Universidade de São Paulo

Escola Superior de Agricultura "Luiz de Queiroz"

Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência

Juliana Betini Fachini

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

Piracicaba

2011

Juliana Betini Fachini

Bacharel em Estatística

Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência

Orientador:

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

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

Piracicaba

2011

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - ESALQ/USP

Fachini, Juliana Betini
Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência / Juliana Betini Fachini. - - Piracicaba, 2011.
140 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2011.

1. Análise de sobrevivência 2. Dados censurados 3. Modelos matemáticos 4. Regressão linear 5. Verosimilhança I. Título

CDD 519.536
F139m

Dedicatória

Aos meus pais,

Edson e Maria de Fátima, pelo exemplo de ser humano que são e por tudo que me ensinaram.

Aos meus avós,

Ezio, Aparecida, Serafim e Aparecida (in memoriam), por acreditarem incondicionalmente em mim como profissional e pessoa.

AGRADECIMENTOS

A Deus, pela força para a realização deste trabalho.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pela concessão de bolsa de estudos.

Ao Prof. Dr. Edwin Mosisés Marcos Ortega pela orientação à elaboração deste trabalho.

Aos Professores Dra. Clarice Garcia Borges Demétrio, Dra. Roseli Aparecida Leandro, Dra. Sônia Maria De Stefano Piedade, Dr. Décio Barbin e Dr. Carlos Tadeu dos Santos Dias por todos os ensinamentos e amizade.

Aos meus amados pais Edson e Maria de Fátima por todo apoio, amor, incentivo, alegrias e paciência com a minha ausência neste período.

Ao Eduardo Monteiro de Castro Gomes por todo apoio, incentivo, paciência, contribuições estatísticas e principalmente por todo amor.

A minha princesa e afilhada Letícia e ao meu príncipe Guilherme por encherem os meus dias de alegria e vida.

Ao meu irmão Ricardo, a minha cunhada Érica e a minha irmã Lívia pelo carinho e alegrias.

A minha prima Andréia por todo o carinho, comidas deliciosas e por me dar a alegria de conviver com os primos Bruno e Bianca.

Ao Vicente, Ana Maria, Mariana e Rodrido por todo carinho, incentivo e ótimos momentos de alegrias.

A minha querida amiga e irmã Elizabeth Francisco Alves pela eterna amizade, carinho, constantes apois e incentivos e por fazer parte de todos os momentos da minha vida.

As minhas queridas amigas Andréia da Silva Meyer e Elizabeth Hashimoto por dividirem as emoções diárias, amizade incondicional e suporte emocional.

As minhas amigas Fernanda e Joseane pelo apoio, constantes incentivos e momentos de muitos risos.

Aos professores do Departamento de Ciência Exatas da ESALQ/USP por todos os ensinamentos.

Aos funcionários do Departamento de Ciência Exatas da ESALQ/USP, em especial a Solange Sabadin, Luciane Brajão, Jorge Wiendl e Eduardo Bonilha, pelos auxílios permanentes.

Aos amigos dos cursos de mestrado e doutorado do Departamento de Ciência Exatas da ESALQ/ USP, em especial a Afrânio, César, Giovana, Lucimary, Vanderly, Glaucia, Simone, Renata e Marina a atenção, a compreensão e a amizade.

A Escola Superior de Agricultura "Luiz de Queiroz", pela oportunidade de realização do meu doutorado.

A todos que de alguma forma contribuíram para a realização deste trabalho.

SUMÁRIO

RESUMO

Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência

Neste trabalho são reunidos diferentes modelos e técnicas para representar situações experimentais ou observacionais de análise de sobrevivência. Para modelar respostas bivariadas e covariáveis foi proposto o modelo de regressão Kumaraswamy-Weibull bivariado. A presença de indivíduos curados foi considerada sob duas diferentes abordagens, originando o modelo de regressão com fração de cura para dados bivariados por meio de cópulas e o modelo de regressão log-linear bivariado com fração de cura. Os parâmetros dos modelos foram estimados pelo método de máxima verossimilhança sujeito a restrição nos parâmetros por meio da função barreira adaptada. Adaptou-se uma análise de sensibilidade de forma a considerar as metodologias de Influência Global, Influência Local e Influência Local Total para verificar vários aspectos que envolvem a formulação e ajuste dos modelos propostos. Utilizou-se um conjunto de dados de insuficiência renal e retinopatia diabética são utilizados para exemplificar a aplicação dos modelos propostos.

