246; 0,0012)

Tabela 6.2 – Estatísticas BIC e AIC_{SUR} dos modelos propostos

Modelo	BIC	AIC_{SUR}
Log-gama-Weibull	112,8612	95,0033
log-gama-logística	110,4400	92,5821

Então, para encontrar possíveis pontos influentes no modelo de regressão Weibull com efeito aleatório log-gama-logística para os dados de tempo de ganho de peso foram calculadas as medidas de influência global e local utilizando a linguagem matricial de programação Ox versão 6.20.

A distância de Cook generalizada (6.12) e afastamento da verossimilhança (6.13) são apresentadas na Figura 6.2. A partir dessa figura, verifica-se que os pontos #16, #72 e #142 são possíveis observações influentes.

Nas Figuras 6.3-6.4, têm-se os gráficos de influência local, considerando as perturbações de caso (grupos) ($C_{d_{max}} = 2,2778$), perturbação do logaritmo do tempo de ganho de peso ($C_{d_{max}} = 0,9549$) e perturbação da covariável x_2 ($C_{d_{max}} = 1,2346$). A partir desses gráficos é possível identificar os pontos #5, #16, #127 e #142.

Figura 6.2 – Gráficos de medidas de influência global para o modelo de regressão Weibull com efeito aleatório log-gama-logística ajustado. (a) Distância de Cook generalizada e (b) Afastamento da verossimilhança

Figura 6.3 – Gráficos de influência local d_{max} do modelo de regressão Weibull com efeito aleatório log-gama-logística ajustado. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_2

Figura 6.4 – Gráficos de influência local total C_i do modelo de regressão Weibull com efeito aleatório log-gama-logística ajustado. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_2

Logo, a análise de sensibilidade (influência global e influência local) detectou como possíveis pontos influentes as observações #16 e #142, pois apareceram com maior frequência. Assim, para analisar o impacto dessas observações nas estimativas dos parâmetros, reajustou-se o modelo eliminando-se, individualmente, cada observação, em seguida, eliminando as duas observações.

Assim, na Tabela 6.3, são apresentadas as mudanças relativas (em porcentagens) de cada parâmetro estimado, definido por $\mathbf{RC}_{\theta_j} = \left[(\hat{\theta}_j - \hat{\theta}_{j(i)}) / \hat{\theta}_j \right] \times 100$, em que $\theta_{j(i)}$ é o estimador de máxima verossimilhança sem a i -ésima observação.

Por meio dos resultados da Tabela 6.3, observa-se que as estimativas de máxima verossimilhança os parâmetros do modelo de regressão log-gama-logística são robustas sobre a deleção das observações influentes. Em geral, a significância dos parâmetros estimados não se alterou (ao nível de 5% de significância) após a remoção dos casos. Portanto, não houve mudanças inferenciais, após a remoção das observações consideradas influentes nos gráficos de influência local e global.

Tabela 6.3 – Mudanças relativas [-**RC**-em %], estimativas dos parâmetros e os correspondentes p -valores entre parênteses, excluindo as observações #16 e #142

Exclusão	$\hat{\phi}$	$\hat{\sigma}_w$	$\hat{\sigma}_{lgw}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Nenhuma	-	-	-	-	-	-
#16	4,6569	0,1668	0,0783	5,2214	-0,1755	-0,0117
	(-)	(-)	(-)	(0,0000)	(0,0000)	(0,0742)
	[0]	[1]	[0]	[0]	[-5]	[3]
#142	4,6570	0,1643	0,0783	5,2171	-0,1837	-0,0114
	(-)	(-)	(-)	(0,0000)	(0,0000)	(0,0755)
	[0]	[1]	[0]	[0]	[-5]	[-5]
#16 e #142	4,6570	0,1646	0,0783	5,2252	-0,1846	-0,0123
	(-)	(-)	(-)	(0,0000)	(0,0000)	(0,0540)
	[0]	[3]	[0]	[0]	[-10]	[-3]
	4,6570	0,1619	0,0783	5,2207	-0,1932	-0,0120
	(-)	(-)	(-)	(0,0000)	(0,0000)	(0,0539)

Essas observações que foram apontadas como possíveis pontos influentes, são animais que tem as seguintes descrições:

- A observação 16 corresponde a um animal macho proveniente do touro 1 que não havia atingindo o peso ideal até os 203 dias.
- A observação 142 corresponde a um animal macho proveniente do touro 4 que não havia atingindo o peso ideal até os 199 dias.

Diante da descrição dos animais, observa-se que os animais não apresentam nenhum padrão incomum, por isso, as observações foram mantidas no conjunto de dados.

Portanto, retirando a covariável não significativa, as estimativas do modelo final são apresentadas na Tabela 6.4. De acordo com esses resultados, observa-se que existe uma

diferença significativa no tempo de ganho de peso dos animais em relação ao sexo, ou seja, os animais machos apresentam probabilidade de ganho de peso menor do que as fêmeas.

Uma medida importante, em melhoramento animal, é a herdabilidade (DUCROCQ, 2001) definida por

$$h^2 = \frac{V_{lgl}}{V_{lgl} + \sigma^2(\pi^2/6)} = \frac{0,0298}{0,0298 + [(0,1670)^2](\pi^2/6)} = 0,3942;$$

em que V_{lgl} é a variância da distribuição log-gama-logística.

A herdabilidade de $h^2 = 0,3942$ indica que aproximadamente 39% da variabilidade desse grupo de animais é devido a variabilidade genética. Entretanto, Yazdi, Visscheir e Ducrocq (2002) ressalta que essa estimativa deve ser interpretada com cautela, pois os dados apresentam censuras.

Tabela 6.4 – Estimativa de máxima verossimilhança dos parâmetros do modelo de regressão Weibull com efeito aleatório log-gama-logística final ajustado aos dados de tempos de ganho de peso

Parâmetro	Estimativa	Erro padrão	p-valor	Intervalo de confiança(95%)
ϕ	4,6568	2,4917	–	(0,0000; 9,54052)
σ	0,1670	0,0166	–	(0,1345; 0,1996)
σ_{lgl}	0,0783	0,0444	–	(0,0000; 0,1654)
β_0	5,1335	0,0297	0,0000	(5,0754; 5,1917)
β_1	-0,1711	0,0412	0,0000	(-0,2518; -0,0904)

Utilizando o estimador de Bayes empírico, os efeitos aleatórios preditos são apresentados na Tabela 6.5. Observa-se por meio dos resultados dessa tabela que apenas o touro 2 apresenta um efeito predito diferente. Também, por meio desses resultados é possível construir as curvas estimadas para cada touro estratificada pelo sexo da sua ninhada, como ilustra as Figuras 6.5-6.6. Essas figuras também indicam que o comportamento dos filhotes do touro 2 é diferente dos demais e que esses filhotes têm maior chance de ganhar peso dos que os filhotes dos demais touros.

Tabela 6.5 – Predição dos efeitos aleatórios

Efeito	Predição
b_1 (Touro=1)	0,2453
b_2 (Touro=2)	0,7375
b_3 (Touro=3)	0,2453
b_4 (Touro=4)	0,2453

Figura 6.5 – Curva de sobrevivência estimada para cada Touro. (a) Touro 1 e (b) Touro 2

Figura 6.6 – Curva de sobrevivência estimada para cada Touro. (a) Touro 3 e (b) Touro 4

7 MODELO GAMA-G INFLACIONADO DE ZEROS

As novas distribuições de probabilidade, gama-Weibull e gama-log-logística, obtidas a partir da função geradora de distribuição gama, foram propostas para analisar dados censurados. Entretanto, essas distribuições podem ser utilizadas para descrever fenômenos ou experimentos que produzem resultados nos reais positivos não nulos. Por exemplo, no projeto de pesquisa FAPESP 2011/04968-4 sobre o manejo de árvores de copaibeiras (*Copaifera* sp.), a variável de interesse é proveniente da quantidade de óleo, em litros, produzida por cada árvore. Nesse caso, é razoável supor que a natureza desses dados sejam os reais positivos não nulos.

No entanto, por alguma razão, muitas dessas copaibeiras em condições saudáveis não produziram óleo, ocasionando um excesso de zeros no conjunto de dados. Segundo Ospina e Ferrari (2010), essa proporção de zeros não permite que as observações sejam provenientes de uma distribuição de probabilidade contínua.

Dessa forma, para analisar dados que são caracterizados pela presença de excesso de zeros, em geral, são utilizados os modelos inflacionados. A análise estatística utilizando esses modelos consiste de dados que são distribuídos de forma contínua no suporte da distribuição, mas que incluem observações em um dos seus extremos.

Os modelos inflacionados são muito utilizados para analisar dados de contagem e nessa linha de pesquisa, destacam-se os trabalhos como de Lambert (2001) que apresentou um modelo Poisson inflacionado de zeros, Vieira, Hinde e Demétrio (2000) que introduziram o modelo de regressão binomial inflacionado de zeros, Ridout, Hinde e Demétrio (2001) que apresentaram um teste escore para dois modelos inflacionados de zeros, Famoye e Singh (2006) que utilizaram um modelo Poisson generalizado inflacionado de zeros para analisar dados de violência doméstica e Garay et al. (2011) que apresentaram uma análise de diagnóstico para um modelo binomial-negativa inflacionado de zeros.

Por outro lado, no contexto de distribuições de probabilidade contínuas, Ospina e Ferrari (2010) propuseram um modelo de regressão beta inflacionado de zeros ou de uns. Esse modelo assume que a distribuição dos dados na forma de proporções é uma mistura entre a distribuição beta e uma distribuição de Bernoulli, a qual atribui-se probabilidades aos inteiros que definem o limite do suporte da distribuição beta.

Assim, utilizando a mesma abordagem adotada por Ospina e Ferrari (2010), neste capítulo, será proposto um modelo gama-G inflacionado de zeros para analisar os dados de produção de óleo de copaíba e dessa forma, verificar quais fatores podem estar influenciando no excesso de zeros e se existem observações discrepantes que podem estar interferindo na análise.

7.1 Especificação do modelo inflacionado de zeros

Assumindo que Y é uma variável aleatória inflacionada de zeros, supõe-se que o componente contínuo dos dados é proveniente de uma amostra com distribuição gama-G (2.4 ou 2.7). Enquanto que o componente discreto (ponto de massa), será descrito por meio de uma distribuição degenerada no ponto zero. Assim, de acordo com Ospina e Ferrari (2010), a função de distribuição acumulada obtida por mistura de distribuições é definida por

$$F(y) = \begin{cases} p & \text{se } y = 0, \\ (1 - p)G(y) & \text{se } y > 0, \end{cases} \quad (7.1)$$

em que $0 < p < 1$ e $G(y)$ é uma função de distribuição acumulada. Logo, a função densidade de probabilidade associada a expressão (7.1) é dada por

$$f(y) = \begin{cases} p & \text{se } y = 0, \\ (1 - p)f(y) & \text{se } y > 0, \end{cases}$$

em que $g(y)$ é uma função densidade de probabilidade.

Portanto, assumindo que as observações diferentes de zero são provenientes de uma amostra com distribuição gama-Weibull ou com distribuição gama-log-logística, os modelos inflacionados de zeros são definidos como:

i. Gama-Weibull inflacionado de zeros

Se Y é uma variável aleatória com distribuição gama-Weibull definida em (2.4), então o modelo gama-Weibull inflacionado de zeros é dado por

$$f(y) = \begin{cases} p & \text{se } y = 0, \\ \frac{(1-p)\alpha}{\lambda^{\phi\alpha}\Gamma(\phi)} y^{\phi\alpha-1} \exp\left[-\left(\frac{y}{\lambda}\right)^{\alpha}\right] & \text{se } y > 0, \end{cases} \quad (7.2)$$

em que $\lambda > 0$ é o parâmetro de escala e $\phi, \alpha > 0$ são os parâmetros de forma. Observa-se que a expressão em (7.2) se reduz em um modelo Weibull inflacionado de zeros.

ii. Gama-log-logística inflacionado de zeros

Se Y é uma variável aleatória com distribuição gama-log-logística definida em (2.7), então o modelo gama-log-logística inflacionado de zeros é expresso por

$$f(y) = \begin{cases} p & \text{se } y = 0, \\ \frac{(1-p)\alpha}{\lambda^{\alpha}\Gamma(\phi)} x^{\alpha-1} \left[1 + \left(\frac{y}{\lambda}\right)^{\alpha}\right]^{-2} \left\{\log\left[1 + \left(\frac{y}{\lambda}\right)^{\alpha}\right]\right\}^{\phi-1} & \text{se } y > 0, \end{cases} \quad (7.3)$$

em que $\lambda > 0$ é o parâmetro de escala e $\phi, \alpha > 0$ são os parâmetros de forma. Para $\phi = 1$, a expressão em (7.3) se reduz em um modelo log-logística inflacionado de zeros.

Então, para verificar se um ou mais fatores, denominados de covariáveis, podem estar relacionados com a ocorrência de excessos de zeros, é preciso associar essas covariáveis com a proporção de zeros do modelo. Uma forma de fazer essa associação é considerar uma função de ligação, que une a parte sistemática com a parte aleatória do modelo (PAULA, 2004).

Assim, Ospina e Ferrari (2010) assumem que a proporção de zeros (parâmetro $p_i, i = 1, \dots, n$) é relacionada a um preditor linear, denotado por η_i , por meio de uma função de ligação logito, tal que

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i, \quad i = 1, \dots, n, \quad (7.4)$$

em que n é o tamanho da amostra, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ é o vetor de parâmetros desconhecidos a serem estimados e $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})^\top$ representa os valores das covariáveis correspondentes à i -ésima observação.

7.2 Estimação por máxima verossimilhança

Na literatura, são propostas várias formas de estimar parâmetros de modelos inflacionados de zeros, por exemplo, Ghosh, Mukhopadhyay e Lu (2006) utilizaram uma abordagem bayesiana para estimar os parâmetros do modelo série de potências inflacionado de zeros e Garay et al. (2011) utilizaram o método da máxima verossimilhança e o algoritmo EM para encontrar as estimativas dos parâmetros.

Neste trabalho, o método da máxima verossimilhança é utilizada para estimar os parâmetros dos modelos gama-G inflacionados de zeros. Assim, para um conjunto de dados que consiste de n observações, $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$, em que $\mathbf{x}_i$ é o vetor de covariáveis associados a i -ésima observação, para $i = 1, \dots, n$, a função de verossimilhança, apresentado por Ospina e Ferrari (2012) é dado por

$$L(\boldsymbol{\theta}) = \prod_{i:y_i=0} p_i \prod_{i:y_i>0} [(1-p_i) f(y_i)],$$

em que p_i é definida na expressão (7.4) e $f(y_i)$ é a função densidade de probabilidade da distribuição gama-Weibull (2.4) ou da distribuição gama-log-logística (2.7).

Dessa forma, o logaritmo da função de verossimilhança para cada modelo inflacionado de zeros é expresso por:

i. Modelo gama-Weibull inflacionado de zeros

$$l(\boldsymbol{\theta}) = r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha\phi \log(\lambda) + \sum_{i:y_i=0} \log\left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}\right] - \sum_{i:y_i>0} \log[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] + \alpha\phi \sum_{i:y_i>0} \log(y_i) - \sum_{i:y_i>0} \log(y_i) - \sum_{i:y_i>0} \left(\frac{y_i}{\lambda}\right)^\alpha, \quad (7.5)$$

em que $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos e r é o número que observações diferentes de zero.

A função escore é obtida, calculando-se a derivada logaritmo da função de verossimilhança (7.5) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Dessa forma

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + r\alpha \log(\lambda) + \alpha \sum_{i:y_i>0} \log(y_i), \quad \frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} = \frac{r\alpha\phi}{\lambda} + \frac{\alpha}{\lambda^{\alpha-1}} \sum_{i:y_i>0} y_i^\alpha$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} = \frac{r}{\alpha} - r\phi \log(\lambda) + \phi \sum_{i:y_i>0} \log(y_i) - \sum_{i:y_i>0} \log\left(\frac{y_i}{\lambda}\right) \left(\frac{y_i}{\lambda}\right)^\alpha \quad \mathbf{e}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = \sum_{i:y_i=0} \frac{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} [\dot{A}]_{\beta_j},$$

em que $A = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}$

ii. Modelo gama-log-logística inflacionado de zeros

$$\begin{aligned} l(\boldsymbol{\theta}) = & r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda) + \sum_{i:y_i=0} \log \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right] + (\alpha - 1) \times \\ & \sum_{i:y_i>0} \log(ty_i) + \sum_{i:y_i>0} \log[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] - 2 \sum_{i:y_i>0} \log \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right] + \\ & (\phi - 1) \sum_{i:y_i>0} \log \left\{ \log \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right] \right\}, \end{aligned} \quad (7.6)$$

em que $\boldsymbol{\theta} = (\phi, \sigma, \boldsymbol{\beta}^\top)^\top$ é o vetor de parâmetros desconhecidos e r é o número que observações diferentes de zero.

A função escore é obtida, calculando-se a derivada logaritmo da função de verossimilhança (7.6) em relação aos parâmetros do vetor $\boldsymbol{\theta}$. Assim,

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = -r\psi(\phi) + \sum_{i:y_i>0} \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\},$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} = & -\frac{r\alpha}{\lambda} + \frac{2\alpha}{\lambda^{\alpha+1}} \sum_{i:y_i>0} y_i^\alpha \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right]^{-1} - \frac{\alpha(\phi - 1)}{\lambda^{\alpha+1}} \sum_{i:y_i>0} y_i^\alpha \log^{-1} \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right] \times \\ & \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right]^{-1}, \end{aligned}$$

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} = & \frac{r}{\alpha} - r \log(\lambda) + \phi \sum_{i:y_i>0} \log(y_i) + \frac{\phi - 1}{\lambda^\alpha} \sum_{i:y_i>0} y_i^\alpha \log \left(\frac{y_i}{\lambda} \right) \log^{-1} \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right] \times \\ & \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right]^{-1} \quad \mathbf{e} \end{aligned}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = \sum_{i: y_i=0} \frac{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})} [\dot{A}]_{\beta_j}, \quad \text{em que} \quad A = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

Porém, as equações $\frac{\partial l(\boldsymbol{\theta})}{\partial \phi} = 0$, $\frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} = 0$, $\frac{\partial l(\boldsymbol{\theta})}{\partial \alpha} = 0$ e $\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = 0$ são não lineares e, em geral, não possuem soluções exatas. Soluções aproximadas são encontradas por métodos iterativos, por exemplo, Newton-Raphson ou um algoritmo de Quase-Newton. Os métodos iterativos para iniciar o algoritmo especificam uma estimativa inicial $\boldsymbol{\theta}^{(0)}$, itera-la sucessivamente até obter convergência.

Os intervalos de confiança e os testes de hipóteses são conduzidas utilizando a distribuição assintótica dos estimadores de máxima verossimilhança, que é uma distribuição normal com a matriz de covariância obtida como a inversa da matriz de informação de Fisher sob certas condições de regularidade satisfeitas. Mais especificamente, a matriz de covariância assintótica é dada por $\mathbf{I}^{-1}(\boldsymbol{\theta})$ em que $\mathbf{I}(\boldsymbol{\theta}) = E[\ddot{\mathbf{L}}(\boldsymbol{\theta})]$ tal que $\ddot{\mathbf{L}}(\boldsymbol{\theta}) = -\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$.

Como não é possível determinar a matriz de informação de Fisher $\mathbf{I}(\boldsymbol{\theta})$ devido a presença de observações censuradas, uma forma alternativa é utilizar a matriz de informação observada $\ddot{\mathbf{L}}(\boldsymbol{\theta})$, avaliada no estimador de máxima verossimilhança $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. A aproximação normal assintótica de $\hat{\boldsymbol{\theta}}$ é expressa como

$$\hat{\boldsymbol{\theta}}^\top \sim N_{(p+3)}[\boldsymbol{\theta}^\top; \ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1}],$$

em que $\ddot{\mathbf{L}}(\boldsymbol{\theta})$ é a matriz de informação observada $(p+3)(p+3)$.

Assim, o intervalo de confiança assintótico para β_j , com $j = 1, \dots, p$, considerando um nível de confiança de $(1 - \alpha)100\%$, é expresso por

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)},$$

em que $z_{\alpha/2}$ é o valor que excede com probabilidade $\alpha/2$ a distribuição normal padrão.

7.3 Análise de sensibilidade

A análise de sensibilidade é uma etapa posterior ao ajuste de um modelo a um conjunto de dados. Por meio dessa análise é possível encontrar possíveis pontos influentes que podem causar distorções nos resultados do ajuste do modelo (PAULA, 2004). Então, para encontrar esses pontos influentes, neste trabalho, considerou-se as medidas de influência global, distância de Cook definida em (3.17) e afastamento da verossimilhança definida em (3.18).

Por outro lado, as medidas de influência local, descritas nos capítulos anteriores, são calculadas após um esquema de perturbação no modelo ou nos dados ser adotado. Além disso, diferentes esquemas de perturbação podem ser considerados de acordo com o objetivo da análise.

Assim, considerando o modelo definido em (7.4) e os logaritmos das funções de verossimilhança expressos em (7.5) e em (7.6), os seguintes esquemas de perturbação foram utilizados nos modelos gama-G inflacionados de zeros:

a. *Perturbação de casos*

Seja $0 \leq \omega_i \leq 1$ e $\omega_0 = (1, \dots, 1)^\top$ o vetor de não perturbação. Então, o logaritmo da função de verossimilhança perturbada que incorpora diferentes pesos para cada elemento do conjunto de dados é definido por:

i. Modelo gama-Weibull inflacionado de zeros

$$l(\theta) = \{r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha\phi \log(\lambda)\} \sum_{i:y_i>0} \omega_i + \sum_{i:y_i=0} \omega_i \times \\ \log \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right] - \sum_{i:y_i>0} \omega_i \log[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] + \alpha\phi \sum_{i:y_i>0} \omega_i \log(y_i) - \\ \sum_{i:y_i>0} \omega_i \log(y_i) - \sum_{i:y_i>0} \omega_i \left(\frac{y_i}{\lambda} \right)^\alpha,$$

ii. Modelo gama-log-logística inflacionado de zeros

$$l(\theta) = \sum_{i:y_i=0} \omega_i \log \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right] + \{r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda)\} \sum_{i:y_i>0} \omega_i + \\ (\alpha - 1) \sum_{i:y_i>0} \omega_i \log(t_i) + \sum_{i:y_i>0} \omega_i \log[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] - 2 \sum_{i:y_i>0} \omega_i \times \\ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i:y_i>0} \omega_i \log \left\{ \log \left[1 + \left(\frac{t_i}{\lambda} \right)^\alpha \right] \right\},$$

b. *Perturbação da variável resposta*

Para verificar a sensibilidade do modelo (7.4), considera-se que a variável resposta $y_i (i = 1, \dots, n)$ é submetida a um esquema de perturbação aditiva por meio da adição de um vetor, denotado por ω , de pequenas perturbações, tal que $y_i^* = y_i + \omega_i \text{sd}(y_i)$, em que $\text{sd}(y_i)$ é um fator escala que pode ser a estimativa do desvio padrão da variável resposta e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Nesse caso, o logaritmo da função de verossimilhança perturbada, em que $\omega_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação, é expresso por:

i. Modelo gama-Weibull inflacionado de zeros

$$l(\theta) = \sum_{i:y_i=0} \log \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right] + r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha\phi \log(\lambda) - \\ \sum_{i:y_i>0} \log[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] + \alpha\phi \sum_{i:y_i>0} \log(y_i^*) - \sum_{i:y_i>0} \log(y_i^*) - \sum_{i:y_i>0} \left(\frac{y_i^*}{\lambda} \right)^\alpha,$$

ii. Modelo gama-log-logística inflacionado de zeros

$$l(\boldsymbol{\theta}) = \sum_{i:y_i=0} \log \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right] + r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda) +$$

$$(\alpha - 1) \sum_{i:y_i>0} \log(y_i^*) + \sum_{i:y_i>0} \log[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})] - 2 \sum_{i:y_i>0} \log \left[1 + \left(\frac{y_i^*}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i:y_i>0} \log \left\{ \log \left[1 + \left(\frac{y_i^*}{\lambda} \right)^\alpha \right] \right\},$$

c. **Perturbação de uma covariável**

Uma outra forma de avaliar a sensibilidade do modelo (7.4) é considerar pequenas perturbações em uma particular covariável contínua, denotada por X_j . Nesse caso, a covariável também é submetida a um esquema de perturbação aditiva, de tal forma que $x_{ij}^* = x_{ij} + \omega_i \text{sd}(x_{ij})$, em que $\text{sd}(x_{ij})$ é um fator de escala que pode ser a estimativa do desvio padrão da covariável perturbada e $\omega_i \in \mathbf{R}$ (SILVA; ORTEGA; CORDEIRO, 2010). Dessa forma, assumindo que $\mathbf{x}_i^{*\top} \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j [x_{ij} + \omega_i \text{sd}(x_{ij})] + \dots + \beta_p x_{ip}$ e $\boldsymbol{\omega}_0 = (0, \dots, 0)^\top$ é o vetor de não perturbação. O logaritmo da função de verossimilhança perturbada é dada por:

i. Modelo gama-Weibull inflacionado de zeros

$$l(\boldsymbol{\theta}) = \sum_{i:y_i=0} \log \left[\frac{\exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})} \right] + r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha\phi \log(\lambda) -$$

$$\sum_{i:y_i>0} \log[1 + \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})] + \alpha\phi \sum_{i:y_i>0} \log(y_i) - \sum_{i:y_i>0} \log(y_i) - \sum_{i:y_i>0} \left(\frac{y_i}{\lambda} \right)^\alpha,$$

ii. Modelo gama-log-logística inflacionado de zeros

$$l(\boldsymbol{\theta}) = \sum_{i:y_i=0} \log \left[\frac{\exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})} \right] + r \log(\alpha) - r \log[\Gamma(\phi)] - r\alpha \log(\lambda) +$$

$$(\alpha - 1) \sum_{i:y_i>0} \log(y_i) + \sum_{i:y_i>0} \log[1 + \exp(\mathbf{x}_i^{*\top} \boldsymbol{\beta})] -$$

$$2 \sum_{i:y_i>0} \log \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right] + (\phi - 1) \sum_{i:y_i>0} \log \left\{ \log \left[1 + \left(\frac{y_i}{\lambda} \right)^\alpha \right] \right\},$$

Para esses três esquemas de perturbação calculou-se, numericamente, a matriz de curvaturas máximas definida por

$$\boldsymbol{\Delta} = (\boldsymbol{\Delta}_{vi})_{(p+4) \times n} = \left[\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial \theta_v \partial \omega_i} \right]_{(p+4) \times n},$$

em que $v = 1, \dots, p + 4$, $i = 1, \dots, n$ e $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top$ é o vetor de perturbação, isto é, os pesos que perturbam os modelos ou as observações.