Palavras-chave: Modelos de regressão; Dados bivariados e censurados; Verossimilhança sujeita a restrição nos parâmetros; Fração de cura; Análise de sensibilidade

ABSTRACT

Models with and without fraction of cure for bivariate data in survival analysis

This work brought together different models and techniques to represent experimental or observational situations in survival analysis. To model bivariate responses and covariates was proposed Kumaraswamy Weibull bivariate regression model. The presence of cured individuals was considered under two different approaches originating the regression model with a cured fraction for bivariate data through copulas and the log-linear bivariate regression model with cured fraction. The parameters of the models were estimated by maximum likelihood method subject to the restriction on the parameters through the adapted barrier function. A sensitivity analysis was adapted considering the methodologies of Global Influence, Local Influence and Total Local Influence to check various aspects of the formulation and adjustment of the models proposed. Data set of renal failure and diabetic retinopathy are used to exemplify the application of the proposed models.

Keywords: Regression models; Bivariate data and censored; Cured fraction; Likelihood subject to restriction on the parameters; Sensitivity analysis

1 INTRODUÇÃO

Modelos estatísticos são ferramentas utilizadas nas mais diversas áres do conhecimento com o objetivo de descrever e explicar diferentes fenômenos por meio de linguagem matemática. A pesquisa e o desenvolvimento desses modelos busca torná-los mais elaborados de forma a considerar nuances e particularidades das diferentes situações em estudo.

Nas áreas das Ciências Biológicas e das Engenharias é comum se estudar o tempo transcorrido até a ocorrência de determinado fenômeno de interesse. A área da Estatística que reune técnicas e métodos para fornecer resposta a esse tipo de indagação é chamada de análise de sobrevivência. Neste trabalho serão abordadas três particularidades que ocorrem com frequência em estudos de análise de sobrevivência: observações censuradas, respostas multivariadas e presença de indivíduos curados ou não suscetíveis.

São chamadas observações censuradas as observações parciais do tempo até a ocorrência do evento para alguns indivíduos em estudo. Essas observações parciais podem ocorrer por diferentes motivos, como o término do período de observação, a saída de indivíduo do estudo por diferentes razões, a impossibilidade de saber quando deve ser o início da contagem do tempo ou, ainda, a ausência do exato momento em que ocorreu um determinado evento em dado intervalo de tempo.

Dentre os problemas estudados em análise de sobrevivência, muitos são os que envolvem vários fatores de forma simultânea. Esses fatores podem ser tratados como respostas ou covariáveis. As respostas são os interesses principais do estudo e devem referir-se ao tempo transcorrido até a ocorrência dos diferentes eventos de interesse ou de recorrência de um mesmo evento. As covariáveis são as características dos indivíduos em estudo que podem influir nas respostas. Neste trabalho serão considerados os modelos de regressão bivariados para relacionar duas variáveis resposta associadas a covariáveis que possam ter efeito sobre as respostas. Nessa linha de pesquisa, destacam-se alguns trabalhos como Gumbel (1960), Freund (1961) e Marshall e Olkin (1967) que estudam a distribuição exponencial bivariada. Além desses, Moeschberger (1974) deriva um modelo Weibull para riscos competitivos, Ryu (1993) propõe um modelo exponencial bivariado, Oliveira (2001) estuda a Extensão da Distribuição Exponencial Bivariada na abordagem Bayes-Empírica, Tarumoto (2001) desenvolve um modelo de riscos competitivos baseado no modelo proposto por Ryu (1993) e Cordeiro et

al.(2010) propõem a distribuição Kumaraswamy generalizada bivariada.

Com base nesses estudos, em especial em Cordeiro et al.(2010), este trabalho propõe um modelo de regressão Kumaraswamy Weibull bivariado com o objetivo de relacionar duas variáveis resposta e obter por meio dos parâmetros de regressão do modelo essa relação e sua associação com as covariáveis, assumindo a distribuição Kumaraswamy Weibull bivariado para as variáveis resposta.

Com relação à terceira particularidade, observa-se que, em estudos experimentais, determinados indivíduos não experimentarão alguns tipos de eventos de interesse, ainda que sejam observados por longos períodos. Esses indivíduos podem ter sido curados ou serem imunes ao evento em estudo. Ao considerar esse fato, Berkson e Gage (1952) propõem a construção de uma função de sobrevivência populacional na forma de mistura. Farewell (1982, 1986), Goldman (1984), Greenhous e Wolfe (1984), Halpern e Brown (1987) e Maller e Zhou (1996), por sua vez, apresentam uma discussão sobre os estimadores de máxima verossimilhança do modelo proposto por Berkson e Gage (1952). Cancho (1999) utiliza o método Bayesiano para estimar os parâmetros do modelo com fração de cura, Ortega et al.(2009a) propõem um modelo de regressão log-gama generalizado com fração de cura e Ortega et al. (2009b) consideram uma análise de Influência Local e resíduos para o modelo de mistura log-gamma generalizado com covariáveis.

Ao considerar duas variáveis resposta, Chatterjee e Shih (2001), Wienke et al. (2003) e Wienke et al. (2006) utilizam a metodologia de cópulas para descrever a função de sobrevivência populacional bivariada na forma de mistura. Para dar continuidade a esses estudos, este trabalho propõe-se a modelar o efeito das variáveis regressoras e sua associação com as variáveis resposta em um modelo de regressão bivariado com fração de cura por meio de cópulas.

Alternativamente à metodologia de fração de cura introduzida por Berkson e Gage (1952), Yakovlev et al. (1993) propõem uma nova classe de modelos de tempo de promoção com a estrutura de riscos competitivos, cuja modelagem é introduzida no contexto biológico. Esse modelo também foi amplamente discutido por Yakovlev et al. (1994), Asselain et al. (1996), Yakovlev e Tsodikov (1996) e Chen et al. (1999) desenvolvem a formulação Bayesiana. Mizoi (2004) aplica uma metodologia de Influência Local em modelos de sobre-

vivência com fração de cura, Cancho et al. (2009) realizam uma análise de resíduo e de Influência Local para o modelo de regressão log-Weibull-exponenciado com fração de cura, Ortega et al. (2009a) desenvolvem uma análise de sensibilidade e resíduo para o modelo de regressão log-gama generalizado com fração de cura. No contexto bivariado, Chen et al.(2002) propõem o modelo de tempo de promoção bivariado com fração de cura, de acordo com essa última abordagem, este trabalho, e considerando a presença de variáveis regressoras, propõe um modelo de regressão log-linear bivariado com fração de cura.

Os parâmetros dos modelos propostos foram estimados pelo método de máxima verossimilhança sujeito a restrição nos parâmetros por meio da função barreira adaptada. Na análise estatística de dados, uma das etapas posterior ao ajuste do modelo a um conjunto de dados, é verificar se as suposições do modelo são válidas e identificar características inesperadas nos dados que possam influênciar as conclusões obtidas. Para detectar observações influentes nos modelos propostos, estudou-se a metodologia de Influência Global (COOK, 1977), medidas de Influência Local baseadas em pequenas perturbações nos dados ou no modelo (COOK, 1986) e a medida de Influência Local Total (LESAFFRE; VERBEKE, 1998). Como essas medidas de análise de sensibilidade não consideram parâmetros com restrições lineares, utilizou-se neste trabalho, com base nos trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009), uma análise de sensibilidade para os modelos desenvolvidos que considera as restrições existentes em alguns parâmetros dos modelos.

O prsente estudo está organizado da seguinte forma: no Capítulo 2 é proposto o modelo de regressão Kumaraswamy Weibull bivariado. A seção 2.2 faz uma revisão da distribuição Kumaraswamy e sua generalização. A seção 2.3.1 aborda a distribuição Kumaraswamy generalizada para o caso bivariado. A seção 2.3.2 propõe o modelo de regressão Kumaraswamy bivariado. Na seção 2.4 é desenvolvida uma técnica de análise de sensiblidade para o modelo proposto na seção 2.3.2. Na seção 2.5 é realizado o ajuste do modelo proposto aos dados de insuficiência renal. Para concluir, a seção 2.6 relata as conclusões obtidas e as pesquisas futuras.

No Capítulo 3 é proposto o modelo de regressão bivariado com fração de cura por meio de cópulas. A seção 3.2 apresenta uma revisão do modelo com fração de cura para o caso univariado. Na seção 3.3 é descrito o modelo bivariado. A seção 3.4 apresenta

o modelo de regressão bivariado com fração de cura prosposto neste trabalho. A seção 3.5 abrange uma descrição da análise de sensibilidade do modelo proposto baseada nas teorias de Influência Global, Local e Local Total sob o enfoque da verossimilhança sujeita a restrições nos parâmetros. Na seção 3.6 é feita a aplicação do modelo proposto. Para finalizar, a seção 3.7 relata as principais conclusões e apresenta sugestões para dar continuidade a este trabalho.