7.4 Aplicação

O exemplo de aplicação é referente a um estudo conduzido em uma área da Floresta Amazônica no estado do Pará. Os dados foram coletados pela pesquisadora Carine Klauberg do Departamento de Ciências Florestais da Escola Superior de Agricultura “Luiz de Queiroz” - ESALQ/USP, Piracicaba/São Paulo e pelo Professor Dr Edson José Vidal da Silva, docente do departamento e pesquisador responsável do projeto FAPESP Processo 2011/04968-4, cujo objetivo da pesquisa era estudar a melhor forma de manejo de espécies com potencial não madeireiro. O estudo inclui 93 árvores produtoras de óleo de copaíba (*Copaifera* sp.) de duas espécies, nos quais as árvores eram perfuradas para extração de óleo. Para esse estudo, foram consideradas as seguintes variáveis:

- i. y : produção de óleo (em litros);
- ii. x_1 : espécie de árvore de copaíba (0=*Copaifera* cf. *reticulada*, 1=*Copaifera* *duckei*);
- iii. x_2 : Diâmetro a altura do peito (1,30 metros) - DAP (em cm);
- iv. x_3 : Presença de pragas (0=ausência, 1=presença);
- v. x_4 : Qualidade do fuste (0=ruim, 1=boa).

Para uma análise descritiva dos dados de produção de óleo de copaíba, encontram-se na Figura 7.1, o histograma de frequências e um gráfico de boxplot da variável resposta. Enquanto que na Figura 7.2 são apresentados gráficos de boxplot em função das covariáveis dicotômicas.

Figura 7.1 – Histograma e boxplot da produção de óleo de copaíba

Figura 7.2 – Boxplot da produção de óleo de copaíba. (a) Por espécie de árvore de copaíba, (b) Por pragas e (c) Por qualidade do fuste

Conforme a Figura 7.1a, pode-se observar que a produção de óleo apresenta uma assimetria e a proporção de árvores que não produziram nenhuma quantidade de óleo corresponde a aproximadamente 53%, como ilustra a linha tracejada. Também é possível observar a presença de pontos discrepantes e nesse caso, a produção de óleo da árvore 27 indica estar bem distante das outras observações (Figura 7.1b).

Por meio das Figuras 7.2a e 7.2b, observa-se que há indicativos de que haja uma diferença entre as espécies de copaíba em relação a produção de óleo de copaíba e que a ausência ou presença de pragas pode estar influenciando a variável resposta. Por outro lado, a qualidade do fuste não está associada a produção de óleo, como ilustra a Figura 7.2c.

Então, de acordo com o que foi observado nas Figuras 7.1 e 7.2, o conjunto de dados é composto por um número grande de zeros, pois metade da amostra de copaibeiras não produziram óleo. Por essa razão, um dos modelos inflacionados de zeros, propostos neste capítulo, seria adequado para avaliar quais os fatores que são responsáveis pela ocorrência excessiva de zeros.

Assim, considerando o logaritmo da função de verossimilhança do modelo gama-Weibull inflacionado de zeros (7.5) e do modelo gama-log-logística inflacionado de zeros (7.3), as estimativas de máxima verossimilhança dos parâmetros desses modelos foram obtidas utilizando a sub-rotina `MaxBFGS` da linguagem matricial de programação `Øx` versão 6.20 com os valores iniciais de $\phi = \lambda = \alpha = 1$ e os coeficientes da regressão como -0.5 , para os dois modelos propostos.

Desse modo, as estimativas de máxima verossimilhança dos modelos ajustados, bem como os erros padrões (entre parênteses) são apresentados na Tabela 7.1.

Nessa situação, observa-se que ao nível de 5% de significância, o diâmetro da árvore e a qualidade do fuste não tem influência sobre a proporção de plantas que não produziram óleo, em ambos os modelos.

Tabela 7.1 – Estimativa de máxima verossimilhança dos parâmetros do modelo gama-G inflacionados de zeros para os dados de produção de óleo de copaíba

Parâmetro	Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)
Gama-Weibull inflacionado de zeros				
ϕ	4,4064	6,5752	–	(0,0000; 17,2935)
α	0,3470	0,2762	–	(0,0000; 0,8884)
λ	0,0161	0,1279	–	(0,0000; 0,2669)
β_0	0,2710	1,0328	0,7930	(-1,7533; 2,2952)
β_1	-1,5913	0,4960	0,0013	(-2,5634; -0,6191)
β_2	0,0191	0,0163	0,2412	(-0,0129; 0,0511)
β_3	-1,0638	0,5298	0,0447	(-2,1022; -0,0253)
β_4	-0,5986	0,5177	0,2476	(-1,6133; 0,4161)
Gama-log-logística inflacionado de zeros				
ϕ	0,5375	0,2583	–	(0,0313; 1,0436)
α	1,5591	0,4182	–	(0,7395; 2,3786)
λ	2,1507	1,2104	–	(0,0000; 4,5230)
β_0	0,2710	1,0328	0,7930	(-1,7532; 2,2952)
β_1	-1,5913	0,4960	0,0013	(-2,5634; -0,6191)
β_2	0,0191	0,0163	0,2413	(-0,0129; 0,0511)
β_3	-1,0638	0,5298	0,0447	(-2,1022; -0,0253)
β_4	-0,5986	0,5177	0,2476	(-1,6133; 0,4161)

Para dar continuidade a análise dos dados de produção de óleo de copaíba, a estatística AIC definida em (2.20) e a estatística BIC definida em (2.21) foram utilizadas para escolher o modelo mais adequado para esse conjunto de dados. Os valores das estatísticas são apresentadas na Tabela 7.2 e por meio desses resultados, observa-se que o modelo gama-Weibull inflacionado de zeros é o mais adequado, entre os modelos propostos, pois apresenta o menor valor das estatísticas.

Tabela 7.2 – Estatísticas AIC e BIC dos modelos gama-G inflacionados de zeros

Modelo	AIC	BIC
Gama-Weibull	268,8106	289,0714
Gama-log-logística	269,7894	290,0502

Por outro lado, utilizando o teste de hipótese para modelos não encaixados apresentados na Seção 3.3, na qual a hipótese nula considera que os modelos são equivalentes. A estatística do teste para esse conjunto de dados foi de $T = 8,7899$ e nesse caso, ao nível de 5% de significância, rejeita-se a hipótese nula. Portanto, o modelo gama-Weibull inflacionado de

zeros é o mais indicado para analisar esse conjunto de dados, em relação ao modelo gama-log-logística inflacionado de zeros.

Dessa forma, para prosseguir com a análise, foram calculadas as medidas de influência global e local para detectar possíveis pontos influentes presentes no conjunto de dados de produção de óleo de copaíba. Para essa etapa, foi utilizado o modelo gama-Weibull inflacionado de zeros e a linguagem matricial de programação Ox versão 6.20.

A distância de Cook generalizada e afastamento da verossimilhança são apresentados na Figura 7.3. A partir dessa figura, verifica-se que o ponto #27 é um possível ponto influente.

Nas Figuras 7.4 e 7.5, têm-se os gráficos de influência local, considerando as perturbações de caso ($C_{d_{max}} = 1,5167$), perturbação na quantidade de óleo extraída ($C_{d_{max}} = 0,2404$) e perturbação da covariável x_3 ($C_{d_{max}} = 0,1431$). A partir desses gráficos, é possível identificar os pontos #12 e #27 como possíveis observações influentes.

Considerando os pontos que apareceram com maior frequência nos gráficos, a análise de sensibilidade (influência global e influência local) detectou como possíveis pontos influentes as observações #12 e #27. Então, para verificar o impacto dessas observações nas estimativas dos parâmetros, reajustou-se o modelo eliminando-se, individualmente, cada observação, em seguida, eliminando as duas observações.

Figura 7.3 – Gráficos de medidas de influência global para o modelo gama-Weibull inflacionado de zeros ajustado aos dados de produção de óleo de copaíba. (a) Distância de Cook generalizada e (b) Afastamento da verossimilhança

Figura 7.4 – Gráficos de influência local $\mathbf{d}_{max}$ do modelo gama-Weibull inflacionado de zeros para os dados de produção de óleo de copaíba. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_2

Figura 7.5 – Gráficos de influência local total C_i do modelo gama-Weibull inflacionado de zeros para os dados de produção de óleo de copaíba. (a) Perturbação de casos, (b) Perturbação da resposta e (c) Perturbação da covariável x_2

Na Tabela 7.3, são apresentadas as mudanças relativas (em porcentagens) de cada parâmetro estimado, definido por $\mathbf{RC}_{\theta_j} = \left[(\hat{\theta}_j - \hat{\theta}_{j(i)}) / \hat{\theta}_j \right] \times 100$, em que $\theta_{j(i)}$ é a estimativa de máxima verossimilhança sem a i -ésima observação.

Os resultados da Tabela 7.3, mostram que a significância dos parâmetros estimados foi alterada, considerando um nível de 5% e logo, houve mudanças inferenciais após a remoção das observações consideradas influentes nos gráficos de diagnósticos, indicando que são pontos influentes no modelo. Segundo a pesquisadora Carine Klauberg, umas das responsáveis do estudo, as árvores apontadas como pontos influentes, tem as seguintes descrições:

- i. A árvore de número 12 (id=1021) é uma árvore jovem, fenotipicamente adequada por apresentar fuste de boa qualidade e não há presença de pragas no seu tronco. Dentre as árvores

manejadas, apresentou a menor produção na primeira coleta, correspondendo a 0,02 litros apenas.

- ii. A árvore de número 27 (id=949) é uma árvore com características fenotípica que se sobressai em relação as outras árvores de copaíba estudadas, produzindo, em média, cerca de oito vezes mais que a produção média de outras copaibeiras.

Tabela 7.3 – Mudanças relativas [-RC-em %], estimativas dos parâmetros e os correspondentes p -valores entre parênteses, excluindo as observações #12 e #27

Exclusão	$\hat{\phi}$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Nenhuma	-	-	-	-	-	-	-	-
	4,4064	0,3470	0,0161	0,2710	-1,5913	0,0191	-1,0638	-0,5986
	(-)	(-)	(-)	(0,7930)	(0,0013)	(0,2412)	(0,0447)	(0,2476)
#12	[-97]	[26]	[-1]	[98]	[3]	[5]	[-2]	[8]
	8,6662	0,2578	0,0003	0,3072	-1,5452	0,0181	-1,0828	-0,5501
	(-)	(-)	(-)	(0,7659)	(0,0019)	(0,2669)	(0,0406)	(0,2700)
#27	[62]	[-82]	[99]	[-3394]	[-7]	[-45]	[13]	[-11]
	1,6864	0,6318	0,5625	-0,2008	-1,7018	0,0277	-0,9288	-0,6616
	(-)	(-)	(-)	(0,8528)	(0,0008)	(0,1152)	(0,0797)	(0,2088)
#12 e #27	[31]	[-38]	[99]	[-620]	[-4]	[-39]	[10]	[-3]
	3,0570	0,4775	0,1159	-0,1604	-1,6550	0,0266	-0,9580	-0,6141
	(-)	(-)	(-)	(0,8821)	(0,0012)	(0,1302)	(0,0733)	(0,2451)

Logo, em acordo com a pesquisadora responsável pelo projeto, optou-se pela retirada da observação #27, pois se refere a uma copaibeira de característica fenotípica incomum. No entanto, é uma árvore que merece uma atenção especial, devido ao seu potencial produtivo. Nesse caso, um estudo genético poderia ser recomendado, pois dessa forma, é possível verificar se a alta produção está ligada os fatores genéticos da planta.

Desse modo, as estimativas do modelo final sem as covariáveis não significativas e sem a observação #27, são apresentas na Tabela 7.4.