Para finalizar esta pesquisa, o Capítulo 4 propõe o modelo de regressão log-linear bivariado com fração de cura. Na seção 4.2 é apresentada uma revisão do modelo com fração de cura univariada seguindo abordagem de Yakovlev et al.(1993). Na seção 4.3 introduz-se o modelo de tempo de promoção bivariado com fração de cura. Na seção 4.4 é proposto o modelo de regressão log-linear bivariado com fração de cura. A seção 4.5 abrange uma descrição da análise de sensibilidade do modelo proposto baseada nas teorias de Influência Global, Local e Local Total. Na seção 4.6 aplica-se o modelo proposto aos dados de retinopotia diabética. Para concluir, a seção 4.7 relata as principais conclusões e sugere possíveis trabalhos relacionados ao modelo proposto.

Referências

BERKSON, J.; GAGE, R.P. Survival curve for cancer patients follwing treatment. **Journal of the American Statistical Association**, Alexandria, v. 47, p. 501-515, 1952.

CASTRO, M.; CANCHO, V. G.; RODRIGUES, J. A bayesian long-term survival model parametrized in the cured fraction, **Biometrical Journal**, Weinheim, v. 51, n. 3, p. 443-455, 2009.

CHATTERJEE N., SHIH J. A bivariate cure-mixture approach for modeling familial association in diseases, **Biometrics**, Washington, v. 57, p. 779-786, 2001.

CHEN, M. H.; IBRAHIM, J; SINHA, D. A new bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Alexandria, v. 94, p. 909-919, 1999.

CHEN, M. H.; IBRAHIM, J; SINHA, D. Bayesian inference for multivariate survival data with a surviving fraction. **Journal of Multivariate Analysis**, New York, v. 80, p. 101-126, 2002.

CLAYTON, D. G. A model for association in bivariate life-tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, London, v. 65, p. 141-151, 1978.

CORDEIRO, G.M; Castro, M. A new family of generalized distributions. **Journal of Statistical Computation and Simulation**, New York, v.0, p. 1-17, 2009.

CORDEIRO, G.M; Ortega, E.M.M., Nadarajah, S. The Kumaraswamy Weibull distributions with application to failure data. **Journal of Franklin Institute**, New York, v.347, p. 1399-1429, 2010.

COOK, R.D. Detection of influential observations in linear regression. **Technometrics**, Alexandria, v. 19, p. 15-18, 1977.

COOK, R.D. Assement of local influence (with discussion). **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 48, n. 2, p. 133-169, 1986.

FAREWELL, V. T. The use mixture models for the analysis of survival data with log-term survivors. **Biometrics**, Washington, v. 38, p. 43-46, 1982.

FREUND, J. E. Bivariate Extension of the Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 56, p. 971-977, 1961.

GOLDMAN, A. Survivorship analysis when cure is a possibility: a Monte Carlo study. **Statistics in Medicine**, Chichester, v. 3, p. 153-163, 1984.

GOMES, E. M. C. **Análise de Sensibilidade e resíduos em modelos de regressão com respostas bivariadas por meio de cópulas.** 2007. 103p. Dissertação (Mestre em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

GREENHOUSE, J.B.; WOLFE, R.A. A competing risks derivation of a mixture model for the analysis of survival data. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 13, p. 3133-3154, 1984.

GU, H; FUNG, W.K. Local influence for the restricted likelihood with applications. **Sankhya: The Indian Journal of Statistical**, Indian, v. 63, pt. 2, p. 250-259, 2001.

GUMBEL, E. J. Bivariate Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 55, p. 698-707, 1960.

HALPERN, J.; BROWN, B. Cure rate models: Power of the log-rank and generalized Wilcoxon tests. **Statistics in Medicine**, Chichester, v. 6, p. 483-489, 1987.

HE, W.; LAWLESS, J. F. Bivariate location-scale models for regression analysis, with applications to lifetime data. **Journal of the Royal Statistical Society**, London, v. 67, n. 1, p. 63-78, 2005.

HOUGAARD, P. Fitting a multivariate failure time distribution. **IEEE Transactions on Reliability**, New York, v. 38, p. 444-448, 1989.

JOHNSON, R. A.; EVANS, J. W.; GREEN, D. W. Some bivariate distributions for modeling the strength properties of lumber. **United States Department of Agriculture**, Washington, FPL-RP-575, 1999.

KWAN, C. W; FUNG, W. K. Assessing local influence for specific restricted likelihood: application to factor analysis. **Psychometrika**, New York, v. 63, n. 1, p. 35-46, 1998.

LANGE, K. **Numerical analysis for statisticians**. New York: Springer, 1999. 356 p.

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

MALLER, R.; ZHOU, X. **Survival Analysis with Long-Term Survivors**. New York: Wiley, 1996. 278 p.

MARSHALL, A. W.; OLKIN, I. A Multivariate Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 62, p. 30-44, 1967.

MIZOI, M. F. **Influência local em modelos de sobrevivência com fração de cura.** 2004. 95p. Tese (Doutorado em Estatística)-Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2004.

MOESCHBERGER, M. L. Life tests under dependent competing causes of failure. **Technometrics**, Alexandria, v. 16, p. 39-47, 1974.

NÚÑEZ, J. S. R. **Modelagem Bayesiana para dados de sobrevivência bivariados através de cópulas.** 2005. 101p. Tese (Doutorado em Estatística)- Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2005.

OLIVEIRA, L. P. **Estudo da extensão do modelo bivariado exponencial de Marshall e Olkin para dados de confiabilidade.** 2001. 166p. Dissertação (Mestrado em Estatística)- Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 2001.

ORTEGA, E. M. M. ; CANCHO, V. G.; PAULA, G. A. Generalized log-gamma regression models with cure fraction. **Lifetime Data Analysis**, Boston, v. 15, p. 79-106, 2009a.

ORTEGA, E. M. M. ; RIZZATO, F. B.; DEMÉTRIO, C. G. B. The generalized log-gamma mixture model with covariates: local influence and residual analysis. **Statistical Methods and Application**, Heidelberg, v. 18, n. 3, p. 305-331, 2009b.

PAULA, G.; CYSNEIROS, F. J. A. Local influence under parameter constraints. **Communications in Statistics: Theory and Methods**, New York, v.88, p. 1-23, 2009.

RYU, K. An Extention of Marshall and Olkin's Bivariate Exponential Distribution. **Journal of the American Statistical Association**, Alexandria, v. 88, p. 1458-1465, 1993.

SHIH, J. H.; LOUIS, T. A. Inferences on the association parameter in copula models for bivariate survival data. **Biometrics**, Washington, v. 51, p. 1384-1399, 1995.

TARUMOTO, M. H. **Um modelo Weibull bivariado para riscos competitivos.** 2001. 154p. Tese (Doutorado em Matemática Aplicada)- Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 2001.

TIBALDI, F. S. **Modeling of Correlated Data and Multivariate Survival Data.** 2004. 160p. Tese de Doutorado, Universidade de Hasselt, Hasselt, 2004.

WIENKE A.; LICHTENSTEIN L.; YASHIN A.I. A bivariate frailty model with a cure fraction for modeling familial correlations in diseases, **Biometrics**, Washington, v.59, p.1178-1183, 2003.

WIENKE A.; LOCATELLI I.; YASHIN A.I. The modelling of a cure fraction in bivariate time-to-event data, **Austrian Journal of Statistics**, Austrian, v.35, p.67-76, 2006.

YAKOVLEV, A.; ASSELAIN, B.; BARDOU, V., FOURQUET, A. HOANG, T. ROCHEFEDIERE; TSODIKOV,A.D. A Stochastic models of tumor latency and their biosta-tistical applications. **Biometrie et analyse de Donnes Spatio-Temporelles**, Paris, v. 12, p. 66-82, 1993.

YAKOVLEV, A. Letter to the Editor. **Statistics in Medicine**, Chichester, v. 13, p. 983-986, 1994.

YAKOVLEV, A.; TSODIKOV,A.D. **Stochastic models of tumor latency and their biostatistical applications**, New Jersey: World Scientific, 1996. 288 p.

2 MODELO DE REGRESSÃO KUMARASWAMY WEIBULL BIVARIADO

Resumo

Neste capítulo foi proposto o modelo de regressão Kumaraswamy Weibull bivariado com base no trabalho de Cordeiro et al. (2010), com o objetivo de verificar uma associação existente entre os tempos de falha e incluir variáveis regressoras. Para estimar os parâmetros do modelo foi implementado o método da função barreira adaptada (LANGE, 1999). Uma análise de Influência Global, Influência Local e Influência Local Total foi desenvolvida para o modelo proposto. Para finalizar, um conjunto de dados de insuficiência renal foi analisado considerando o modelo de regressão Kumaraswamy Weibull bivariado.