Tabela 7.4 – Estimativa de máxima verossimilhança dos parâmetros do modelo de regressão gama-Weibull inflacionado de zeros final ajustado aos dados de produção óleo de copaíba sem a observação #27

Parâmetro	Estimativa	Erro padrão	p -valor	Intervalo de confiança(95%)
ϕ	1,6864	1,9623	–	(0,0000; 5,5324)
α	0,6318	0,4297	–	(0,0000; 1,4740)
λ	0,5625	1,5134	–	(0,0000; 3,5286)
β_0	1,1063	0,3753	0,0032	(0,3706; 1,8418)
β_1	-1,4385	0,4668	0,0021	(-2,3534; -0,5236)
β_3	-1,0557	0,5234	0,0437	(-2,0816; -0,0299)

Por meio dos resultados da Tabela 7.4, observa-se que a chance de uma árvore não produzir óleo é aproximadamente 24%(exp(-1,4385)) menor em árvores da espécie *Copaifera*

duckei do que em relação às árvores da espécie *Copaifera* cf. *Reticulada*. Por outro lado, a chance de uma árvore não produzir óleo é aproximadamente $35\%(\exp(-1,0557))$ menor em árvores com presença de pragas do que em árvores na ausência de pragas.

Além disso, substituindo os parâmetros do modelo em (7.2) por suas estimativas (Tabela 7.4), é possível determinar as frequências da produção de óleo de copaíba, que são apresentadas na Tabela 7.5. Por meio dessas frequências, observa-se que as copaibeiras apresentam produção inferior a 1 litro e as copaibeiras da espécie *Copaifera duckei* apresentam desempenho de produção melhor do que da espécie *Copaifera* cf. *Reticulada*, na presença de pragas.

Tabela 7.5 – Frequência da produção de óleo de copaíba

$f(y)$	<i>Copaifera</i> cf. <i>Reticulada</i>		<i>Copaifera duckei</i>	
	Ausência de pragas	Presença de pragas	Ausência de pragas	Presença de pragas
0,5	0,1208	0,2369	0,2831	0,3890
1,0	0,0759	0,1489	0,1779	0,2444
5,0	0,0067	0,0131	0,0156	0,0215

Finalmente, utilizando a expressão definida em (7.4), pode-se estimar a proporção de zeros presentes no conjunto de dados de produção de óleo de copaíba. Dessa forma, substituindo os parâmetros por suas estimativas, têm-se

$$\hat{p}_i = \frac{\exp(1,1063 - 1,4385x_{1i} - 1,0557x_{3i})}{1 + \exp(1,1063 - 1,4385x_{1i} - 1,0557x_{3i})} \quad \text{e} \quad \hat{p} = \frac{1}{93} \sum_{i=1}^{93} \hat{p}_i = 0,5326.$$

Logo, a estimativa da proporção de zeros é dada por $\hat{p} = 0,5326$; ou seja, aproximadamente 53% da amostra é formada por árvores que não produziram óleo.

8 CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentada duas novas distribuições de probabilidade, denominadas de distribuição gama-Weibull e distribuição gama-log-logística, obtidas a partir da função geradora de distribuições gama proposta por Zografos e Balakrishnan (2009). Uma das vantagens dessas distribuições é a flexibilidade em acomodar diferentes formas da função de taxa de falha e, também, serve para discriminação de modelos, pois possui como casos particulares, a distribuição Weibull, exponencial e gama, no caso da distribuição gama-Weibull, e a distribuição log-logística é um caso particular da distribuição gama-log-logística.

Dessa forma, podem-se destacar como principais contribuições deste trabalho:

- i. A apresentação de uma nova função geradora de distribuições proposta por Zografos e Balakrishnan (2009) que possui apenas um parâmetro, o que facilita na generalização de distribuições usuais.
- ii. Fornece material para auxiliar na análise de dados de sobrevivência no contexto de modelos de regressão locação-escala, no modelos de longa-duração e modelos com efeito aleatório.
- iii. Um estudo de simulação considerando algumas porcentagens de censuras permitiu encontrar, para o modelo regressão log-gama-Weibull e log-gama-logística, um resíduo cuja distribuição empírica sugere uma concordância razoável com a distribuição normal padrão.
- iv. A apresentação de um modelo inflacionado de zeros no contexto de variáveis contínuas, utilizada para analisar dados de produção de óleo copaiba.
- v. A utilização dos modelos em conjunto de dados reais, na área de zootecnia, florestal e ciências sociais.

8.1 Pesquisas futuras

Dando continuidade ao trabalho desenvolvimento durante o doutorado, os seguintes temas podem ser abordados:

- i. Considerando a função geradora de distribuição gama, desenvolver uma biblioteca no pacote estatístico R;
- ii. Propor um modelo de regressão log-gama-G bivariado para dados com censura à direita e censura intervalar;
- iii. Propor um modelo de longa-duração com efeito aleatório, utilizando as distribuições: gama-Weibull e gama-log-logística;
- iv. Estudar os resíduos para os modelos gama-G inflacionados de zeros.

REFERÊNCIAS

- AARSET, M.V. How to identify bathtub hazard rate. **IEEE Transactions on Reliability**, New York, v.36, p.106-108, 1987.
- ABRAMS, K.; ASHBY, D.; ERRINGTON, D. A Bayesian approach to Weibull survival models-application to a cancer clinical trial. **Lifetime Data Analysis**, Hingham, v.2, p.159-174, 1996.
- ALEXANDER, C.; CORDEIRO, G.M.; ORTEGA, E.M.M.; SARABIA, J.M. Generalized beta-generated distributions. **Computational Statistics and Data Analysis**, Amsterdam, v.56, p.1880-1897, 2012.
- ASHKAR, F.; AUCOIN, F. Discriminating between the lognormal and the log-logistic distributions for hydrological frequency analysis. **Journal of Hydrologic Engineering**, New York, v.17, p.160-167, 2012.
- ATKINSON, A.C. **Plots, transformations, and regression**. Oxford: University Press, 1985. 282p.
- BARLOW, W.E.; PRENTICE, R.L. Residuals for relative risk regression. **Biometrika**, Cambridge, v.75, p.65-74, 1988.
- BARRETO-SOUZA, W; CRIBARI-NETO, F. A generalization of the exponential-Poisson distribution. **Statistics and Probability Letters**, Amsterdam, v.79, p.2493-2500, 2009.
- BENNETT, S. Log-logistic regression model for survival data. **Journal of the Royal Statistical Society - Series C**, London, v.32, p.165-171, 1983.
- BERKSON, J.; GAGE, R.P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Boston, v.47, p.501-515, 1952.
- BONAT, W.H.; KRAINSKI, E.T.; RIBEIRO JR, P.J.; ZEVIANI, W.M. **Métodos computacional em inferência estatística**. Piracicaba: Região Brasileira da Sociedade Internacional de Biometria, 2012. 289p.
- BUTLER, R.J.; McDONALD, J.B. Using incomplete moments to measure inequality. **Journal of Econometrics**, Amsterdam, v.42, p.109-119, 1989.
- CABRAL, C.R.B.; LACHOS, V.H.; MADRUGA, M.R. Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population. **Journal of Statistical Planning and Inference**, Amsterdam, v.142, p.181-200, 2012.
- CANCHO, V.G.; LOUZADA-NETO, F.; BARRIGA, G.D.C. The Geometric Birnbaum-Saunders regression model with cure rate. **Journal of Statistical Planning and Inference**, Amsterdam, v.142, p.993-1000, 2012.

- CANCHO, V.G.; ORTEGA, E.M.M.; BARRIGA, G.D.C.; HASHIMOTO, E.M. The Conway-Maxwell-Poisson-generalized gamma regression model with long-term survivors. **Journal of Statistical Computation and Simulation**, New York, v.81, p.1461-1481, 2011.
- CANCHO, V.G.; ORTEGA, E.M.M.; PAULA, G.A. On estimation and influence diagnostics for log-Birnbaum Saunders Student-t regression models: full bayesian analysis. **Journal of Statistical Planning and Inference**, Amsterdam, v.140, p.2486-2496, 2010.
- CARRASCO, J.M.F.; ORTEGA, E.M.M.; PAULA, G.A. Log-Modified Weibull regression models with censored data: sensitivity and residual analysis. **Computational Statistics and Data Analysis**, Amsterdam, v.52, p.4021-4029, 2008.
- CHEN, M.H.; HARRINGTON, D.; IBRAHIM, J. Bayesian cure rate models for malignant melanoma: A case study of ECOG Trial E1690. **Applied Statistics**, London, v.51, p.135-150, 2002.
- CHEN, M.H., IBRAHIM, J.; SINHA, D. A new bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Boston, v.94, p.909-919, 1999.
- COLLETT, D. **Modelling survival data in medical research**. 2nd ed. London: Chapman & Hall, 2003. 391p.
- COLOSIMO, E.A.; GIOLO, S.R. **Análise de sobrevivência aplicada**. São Paulo: Edgard Blücher, 2006. 367p.
- COOK, R. D. Detection of influential observations in linear regression. **Technometrics**, Alexandria, v.19, p.15-18, 1977.
- _____. Assessment of local influence (with discussion). **Journal of the Royal Statistical Society**, London, v.48, p.133-169, 1986.
- COOK, R.D.; WEISBERG, S. **Residuals and influence in regression**. New York: Chapman and Hall, 1982. 230p.
- CORDEIRO, M.G.; de CASTRO, M. A new family of generalized distributions. **Journal of Statistical Computation and Simulation**, New York, v.81, p.883-898, 2011.
- CORDEIRO, G.M.; HASHIMOTO, E.M.; ORTEGA, E.M.M.; PASCOA, M.A.R. The McDonald extended distribution: properties and applications. **AStA Advances in Statistical Analysis**, New York, v.96, p.409-433, 2012.
- COX, D.R.; OAKES, D. **Analysis of survival data**. New York. Chapman and Hall, 1984. 201p.
- COX, D.R.; SNELL, E.J. A general definition of residuals. **Journal of the Royal Statistical Society - Series B**, Oxford, v.30, p.248-275, 1968.

de CASTRO, M.; CANCHO, V.G.; RODRIGUES, J. A Bayesian long-term survival model parametrized in the cured fraction. **Biometrical Journal**, Berlin, v.51, p.443-445, 2009.

DOORNIK, J. **Ox: Object-oriented matrix programming using Ox**. London : Timberlake Consultants Ltd, 2007. 461p.

ELLIOTT, J.; RAWLINGS, J.M.; MARKWELL, P.J.; BARBER, P.J. Survival of cats with naturally occurring chronic renal failure: effect of dietary management. **Journal of Small Animal Practice**, Oxford, v.41, p.235-242, 2000.

ENGLEHARDT, J.D.; LI, R. The discrete Weibull distributions: an alternative for correlated counts with confirmation for microbial counts in water. **Risk Analysis**, New York, v.31, p.370-381, 2011.

ESCOBAR, L.A.; MEEKER, W.Q. Assessing influence in regression analysis with censored data. **Biometrics**, Washington, v.48, p.507-528, 1992.

EUGENE, N; LEE, C; FAMOYE, F. Beta-normal distribution and its applications. **Communications in Statistics - Theory and Methods**, New York, v.31, p.497-512, 2002.

EVERITT, B.S.; HOTHORN, T. **A handbook of statistical analyses using R**. Boca Raton: Chapman & Hall, 2010. 335p.

FABIO, L.Z.; PAULA, G.A.; de CASTRO, M. A Poisson mixed model with nonnormal random effect distribution. **Computational Statistics and Data Analysis**, Amsterdam, v.56, p.1499-1510, 2012.

FACHINI, J.B.; ORTEGA, E.M.M.; LOUZADA-NETO, F. Influence diagnostics for polyhazard models in the presence of covariates. **Statistical Methods and Applications**, New York, v.17, p.413-433, 2008.

FAMOYE, F.; SINGH, K. Zero-inflated generalized Poisson regression model with an application to domestic violence data. **Journal of Data Science**, New York, v.4, p.117-130, 2006.

FAREWELL, V.T. The use mixture models for the analysis of survival data with long-term survivors. **Biometrics**, Washington, v.38, p.43-46, 1982.

FARRINGTON, C. P. Residuals for proportional hazards models with interval-censored survival data. **Biometrics**, Washington, v.56, p.473-482, 2000.

FLEMING, T.T.; HARRINGTON, D.P. **Counting process and survival analysis**. New York: Wiley, 1991. 448p.

GARAY, A.M.; HASHIMOTO, E.M.; ORTEGA, E.M.M.; LACHOS, V.H. On estimation and influence diagnostics for zero-inflated negative binomial regression models. **Computational Statistics and Data Analysis**, Amsterdam, v.55, p.1304-1318, 2011.

GIMÉNEZ, P.; COLOSIMO, E.; BOLFARINE, H. Asymptotic relative efficiency of score tests in Weibull models with measurement errors. **Statistical Papers**, New York, v.47, p.461-470, 2006.

GIOLO, S.R.; HENDERSON, R.; DEMÉTRIO, C.G.B.; MANSMANN, U. **Modelling time to gain a specified pre-weaning weight in Nelore cattle using survival models**. Curitiba: UFPR, Departamento de Estatística, 2004. Relatório Técnico.

GHOSH, S.K.; MUKHOPADHYAY, P.; LU, J.C. Bayesian analysis of zero-inflated regression models. **Journal of Statistical Planning and Inference**, Washington, v.136, p.1360-1375, 2006.

GRANA, C.; CHINOL, M.; ROBERTSON, C. et al. Pretargeted adjuvant radioimmunotherapy with Yttrium-90-biotin in malignant glioma patients: A pilot study. **British Journal of Cancer**, London, v.86, p.207-212, 2002.

HASHIMOTO, E.M.; ORTEGA, E.M.M.; CANCHO, V.G.; CORDEIRO, G.M. The log-exponentiated Weibull regression model for interval-censored data. **Computational Statistics and Data Analysis**, Amsterdam, v.54, p.1017-1035, 2010.

_____. On estimation and diagnostics analysis in generalized log-gamma regression model for interval-censored data. **Statistics**, In press, 2012.

HASHIMOTO, E.M.; ORTEGA, E.M.M.; CORDEIRO, G.M.; BARRETO, M.L. The log-burr XII regression model for grouped survival data. **Journal of Biopharmaceutical Statistics**, Monticello, v.22, p.141-159, 2012.

HEDEKER, D.; SIDDIQUI, O.; HU, F.B. Random-effects regression analysis of correlated grouped-time survival data. **Statistical Methods in Medical Research**, London, v.9, p.161-179, 2000.

HOGGART, C.J.; GRIFFIN, J.E. A bayesian partition model for customer attrition. In: George, E.I. (Ed.). **Bayesian Methods with Applications to Science, Policy, and Official Statistics (Select Papers from ISBA 2000)**, Creta: International Society for Bayesian Analysis, 2001. p.61-70.

HORTA, J.S.B.; MORAES, M.C.M. O sistema CAPES de avaliação da pós-graduação: da área de educação à grande área de ciências humanas. **Revista Brasileira de Educação**, Belo Horizonte, v.30, p.95-181, 2005.

IBRAHIM, J.G.; CHEN, M.H.; SINHA, D. Bayesian semi-parametric models for survival data with a cure fraction. **Biometrics**, Washington, v.57, p.383-388, 2001.

JOHNSON, N.L.; KEMP, A.W.; KOTZ, S. **Univariate discrete distributions**. New Jersey: John Wiley & Sons, 2005. 592p.

JUNG, K. Local influence in generalized estimating equations. **Scandinavian Journal of Statistics**, Oxford, v.35, p.286-294, 2008.

KALBFLEISH, J.D.; PRENTICE, R.L. **The statistical analysis of failure time data**. John Wiley: New York, 2002. 439p.

KENNEY, P.F.; KEEPING, E.S. **Mathematics of statistics**, New Jersey: Van Nostrand, 1962. 260p.

KLEIN, C.A. Tensile strength of randomly perforated aluminium plates: Weibull distribution parameters. **Journal of Applied Physics**, Washington, v.104, p.13501-13506, 2008.

KLEIN, J.P.; MOESCHBERGER, M.L. **Survival analysis techniques for censored and truncated data**. New York: Springer-Verlang, 2003. 536p.

LAGAKOS, S.W. The graphical evaluation of explanatory variables in proportional hazards regression. **Biometrika**, Cambridge, v.568, p.93-98, 1980.

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. **Technometrics**, Alexandria, v.34, p.1-14, 1992.

LAWLESS, J.F. **Statistical models and methods for lifetime data**. 2nd ed. New Jersey: John Wiley & Sons, 2003. 664p.

LEE, K.J.; THOMPSON, S. Flexible parametric models for random-effects distributions. **Statistics in Medicine**, Chichester, v.27, p.418-434, 2008.

LEIVA, V.; BARROS, M.; PAULA, G.A.; GALEA, M. Influence Diagnostics in Log-Birnbaum-Saunders Regression Models with Censored Data. **Computational Statistics and Data Analysis**, Amsterdam, v.51, p.5694-5707, 2007.

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

LECOUTRE, B.; MABIKA, B.; DERZKO, G. Assessment and monitoring in clinical trials when survival curves have distinct shapes: a Bayesian approach with Weibull modelling. **Statistics in Medicine**, Chichester, v.21, p.663-674, 2002.

LIANG, H.; ZOU, G. Improved AIC selection strategy for survival analysis. **Computational Statistics and Data Analysis**, Amsterdam, v.52, p.2538-2548, 2008.

LINDSEY, J.C.; RYAN, L.M. Tutorial in biostatistics methods for interval-censored data. **Statistics in Medicine**, Chichester, v.17, p.129-238, 1998.

LU, B.; SONG, X. Local influence analysis of multivariate probit latent variable models. **Journal of Multivariate Analysis**, New York, v.97, p.1783-1798, 2006.

MAGALHÃES, M.N. **Probabilidade e variáveis aleatórias**. São Paulo: EDUSP, 2011. 428p.

- MALLER, R.; ZHOU, X. **Survival analysis with long-term survivors**. New York: Wiley, 1996. 278p.
- McCULLOCH, C.E.; NEUHAUS, J.M. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. **Statistical Science**, Hayward, v.26, p.338-402, 2011.
- MEEKER, W.Q.; ESCOBAR, L.A. **Statistical methods for reliability data**. New York: John Wiley & Sons, 1998. 712p.
- MOOD, A.M.; GRAYBILL, F.A.; BOES, D. **Introduction to the theory of statistics**. Open University Pres, 1974. 564p.
- MOORS, J.J.A. A quantile alternative for kurtosis. **Journal of the Royal Statistical Society - Series D**, Abingdon, v.37, p.25-32, 1988.
- MOREIRA, N.P.; SILVEIRA, S.F.R.; FERREIRA, M.A.M.; CUNHA, N.R.S. Eficiência e qualidade dos programas de pós-graduação das instituições federais de ensino superior usuárias do programa de fomento à pós-graduação (prof). **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v.18, p.365-388, 2010.
- MUDHOLKAR, G.S.; SRIVASTAVA, D.K.; KOLLIA, G.D. A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data. **Journal of the American Statistical Association**, Boston, v.91, p.1575-1583, 1996.
- NADARAJAH, S.; KOTZ, S. The beta exponential distribution. **Reliability Engineering and System Safety**, Barking, v.91, p.689-697, 2006.
- ODELL, P.M.; ANDERSON, K.M; D'AGOSTINO, R.B. Maximum likelihood estimation for interval-censored data using a weibull-based accelerated failure time model. **Biometrics**, Washington, v.48, p.951-959, 1992.
- ORTEGA, E.M.M.; BOLFARINE, H.; PAULA G.A. Influence diagnostics in generalized log-gamma regression models. **Computational Statistics and Data Analysis**, Amsterdam, v.42, p.165-186, 2003.
- ORTEGA, E.M.M.; CANCHO, V.G.; BOLFARINE, H. Influence diagnostics in exponentiated -Weibull regression models with censored data. **Statistics and Operations Research Transactions**, Catalunya, v.30, p.172-192, 2006.
- ORTEGA, E.M.M.; CORDEIRO, G.M.; HASHIMOTO, E.M. A log-linear regression model for the beta-Weibull distribution. **Communication in Statistics: Simulation and Computation**, New York, v.40, p.1206-1235, 2011.
- ORTEGA, E.M.M.; PAULA, G.A.; BOLFARINE, H. Deviance residuals in generalised log-gamma regression models with censored observations. **Journal of Statistical Computation and Simulation**, New York, v. 78, p.747-764, 2008.

OSPINA, R.; FERRARI, S.L.P. Inflated beta distributions. **Statistical Papers**, New York, v.51, p.111-126, 2010.

_____. A general class of zero-or-one inflated beta regression models. **Computational Statistics and Data Analysis**, Amsterdam, v.56, p.1609-1623, 2012.

PASCOA, M.A.R.; ORTEGA, E.M.M.; CORDEIRO, G.M. The Kumaraswamy generalized gamma distribution with application in survival analysis. **Statistical Methodology**, Amsterdam, v.8, p.411-433, 2011.

PAULA, G.A. **Modelos de regressão com apoio computacional**. São Paulo, 2004. 403p.

PETRIE, A.; WATSON, P. **Estatística em ciência animal e veterinária**. ROCA, 2009. 238p.

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

PESCIM, R.R.; DEMÉTRIO, C.G.B.; CORDEIRO, G.M.; ORTEGA, E.M.M.; URBANO, M.R. The beta generalized half-normal distribution. **Computational Statistics and Data Analysis**, Amsterdam, v.54, p.945-957, 2010.

POON, W.; POON, Y.S. Local Conditional Influence. **Journal of Applied Statistics**, Abingdon, v.34, p.997-1009, 2007.

RIDOUT, M.S.; HINDE, J.P.; DEMÉTRIO, C.G.B. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. **Biometrics**, Washington, v.57, p.219-223, 2001.

RODRIGUES, J.; CANCHO, V.G.; de CASTRO, M. Unified theory of survival analysis. **Brazilian Statistical Association**, São Paulo, 2007. 94p.

RODRIGUES, J.; CANCHO, V.G.; de CASTRO, M.; LOUZADA-NETO, F. On the unification of the long-term survival models. **Statistics and Probability Letters**, Amsterdam, v.79, p.753-759, 2009.

ROMEO, J.S.; TANAKA, N.I.; LIMA, A.C.P. Bivariate survival modeling: a Bayesian approach based on Copulas. **Lifetime Data Analysis**, Hingham, v.12, p.205-222, 2006.

ROSS, S.M. **Simulation**. Boston: Elsevier Academic Press, 2006. 282p.

SANTANA, T.V.F.; ORTEGA, E.M.M.; CORDEIRO, G.M.; SILVA, G.O. The Kumaraswamy-log-logistic distribution. **Journal of Statistical Theory and Applications**, Holland, v.11, p.265-291, 2012.

SPARLING, Y.H.; YOUNES, N.; LACHIN, M.J. Parametric survival models for interval-censored data with time dependent covariates. **Biostatistics**, Oxford, v.7, p.599-614, 2006.

SHAYAN, Z.; AYATOLLAHI, S.M.T.; ZARE, N. A parametric method for cumulative incidence modeling with a new four-parameter log-logistic distribution. **Theoretical Biology and Medical Modeling**, London, v.8, p.1-11, 2011.

SHAWKY, A.I.; ABU-ZINADAH, H.H. Characterizations of the exponentiated pareto distribution based on record values. **Applied Mathematical Sciences**, New York, v.2, p.1283-1290, 2008.

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

SHOUKRI, M.M.; MIAN, I.U.M; TRACY, D.S. Sampling properties of estimators of the log-logistic distribution with application to Canadian precipitation data. **The Canadian Journal of Statistics**, Ottawa, v.16, p.223-236, 1988.

SILVA, G.O.; ORTEGA, E.M.M.; CANCHO, V.G.; BARRETO, M.L. Log-Burr XII regression models with censored data. **Computational Statistics and Data Analysis**, Amsterdam, v.52, p.3820-3842, 2008.

SILVA, G.O.; ORTEGA, E.M.M.; CORDEIRO, G.M. A log-extended Weibull regression model. **Computational Statistics and Data Analysis**, Amsterdam, v.53, p.4482-4489, 2009.

_____. The beta modified Weibull distribution. **Lifetime Data Analysis**, Hingham, v.16, p.409-430, 2010.

SILVA, G.O.; ORTEGA, E.M.M.; PAULA, G.A. Residuals for log-Burr XII regression models in survival analysis. **Journal of Applied Statistics**, Abingdon, v.38, p.1435-1445, 2011.

THERNEAU, T.M.; GRAMBSCH, P.M.; FLEMING, T.R. Martingale-based residuals for survival models. **Biometrika**, Cambridge, v.77, p.147-160, 1990.

TSODIKOV, A.D.; IBRAHIM, J.G.; YAKOVLEV, A.Y. Estimating cure rates from survival data: an alternative to two-component mixture models, **Journal of the American Statistical Association**, Boston, v.98, p.1063-1078, 2003.

VAIDA, F.; XU, R. Proportional hazards model with random effects, **Statistics in Medicine**, Chichester, v.19, p.3309-3324, 2000.

VERBEKE, G.; LESAFFRE, E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data, **Computational Statistics and Data Analysis**, Amsterdam, v.23, p.541-556, 1997.

VIEIRA, A.M.C.; HINDE, J.P.; DEMÉTRIO, C.G.B. Zero-inflated proportion data models applied to a biological control assay. **Journal of Applied Statistics**, Abingdon, v.27, p.373-389, 2000.

VILLEGAS, C.; PAULA, G.A.; LEIVA, V. Birnbaum-Saunders Mixed Models for Censored Reliability Data Analysis. **IEEE Transactions on Reliability**, New York, v.60, p.748-758, 2011.

WEIBULL, W. A statistical distribution of wide applicability. **Journal of Applied Mechanics**, New York, v.18, p.293-297, 1951.

WEISBERG, S. **Applied linear regression**. New York: John Wiley & Sons, 2005. 336p.

XIANG, L.; YAU, K.K.W.; TSE, S.K.; LEE, H. Influence diagnostics for random effect survival models: application to a recurrent infection study for kidney patients on portable dialysis. **Computational Statistics and Data Analysis**, Amsterdam, v.51, p.5977-5993, 2007.

XIE, F.; WEI, B. Diagnostics analysis for log-Birnbaum-Saunders regression models. **Computational Statistics and Data Analysis**, Amsterdam, v.51, p.4692-4706, 2007.

YAKOVLEV, A.Y.; TSODIKOV, A.D. **Stochastic models of tumor latency and their biostatistical applications, volume 1 of Mathematical Biology and Medicine**. World Scientific: New Jersey, 1996.

YAMAGUCHI, K. Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "permanent employment" in Japan, **Journal of the American Statistical Association**, Boston, v.87, p.284-292, 1992.

YAZDI, M.H.; VISSCHER, P.M.; DUCROCQ, V.; THOMPSON, R. Heritability, reliability of genetic evaluations and response to selection in proportional hazard models. **Journal of Dairy Science**, Lancaster, v.85, p.1563-1577, 2002.

ZARE, N.; DOOSTFATEMEH, M.; REZAIANZADEH, A. Modeling of breast cancer prognostic factors using a parametric log-logistic model in Fars Province, Southern Iran. **Asian Pacific Journal of Cancer Prevention**, Bangkok, v.13, p.1533-1537, 2012.

ZEVIANI, W.M. **Estatística computacional**. Disponível em: <<http://www.leg.ufpr.br/~walmes/ensino/ce083-2012-01/>>. Acesso em: 30 set. 2012.

ZHANG, D.; DAVIDIAN, M. Linear mixed models with flexible distributions of random effects for longitudinal data. **Biometrics**, Washington, v.57, p.795-802, 2001.

ZOGRAFOS, K.; BALAKRISHNAN, N. On families of beta-and generalized gama-generated distributions and associated inference, **Statistical Methodology**, Amsterdam, v.6, p.344-362, 2009.

APÊNDICES

APÊNDICE A- Programa computacional para estimação de parâmetros

A.1. Estimação dos parâmetros da distribuição gama-Weibull e gama-log-logística

```

/*****
PROGRAMA: GDG_censored.ox
*****/
#include<oxstd.h> #include<oxdraw.h> #include<oxprob.h> #include<oxfloat.h>
#include<maximize.h> #include<simula.h> #pragma link("maximize.oxo")
static decl n,g_mY;
//***** Log-verossimilhanca GWeibull *****/
lvGW(const vP, const adFunc, const avScore, const amHessian){
decl cont,t,x,div;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0];
decl lambda=vP[1];
if(lambda<0.01){
lambda=0.01; }
decl alfa=vP[2];
for(cont=0;cont<n;++cont){
t=g_mY[cont][0];
x=(t/lambda)^alfa;
if(g_mY[cont][1]==1){
vero[0][cont]=log(alfa)-((alfa*a)*log(lambda))-loggamma(a)+(((alfa*a)-1)*log(t))-((t/lambda)^alfa);
}
else{
vero[0][cont]=log(1-probgamma(x,a,1));
}
}
adFunc[0]=double(vero*uns);
return 1;
}
//***** Log-verossimilhanca GLL *****/
lvGLL(const vP, const adFunc, const avScore, const amHessian){
decl cont,t,x,div;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0];
decl lambda=vP[1];

```

```

if(lambda<0.01){
lambda=0.01; }
decl alfa=vP[2];
for(cont=0;cont<n;++cont){
t=g_mY[cont][0];
x=log(1+((t/lambda)^alfa));
if(g_mY[cont][1]==1){
vero[0][cont]=log(alfa)-(alfa*log(lambda))-loggamma(a)+((alfa-1)*log(t))-
(2*log(1+((t/lambda)^alfa)))+((a-1)*log(x));
}
else{
vero[0][cont]=log(1-probgamma(x,a,1));
}
}
adFunc[0]=double(vero*uns);
return 1;
}
main(){
g_mY=loadmat("gatos.txt");
print("Datos", g_mY);
n=rows(g_mY);
decl dfunc1,dfunc2,dfunc3,ir1,ir2,ir3,i,j,ep1,ep3,ep2;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
print("\n===== Gama Weibull ===== \n");
decl vP1=<1;673.129;1.519177>;
lvGW(vP1,&dfunc1,0,0);
decl mhess1=0.000001*unit(3);
MaxControl(-1,20);
ir1=MaxBFGS(lvGW,&vP1,&dfunc1,&mhess1,1);
if (ir1 == MAX_CONV){
if (Num2Derivative(lvGW,vP1,&mhess1)){
mhess1 = invertgen(-mhess1,0);
ep1=sqrt(diagonal(mhess1));
}
}
decl m=rows(vP1);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){

```

```

ic1[i][0]=vP1[i]-zi*ep1[i];
ic1[i][1]=vP1[i]+zi*ep1[i];
}
print("\n  Estimativa  Erro padrao  p_valor  IC(95)", "%13.5f",
vP1~ep1'~(2*(1-probn(fabs(vP1'./ep1))))~ic1);
print("\n=====  Gama log-logistica  ===== \n");
decl vP2=<1;673.129;1.519177>;
lvGLL(vP2,&dfunc2,0,0);
decl mhess2=0.000001*unit(3);
ir2=MaxBFGS(lvGLL,&vP2,&dfunc2,&mhess2,1);
if (ir2 == MAX_CONV){
if (Num2Derivative(lvGLL,vP2,&mhess2)){
mhess2 = invertgen(-mhess2,0);
ep2=sqrt(diagonal(mhess2));
}
}
decl ic2=zeros(m,2);
for(i=0;i<m;++i){
ic2[i][0]=vP2[i]-zi*ep2[i];
ic2[i][1]=vP2[i]+zi*ep2[i];
}
print("\n  Estimativa  Erro padrao  p_valor  IC(95)", "%13.5f",
vP2~ep2'~(2*(1-probn(fabs(vP2'./ep2))))~ic2);
}

```

A.2. Estimação dos parâmetros da distribuição log-gama-Weibull e log-gama-logística

```

/*****
PROGRAMA: GDG_DOinter.ox
*****/
#include<oxstd.h> #include<oxdraw.h> #include<oxprob.h> #include<oxfloat.h>
#include<maximize.h> #include<simula.h> #pragma link("maximize.oxo")
static decl n, g_mY;
//***** Log-verossimilhanca LGW *****/
lvGW(const vP, const adFunc, const avScore, const amHessiam){
decl cont,y,x,xx1,xx2,xx3,xx12,xx13,xx23,xbeta,z;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0];
decl sigma=vP[1];

```

```

decl beta0=vP[2];
decl beta1=vP[3];
decl beta2=vP[4];
decl beta3=vP[5];
for(cont=0;cont<n;++cont){
y=log(g_mY[cont][0]);
xx1=g_mY[cont][2];
xx2=g_mY[cont][3];
xx3=xx1*xx2;
xbeta=beta0+(beta1*xx1)+(beta2*xx2)+(beta3*xx3);
z=(y-xbeta)/sigma;
x=exp(z);
if(g_mY[cont][1]==1){
vero[0][cont]=-log(sigma)-loggamma(a)+(a*z)-x;
}
else{
vero[0][cont]=log(1-probgamma(x,a,1)+1e-10);
}
}
adFunc[0]=double(vero*uns);
return 1;
}
/***** Log-verossimilhanca LGL *****/
lvGLL(const vP, const adFunc, const avScore, const amHessian){
decl cont,y,x,xx1,xx2,xx3,xx12,xx13,xx23,xbeta,z;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0];
decl sigma=vP[1];
decl beta0=vP[2];
decl beta1=vP[3];
decl beta2=vP[4];
decl beta3=vP[5];
for(cont=0;cont<n;++cont){
y=log(g_mY[cont][0]);
xx1=g_mY[cont][2];
xx2=g_mY[cont][3];
xx3=xx1*xx2;
xbeta=beta0+(beta1*xx1)+(beta2*xx2)+(beta3*xx3);

```

```

z=(y-xbeta)/sigma;
x=log(1+exp(z));
if(g_mY[cont][1]==1){
vero[0][cont]=-log(sigma)-loggamma(a)+z-(2*x)+((a-1)*log(x+1e-10));
}
else{
vero[0][cont]=log(1-probgamma(x,a,1)+1e-10);
}
}
adFunc[0]=double(vero*uns);
return 1;
}
main(){
g_mY=loadmat("DadosDO1999.txt");
n=rows(g_mY);
print("Tamanho da amostra", n);
decl dfunc1,dfunc2,ir1,ir2,i,j,ep1,ep2;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
print("\n===== Log-gama-Weibull ===== \n");
decl vP1=<0.5;0.064;4.1289;-0.0044;-0.4227;0.0101>;
lvGW(vP1,&dfunc1,0,0);
decl mhess1=0.000000000001*unit(6);
//MaxControl(-1,20);
MaxControlEps(0.01,0.01);
ir1=MaxBFGS(lvGW,&vP1,&dfunc1,&mhess1,1);
Num2Derivative(lvGW,vP1,&mhess1);
ep1= sqrt(diagonal(-1/mhess1));
decl m=rows(vP1);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP1[i]-zi*ep1[i];
ic1[i][1]=vP1[i]+zi*ep1[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir1) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP1~ep1'~(2*(1-probn(fabs(vP1'./ep1))))'~ic1);
print("\n===== Log-gama-logistica ===== \n");
decl vP2=<0.5;0.0454;4.0281;-0.0025;-0.4390;0.0104>;

```



```

lvGLL(vP2,&dfunc2,0,0);
decl mhess2=0.000001*unit(6);
ir2=MaxBFGS(lvGLL,&vP2,&dfunc2,&mhess2,1);
Num2Derivative(lvGLL,vP2,&mhess2);
ep2= sqrt(diagonal(-1/mhess2));
decl ic2=zeros(m,2);
for(i=0;i<m;++i){
ic2[i][0]=vP2[i]-zi*ep2[i];
ic2[i][1]=vP2[i]+zi*ep2[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir2) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP2~ep2'~(2*(1-probn(fabs(vP2'./ep2)))')~ic2);
/***** Residuos *****/
decl a1,sigma1,beta01,beta11,beta21,beta31,xbeta1,z1,x1,a2,sigma2,
beta02,beta12,beta22,beta32,xbeta2,z2,x2;
decl rdi1=zeros(n,1);
decl rmi1=zeros(n,1);
decl rdi2=zeros(n,1);
decl rmi2=zeros(n,1);
decl SLGW=zeros(n,1);
decl SLGLL=zeros(n,1);
for(i=0;i<n;i++){
a1=vP1[0];
sigma1=vP1[1];
beta01=vP1[2];
beta11=vP1[3];
beta21=vP1[4];
beta31=vP1[5];
xbeta1=beta01+(beta11*g_mY[i][2])+(beta21*g_mY[i][3])+(beta31*(g_mY[i][2]*g_mY[i][3]));
z1=(log(g_mY[i][0])-xbeta1)/sigma1;
x1=exp(z1);
SLGW[i]=1-probgamma(x1,a1,1)+1e-3;
a2=vP2[0];
sigma2=vP2[1];
beta02=vP1[2];
beta12=vP1[3];
beta22=vP1[4];
beta32=vP1[5];

```

```

xbeta2=beta02+(beta12*g_mY[i][2])+(beta22*g_mY[i][3])+(beta32*(g_mY[i][2]*g_mY[i][3]));
z2=(log(g_mY[i][0])-xbeta2)/sigma2;
x2=log(1+exp(z2));
SLGLL[i]=1-probgamma(x2,a2,1)+1e-3;
if(g_mY[i][1]==1){
rmi1[i]=1+log(SLGW[i]);
rmi2[i]=1+log(SLGLL[i]);
rdi1[i]=(rmi1[i]/fabs(rmi1[i]))*((-2*(rmi1[i]+(log(g_mY[i][1]-rmi1[i]))))^(1/2));
rdi2[i]=(rmi2[i]/fabs(rmi2[i]))*((-2*(rmi2[i]+(log(g_mY[i][1]-rmi2[i]))))^(1/2));
}
else{
rmi1[i]=log(SLGW[i]);
rmi2[i]=log(SLGLL[i]);
if(rmi1[i]<1e-2 && rmi1[i]>-1e-4)
rdi1[i]=(rmi1[i]/fabs(rmi1[i]))*((2*rmi1[i])^(1/2));
rdi2[i]=(rmi1[i]/fabs(rmi1[i]))*((2*rmi2[i])^(1/2));
}
else{
rdi1[i]=(rmi1[i]/fabs(rmi1[i]))*((-2*rmi1[i])^(1/2));
rdi2[i]=(rmi2[i]/fabs(rmi2[i]))*((-2*rmi2[i])^(1/2));
}
}
}
print("\n Residuos ",rmi1~rdi1~rmi2~rdi2);
}

```

A.3. Estimação dos parâmetros da distribuição log-gama-Weibull e log-gama-logística para dados com censura intervalar

```

/*****
PROGRAMA: GDG_ODE.ox
*****/
#include<oxstd.h> #include<oxdraw.h> #include<oxprob.h> #include<oxfloat.h>
#include<maximize.h> #include<simula.h> #pragma link("maximize.oxo")
static decl n, g_mY;
//***** Log-verossimilhanca LGW *****/
lvGW(const vP, const adFunc, const avScore, const amHessian){
decl cont,tu,tv,xx0,xx1,xx2,xx3,xx4,xx5;
decl uns=ones(n,1);
decl vero=zeros(1,n);

```

```

decl a=vP[0];
decl sigma=vP[1];
decl beta0=vP[2];
decl beta1=vP[3];
decl beta2=vP[4];
decl beta3=vP[5];
decl beta4=vP[6];
decl beta5=vP[7];
for(cont=0;cont<n;++cont){
tu=g_mY[cont][1];
tv=g_mY[cont][2];
xx1=g_mY[cont][4];
xx2=g_mY[cont][5];
xx3=g_mY[cont][7];
xx4=g_mY[cont][8];
xx5=g_mY[cont][9];
decl xbeta=beta0+(beta1*xx1)+(beta2*xx2)+(beta3*xx3)+(beta4*xx4)+(beta5*xx5);
decl zu=(log(tu)-xbeta)/sigma;
decl zv=(log(tv)-xbeta)/sigma;
decl xu=exp(zu);
decl xv=exp(zv);
if(g_mY[cont][3]==3 || g_mY[cont][3]==2){
vero[0][cont]=log(probgamma(xv,a,1)-probgamma(xu,a,1)+1e-10);
}
else{
vero[0][cont]=-log(sigma)-loggamma(a)+(a*zu)-xu;
}
}
adFunc[0]=double(vero*uns);
return 1;
}
//***** Log-verossimilhanca LGL *****
lvGLL(const vP, const adFunc, const avScore, const amHessian){
decl cont,tu,tv,xx0,xx1,xx2,xx3,xx4,xx5;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0];
decl sigma=vP[1];
decl beta0=vP[2];

```

```

decl beta1=vP[3];
decl beta2=vP[4];
decl beta3=vP[5];
decl beta4=vP[6];
decl beta5=vP[7];
for(cont=0;cont<n;++cont)
tu=g_mY[cont][1];
tv=g_mY[cont][2];
xx1=g_mY[cont][4];
xx2=g_mY[cont][5];
xx3=g_mY[cont][7];
xx4=g_mY[cont][8];
xx5=g_mY[cont][9];
decl xbeta=beta0+(beta1*xx1)+(beta2*xx2)+(beta3*xx3)+ (beta4*xx4)+(beta5*xx5);
decl zu=(log(tu)-xbeta)/sigma;
decl zv=(log(tv)-xbeta)/sigma;
decl xu=log(1+exp(zu));
decl xv=log(1+exp(zv));
if(g_mY[cont][3]==3 || g_mY[cont][3]==2){
vero[0][cont]=log(probgamma(xv,a,1)-probgamma(xu,a,1)+1e-10);
}
else{
vero[0][cont]=-log(sigma)-loggamma(a)+zu-(2*xu)+((a-1)*log(xu+1e-10));
}
}
adFunc[0]=double(vero*uns);
return 1;
}
main(){
g_mY=loadmat("Ovulacao_ODE.txt");
n=rows(g_mY);
decl dfunc1,dfunc2,dfunc3,ir1,ir2,i,j,ep1,ep2;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
print("\n===== Log-gama-Weibull ===== \n");
//decl vP1=<1;0.50666;3.6148;0.043700;0.10160;-0.087610;-0.11438;0.5>;
lvGW(vP1,&dfunc1,0,0);
decl mhess1=0.000001*unit(8);
//MaxControl(-1,20);

```

```

ir1=MaxBFGS(lvGW,&vP1,&dfunc1,&mhess1,1);
Num2Derivative(lvGW,vP1,&mhess1);
ep1=sqrt(diagonal(-1/mhess1));
print(mhess1);
decl m=rows(vP1);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP1[i]-zi*ep1[i];
ic1[i][1]=vP1[i]+zi*ep1[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir1) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP1~ep1'~(2*(1-probn(fabs(vP1'./ep1))))'~ic1);
print("\n===== Log-gama-logistica ===== \n");
decl vP2=<1;0.50666;3.6148;0.043700;0.10160;-0.087610;-0.11438;0.5>;
lvGLL(vP2,&dfunc2,0,0);
decl mhess2=0.000001*unit(8);
ir2=MaxBFGS(lvGLL,&vP2,&dfunc2,&mhess2,1);
Num2Derivative(lvGLL,vP2,&mhess2);
ep2=sqrt(diagonal(-1/mhess2));
decl ic2=zeros(m,2);
for(i=0;i<m;++i){
ic2[i][0]=vP2[i]-zi*ep2[i];
ic2[i][1]=vP2[i]+zi*ep2[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir2) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP2~ep2'~(2*(1-probn(fabs(vP2'./ep2))))'~ic2);
}

```

A.4. Estimação dos parâmetros do modelo de longa-duração Poisson gama-Weibull e Poisson gama-log-logistica

```

/*****
PROGRAMA: PoissonGW_Glioma.ox
*****/
#include<oxstd.h> #include<oxdraw.h> #include<oxprob.h> #include<oxfloat.h>
#include<maximize.h> #include<simula.h> #pragma link("maximize.oxo")
static decl g_mX,g_mY,g_mU,g_mC,g_mD;
/***** Log-verossimilhanca Poisson GWeibull *****/

```

```

lvGW(const vP,const adFunc, const avScore,const amHessian){
  decl n=rows(g_mY);
  decl cont,t,xx0,xx1,xx2,xx3,xx4,xx5,xx6,xx7,xx8,ft,Ft,p0;
  decl uns=ones(n,1);
  decl vero=zeros(1,n);
  decl a=vP[0][0];
  decl lambda=vP[1][0];
  if(lambda<0.01){
    lambda=0.01; }
  decl alfa=vP[2][0];
  decl beta0=vP[3][0];
  decl beta1=vP[4][0];
  decl beta2=vP[5][0];
  decl beta3=vP[6][0];
  for(cont=0;cont<n;++cont){
    t=g_mY[cont][0];
    xx1=g_mY[cont][2];
    xx2=g_mY[cont][3];
    xx3=g_mY[cont][4];
    decl xbeta=beta0+beta1*xx1+beta2*xx2+beta3*xx3;
    decl k=exp(xbeta);
    decl mt=(t/lambda)^alfa;
    Ft=probgamma(mt,a,1);
    if(g_mY[cont][1]==1){
      vero[0][cont]=log(k)+log(alfa)-((alfa*a)*log(lambda))-loggamma(a)+(((alfa*a)-1)*log(t))-
      ((t/lambda)^alfa)-(k*Ft);
    }
    else{
      vero[0][cont]=-(k*Ft);
    }
  }
  adFunc[0]=double(vero*uns);
  return 1;
}
//***** Log-verossimilhanca Poisson GLL *****
lvGLL(const vP,const adFunc, const avScore,const amHessian){
  decl n=rows(g_mY);
  decl cont,t,xx0,xx1,xx2,xx3,xx4,xx5,xx6,xx7,xx8,ft,Ft,p0;
  decl uns=ones(n,1);

```

```

decl vero=zeros(1,n);
decl a=vP[0][0];
decl lambda=vP[1][0];
if(lambda<0.01){
lambda=0.01; }
decl alfa=vP[2][0];
decl beta0=vP[3][0];
decl beta1=vP[4][0];
decl beta2=vP[5][0];
decl beta3=vP[6][0];
for(cont=0;cont<n;++cont){
t=g_mY[cont][0];
xx1=g_mY[cont][2];
xx2=g_mY[cont][3];
xx3=g_mY[cont][4];
decl xbeta=beta0+beta1*xx1+beta2*xx2+beta3*xx3;
decl k=exp(xbeta);
decl mt=log(1+((t/lambda)^alfa));
Ft=probgamma(mt,a,1);
if(g_mY[cont][1]==1){
vero[0][cont]=log(k)+log(alfa)-(alfa*log(lambda))-loggamma(a)+((alfa-1)*log(t))-
(2*log(1+((t/lambda)^alfa)))+(a-1)*log(mt)-(k*Ft);
}
else{
vero[0][cont]=-(k*Ft);
}
}
adFunc[0]=double(vero*uns);
return 1;
}
main(){
g_mY = loadmat("glioma_todos.txt");
decl nc=rows(g_mY);
decl dfunc1,dfunc2,dfunc3,ir1,ir2,ir3,i,j,ep1,ep3,ep2;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
print("\n===== Log-gama-Weibull ===== \n");
decl vP1=<1;1;1;6.014;-0.034;0.006;-1.441>;
lvGW(vP1,&dfunc1,0,0);

```

```

decl mhess1=0.000001*unit(7);
MaxControl(-1,20);
ir1=MaxBFGS(lvGW,&vP1,&dfunc1,&mhess1,1);
if (ir1 == MAX_CONV){
if (Num2Derivative(lvGW,vP1,&mhess1)){
mhess1 = invertgen(-mhess1,0);
ep1=sqrt(diagonal(mhess1));
}
}
decl m=rows(vP1);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP1[i]-zi*ep1[i];
ic1[i][1]=vP1[i]+zi*ep1[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir1) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP1~ep1'*(2*(1-probn(fabs(vP1'./ep1))))~ic1);
print("\n===== Log-gama-logistica ===== \n");
decl vP2=<1;1;1;6.014;-0.034;0.006;-1.441>;
lvGLL(vP2,&dfunc2,0,0);
decl mhess2=0.000001*unit(7);
ir2=MaxBFGS(lvGLL,&vP2,&dfunc2,&mhess2,1);
if (ir2 == MAX_CONV){
if (Num2Derivative(lvGLL,vP2,&mhess2)){
mhess2 = invertgen(-mhess2,0);
ep2=sqrt(diagonal(mhess2));
}
}
decl ic2=zeros(m,2);
for(i=0;i<m;++i){
ic2[i][0]=vP2[i]-zi*ep2[i];
ic2[i][1]=vP2[i]+zi*ep2[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir2) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP2~ep2'*(2*(1-probn(fabs(vP2'./ep2))))~ic2);
}

```


A.5. Estimação dos parâmetros do modelo de longa-duração geométrica gama-Weibull e geométrica gama-log-logística

```

/*****
PROGRAMA: GeometricaGDG_Glioma.ox
*****/

#include<oxstd.h> #include<oxdraw.h> #include<oxprob.h> #include<oxfloat.h>
#include<maximize.h> #include<simula.h> #pragma link("maximize.oxo")
static decl g_mX,g_mY,g_mU,g_mC,g_mD;
/*****      Log-verossimilhanca Geometrica GWeibull      *****/
lvGW(const vP,const adFunc, const avScore,const amHessian)
decl n=rows(g_mY);
decl cont,t,xx0,xx1,xx2,xx3,xx4,xx5,xx6,xx7,xx8,ft,Ft,p0;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0][0];
decl lambda=vP[1][0];
if(lambda<0.01){
lambda=0.01; }
decl alfa=vP[2][0];
decl beta0=vP[3][0];
decl beta1=vP[4][0];
decl beta2=vP[5][0];
decl beta3=vP[6][0];
for(cont=0;cont<n;++cont){
t=g_mY[cont][0];
xx1=g_mY[cont][2];
xx2=g_mY[cont][3];
xx3=g_mY[cont][4];
decl xbeta=beta0+beta1*xx1+beta2*xx2+beta3*xx3;
decl p=exp(xbeta)/(1+exp(xbeta));
decl mt=(t/lambda)^alfa;
Ft=probgamma(mt,a,1);
if(g_mY[cont][1]==1){
vero[0][cont]=log(p)+log(1-p)+log(alfa)-((alfa*a)*log(lambda))-loggamma(a)+(((alfa*a)-1)*log(t))-
((t/lambda)^alfa)-2*log(p+((1-p)*Ft));
}
else{
vero[0][cont]=log(p)-log(p+((1-p)*Ft));
}
}

```

```

}
adFunc[0]=double(vero*uns);
return 1;
}
//***** Log-verossimilhanca Geometrica GLL *****
lvGLL(const vP,const adFunc, const avScore,const amHessian){
decl n=rows(g_mY);
decl cont,t,xx0,xx1,xx2,xx3,xx4,xx5,xx6,xx7,xx8,ft,Ft,p0;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl a=vP[0][0];
decl lambda=vP[1][0];
if(lambda<0.01)
lambda=0.01; }
decl alfa=vP[2][0];
decl beta0=vP[3][0];
decl beta1=vP[4][0];
decl beta2=vP[5][0];
decl beta3=vP[6][0];
for(cont=0;cont<n;++cont)
t=g_mY[cont][0];
xx1=g_mY[cont][2];
xx2=g_mY[cont][3];
xx3=g_mY[cont][4];
decl xbeta=beta0+beta1*xx1+beta2*xx2+beta3*xx3;
decl p=exp(xbeta)/(1+exp(xbeta));
decl mt=log(1+((t/lambda)^alfa));
Ft=probgamma(mt,a,1);
if(g_mY[cont][1]==1){
vero[0][cont]=log(p)+log(1-p)+log(alfa)-(alfa*log(lambda))-loggamma(a)+((alfa-1)*log(t))-
(2*log(1+((t/lambda)^alfa)))+((a-1)*log(mt))-2*log(p+((1-p)*Ft));
}
else{
vero[0][cont]=log(p)-log(p+((1-p)*Ft));
}
}
adFunc[0]=double(vero*uns);
return 1;
}

```

```

main(){
g_mY = loadmat("glioma.txt");
decl nc=rows(g_mY);
decl dfunc1,dfunc2,dfunc3,ir1,ir2,ir3,i,j,ep1,ep3,ep2;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
print("\n===== Log-gama-logistica ===== \n");
decl vP1=<1;1;1;6.014;-0.034;0.006;-1.441>;
lvGW(vP1,&dfunc1,0,0);
decl mhess1=0.000001*unit(7);
MaxControl(-1,20); //MaxControlEps(0.0001,0.01);
ir1=MaxBFGS(lvGW,&vP1,&dfunc1,&mhess1,1);
if (ir1 == MAX_CONV){
if (Num2Derivative(lvGW,vP1,&mhess1)){
mhess1 = invertgen(-mhess1,0);
ep1=sqrt(diagonal(mhess1));
}
}
decl m=rows(vP1);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP1[i]-zi*ep1[i];
ic1[i][1]=vP1[i]+zi*ep1[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir1) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP1~ep1'~(2*(1-probn(fabs(vP1'./ep1))))'~ic1);
print("\n===== Log-gama-logistica ===== \n");
decl vP2=<0.5;0.5;0.5;6.014;-0.034;0.006;-1.441>;
lvGLL(vP2,&dfunc2,0,0);
decl mhess2=0.000001*unit(7);
ir2=MaxBFGS(lvGLL,&vP2,&dfunc2,&mhess2,1);
if (ir2 == MAX_CONV){
if (Num2Derivative(lvGLL,vP2,&mhess2)){
mhess2 = invertgen(-mhess2,0);
ep2=sqrt(diagonal(mhess2));
}
}
decl ic2=zeros(m,2);

```

```

for(i=0;i<m;++i){
ic2[i][0]=vP2[i]-zi*ep2[i];
ic2[i][1]=vP2[i]+zi*ep2[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir2) );
print("\n Estimativa   Erro padrao   p_valor   IC(95)", "%13.5f",
vP2~ep2'~(2*(1-probn(fabs(vP2'./ep2)))')~ic2);
}

```

A.6. Estimação dos parâmetros do modelo de regressão com efeito aleatório log-gama-Weibull

```

/*****
PROGRAMA: EALGW_Cattle.ox
*****/

#include<oxstd.h>  #include<oxdraw.h>  #include<oxprob.h>  #include<oxfloat.h>
#include<maximize.h>  #include<simula.h>  #pragma link("maximize.oxo")
static decl g_mY,g_mH,n,ngh;
log_vero(const vP,const adFunc, const avScore,const amHessian){
decl j,k,i,integrando,lj,yij,x,w,gk,integral,vero,xbeta,alfa,sy,fy,z;
decl phi=vP[0][0];
decl sig1=vP[1][0];
decl sig2=vP[2][0];
decl b0=vP[3][0];
decl b1=vP[4][0];
decl b2=vP[5][0];
decl mu=0;
decl N=4;
integrando=zeros(N,ngh);
for(j=0; j<N;++j){
for(k=0; k<ngh;++k){
lj=zeros(n,1);
x=g_mGH[k][1];
for(i=0; i<n; ++i){
if(g_mY[i][1]==j+1){
yij=log(g_mY[i][2]);
xbeta=b0+(b1*g_mY[i][4])+(b2*g_mY[i][5]);
z=(yij-xbeta-x)/sig1;
if(g_mY[i][3]==1){
lj[i]=-log(sig1)+(z-exp(z));

```

```

}
else{
lj[i]=-exp(z);
}
}
else{
lj[i]=0;
}
w=g_mGH[k][0];
gk=(1/(sig2*gammafact(phi)))*exp((phi*((x-mu)/sig2))-exp((x-mu)/sig2));
integrando[j][k]=w*(gk/exp(-(x^2)))*exp(sumc(lj));
}
}
integral=sumr(integrando);
}
vero=log(integral);
adFunc[0]=double(sumc(vero));
return 1;
}
main(){
g_mY=loadmat("cattle.txt");
n=rows(g_mY);
g_mGH=loadmat("gaussHermite20.txt");
ngh=rows(g_mGH);
decl dfunc,ir,i,j,ep;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
decl vP=<1;0.16263;0.5;5.27975;-0.17256;-0.01229>;
log_vero(vP,&dfunc,0,0);
decl mhess1=0.01*unit(6);
MaxControl(-1,20);
ir=MaxBFGS(log_vero,&vP,&dfunc,&mhess1,1);
Num2Derivative(log_vero,vP,&mhess1);
ep= sqrt(diagonal(-1/mhess1));
decl m=rows(vP);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP[i]-zi*ep[i];
ic1[i][1]=vP[i]+zi*ep[i];
}
}

```

```

}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir) );
print("\n Estimativa   Erro padrao   p_valor   IC(95)", "%13.5f",
vP~ep'~(2*(1-probn(fabs(vP'./ep)))')~ic);
}

```

A.7. Estimação dos parâmetros do modelo de regressão com efeito aleatório log-gama-logística

```

/*****
PROGRAMA: EALogGLL_Cattle.ox
*****/

#include<oxstd.h>  #include<oxdraw.h>  #include<oxprob.h>  #include<oxfloat.h>
#include<maximize.h>  #include<simula.h>  #pragma link("maximize.oxo")
static decl g_mY,g_mH,n,ngh;
log_vero(const vP,const adFunc, const avScore,const amHessian){
decl j,k,i,integrando,lj,yij,x,w,gk,integral,vero,xbeta,alfa,sy,fy,z;
decl phi=vP[0][0];
decl sig1=vP[1][0];
decl sig2=vP[2][0];
decl b0=vP[3][0];
decl b1=vP[4][0];
decl b2=vP[5][0];
decl mu=0;
decl N=4;
integrando=zeros(N,ngh);
for(j=0; j<N;++j){
for(k=0; k<nggh;++k){
lj=zeros(n,1);
x=g_mGH[k][1];
for(i=0; i<n; ++i){
if(g_mY[i][1]==j+1){
yij=log(g_mY[i][2]);
xbeta=b0+(b1*g_mY[i][4])+(b2*g_mY[i][5]);
z=(yij-xbeta-x)/sig1;
if(g_mY[i][3]==1){
lj[i]=-log(sig1)+(z-exp(z));
}
else{
lj[i]=-exp(z);

```

```

}
}
else{
lj[i]=0;
}
w=g_mGH[k][0];
gk=(1/(sig2*gammafact(phi)))*exp((x-mu)/sig2)*((1+exp((x-mu)/sig2))^(-2))*
((log(1+exp((x-mu)/sig2)))^(phi-1));
integrando[j][k]=w*(gk/exp(-(x^2)))*exp(sumc(lj));
}
}
integral=sumr(integrando);
}
vero=log(integral);
adFunc[0]=double(sumc(vero));
return 1;
}
main(){
g_mY=loadmat("cattle.txt");
n=rows(g_mY);
g_mGH=loadmat("gaussHermite20.txt");
ngh=rows(g_mGH);
decl dfunc,ir,i,j,ep;
decl alfa=0.05;
decl zi=quann(1-(alfa/2));
decl vP=<1;0.16263;1;5.27975;-0.17256;-0.01229>; log_vero(vP,&dfunc,0,0);
decl mhess1=0.01*unit(6);
MaxControl(-1,20);
ir=MaxBFGS(log_vero,&vP,&dfunc,&mhess1,1);
Num2Derivative(log_vero,vP,&mhess1);
ep= sqrt(diagonal(-1/mhess1));
decl m=rows(vP);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP[i]-zi*ep[i];
ic1[i][1]=vP[i]+zi*ep[i];}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir) );
print("\n Estimativa Erro padrao p_valor IC(95)", "%13.5f",
vP~ep'~(2*(1-probn(fabs(vP'./ep))))'~ic);}

```

A.8. Estimação dos parâmetros do modelo GDG inflacionado de zeros

```

/*****
PROGRAMA: ZIgammaG.ox
*****/

#include<oxstd.h>  #include<oxdraw.h>  #include<oxprob.h>  #include<oxfloat.h>
#include<maximize.h>  #include<simula.h>  #pragma link("maximize.oxo")

static decl n, g_mY;

//*****      Log-verossimilhanca ZIGWeibull      *****/

log_veroGW(const vP,const adFunc, const avScore,const amHessian){
decl cont,y,xx0,xx1,xx2,xx3,xx4,xx5,xx6,xx7,xx8,ind;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl phi=vP[0][0];
decl a=vP[1][0];
decl lambda=vP[2][0];
decl b0=vP[3][0];
decl b1=vP[4][0]; //especie
decl b2=vP[5][0]; //DAP
decl b3=vP[6][0]; //praga
decl b4=vP[7][0]; //QF
for(cont=0;cont<n;++cont){
y=g_mY[cont][2];
xx1=g_mY[cont][4];
xx2=g_mY[cont][5];
xx3=g_mY[cont][7];
xx4=g_mY[cont][8];
decl xbeta=b0+(b1*xx1)+(b2*xx2)+(b3*xx3)+(b4*xx4);
decl pi=exp(xbeta)/(1+exp(xbeta));
decl lfy=log(a)-((phi*a)*log(lambda))-loggamma(phi)+(((phi*a)-1)*log(y))-((y/lambda)^a);
if(g_mY[cont][3]==1){
vero[0][cont]=log(pi);
}
else{
vero[0][cont]=log(1-pi)+lfy;
}
}
adFunc[0]=double(vero*uns);
return 1;
}

```



```

*****      Log-verossimilhanca ZIGLL      *****

log_veroGLL(const vP,const adFunc, const avScore,const amHessian){
decl cont,y,xx0,xx1,xx2,xx3,xx4,xx5,xx6,xx7,xx8,ind;
decl uns=ones(n,1);
decl vero=zeros(1,n);
decl phi=vP[0][0];
decl a=vP[1][0];
decl lambda=vP[2][0];
decl b0=vP[3][0];
decl b1=vP[4][0]; //especie
decl b2=vP[5][0]; //DAP
decl b3=vP[6][0]; //praga
decl b4=vP[7][0]; //QF
for(cont=0;cont<n;++cont){
y=g_mY[cont][2];
xx1=g_mY[cont][4];
xx2=g_mY[cont][5];
xx3=g_mY[cont][7];
xx4=g_mY[cont][8];
decl xbeta=b0+(b1*xx1)+(b2*xx2)+(b3*xx3)+(b4*xx4);
decl z=1+((y/lambda)^a);
decl pi=exp(xbeta)/(1+exp(xbeta));
decl lfy=log(a)-loggamma(phi)-(a*log(lambda))+((a-1)*log(y))-2*log(z)+
((phi-1)*log(log(z+1e-10)));
if(g_mY[cont][3]==1){
vero[0][cont]=log(pi);
}
else{
vero[0][cont]=log(1-pi)+lfy;
}
}
adFunc[0]=double(vero*uns);
return 1;
}

main(){
g_mY=loadmat("DadosCarine_new.txt");
n=rows(g_mY);
decl dfunc1,dfunc2,ir1,ir2,i,j,ep1,ep2;
decl alfa=0.05;

```

```

decl zi=quann(1-(alfa/2));
print("\n=====  ZI gama-Weibull  ===== \n");
decl vP1=<1;1;1;-0.5;-0.5;-0.5;-0.5>;
log_veroGW(vP1,&dfunc1,0,0);
decl mhess1=0.000001*unit(8);
MaxControl(-1,20);
ir1=MaxBFGS(log_veroGW,&vP1,&dfunc1,&mhess1,1);
if (ir1 == MAX_CONV){
Num2Derivative(log_veroGW,vP1,&mhess1);
ep1=sqrt(diagonal(-1/mhess1));
}
decl m=rows(vP1);
decl ic1=zeros(m,2);
for(i=0;i<m;++i){
ic1[i][0]=vP1[i]-zi*ep1[i];
ic1[i][1]=vP1[i]+zi*ep1[i];
}
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir1) );
print("\n Estimativa  Erro padrao  p_valor  IC(95)", "%13.5f",
vP1~ep1'~(2*(1-probn(fabs(vP1'./ep1)))')~ic1);
print("\n=====  ZI gama-log-logistica  ===== \n");
decl vP2=<1;1;1;-0.5;-0.5;-0.5;-0.5>;
log_veroGLL(vP2,&dfunc2,0,0);
decl mhess2=0.000001*unit(8);
ir2=MaxBFGS(log_veroGLL,&vP2,&dfunc2,&mhess2,1);
if (ir2 == MAX_CONV){
Num2Derivative(log_veroGLL,vP2,&mhess2);
ep2=sqrt(diagonal(-1/mhess2));
}
decl ic2=zeros(m,2);
for(i=0;i<m;++i){
ic2[i][0]=vP2[i]-zi*ep2[i];
ic2[i][1]=vP2[i]+zi*ep2[i]; }
print("\n CONVERGENCE: ",MaxConvergenceMsg(ir2) );
print("\n Estimativa  Erro padrao  p_valor  IC(95)", "%13.5f",
vP2~ep2'~(2*(1-probn(fabs(vP2'./ep2)))')~ic2);}

```