Palavras-chave: Distribuição Kumaraswamy generalizada; Dados bivariados e censurados; Modelos de regressão; Distribuição Kumaraswamy bivariada; Verossimilhança sujeita a restrição nos parâmetros; Análise de sensibilidade

Abstract

In this chapter we proposed a Kumaraswamy Weibull bivariate regression model based on the work of Cordeiro et al. (2010), with the objective to verify an association between the failure times and include regressive variables. To estimate the parameters of the model was implemented adapted barrier function method (LANGE, 1999). An analysis of Global Influence, Local Influence and Total Local Influence were developed for the proposed model. Finally, a data set of renal failure was analyzed considering the Kumaraswamy Weibull bivariate regression model.

Keywords: Bivariate data and censored; Regression models; Sensitivity analysis; Generalized Kumaraswamy Distribution; Likelihood subject to restriction on the parameters; Kumaraswamy bivariate distribution

2.1 Introdução

Aplicações estatísticas que têm como variável resposta o tempo até a ocorrência de um evento de interesse usam, em geral, as distribuições Exponencial, Weibull, Gama, Valor

extremo, Log-normal e Log-logística para modelar os dados. Contudo, há um intenso crescimento de trabalhos propondo novas distribuições. Entre elas, destacam-se as distribuições Weibull exponenciada, Weibull modificada generalizada, Beta Weibull modificada, Beta generalizada e Kumaraswamy-Weibull, sendo a última a que receberá maior atenção neste capítulo e será denotada por K-Weibull.

Em 1980, Kumaraswamy propôs uma nova distribuição com dois parâmetros que recebe seu nome. O desempenho dessa distribuição Kumaraswamy é comparado com outras distribuições por Jones (2009). Recentemente, Cordeiro e Castro (2010) apresentaram uma extensão da distribuição Kumaraswamy, conhecida como distribuição Kumaraswamy generalizada, denotada por K-G.

Ao considerar mais de uma variável resposta no experimento, Cordeiro et al. (2010) propõem, para o caso bivariado, a distribuição Kumaraswamy generalizada bivariada, com enfoque maior na distribuição K-Weibull bivariada tipo I, construída a partir da distribuição Weibull bivariada de Hougaard (1986), e da distribuição Kumaraswamy Weibull bivariada tipo II, formulada com base na distribuição Weibull bivariada proposta por Lu e Bhattacharyya (1990).

Na prática observa-se que as variáveis resposta de cada observação estão associadas a variáveis regressoras. Assim, a análise estatística dos dados procura trabalhar com modelos que consideram todas as informações obtidas a partir das observações. Por essa razão, este capítulo tem como objetivo principal extender a distribuição K-Weibull bivariada tipo I para inserir covariáveis por meio do parâmetro de escala, dando origem ao modelo de regressão Kumaraswamy Weibull bivariado.

Ao considerar métodos de estimação para os parâmetros do modelo proposto é necessário selecionar métodos que levem em conta as restrições existentes no espaço paramétrico associado à distribuição. Neste trabalho utiliza-se o método de máxima verossimilhança sujeito a restrição nos parâmetros. Sua implementação foi baseada na função barreira adaptada, que é uma combinação do método barreira com o algoritmo EM. Para maiores detalhes sobre esse método ver, por exemplo, Lange (1999).

Uma importante etapa após o ajuste de um modelo a um conjunto de dados consiste em verificar a adequabilidade das suposições feitas ao modelo. Para tanto, o se-

gundo objetivo deste capítulo consiste em desenvolver e discutir técnicas de diagnóstico para o modelo de regressão Kumaraswamy Weibull bivariado. Algumas das técnicas mais comumente utilizadas são Influência Global proposta por Cook (1977), Influência Local introduzida por Cook (1986) e Influência Local Total desenvolvida por Lesaffre e Verbeke (1998). Vale notar que essas técnicas de diagnóstico não consideram parâmetros com restrição no espaço paramétrico. Por esse motivo, para propor uma análise de sensibilidade ao modelo desenvolvido neste capítulo são utilizados como referências os trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009), que utilizam Influência Local na estrutura de verossimilhança restrita.

Este trabalho está organizado da seguinte forma: a seção 2.2 consiste em uma revisão da distribuição Kumaraswamy e sua generalização. Na seção 2.3.1 aborda-se a distribuição Kumaraswamy generalizada para o caso bivariado. Na seção 2.3.2 apresenta-se o modelo de regressão Kumaraswamy bivariado. A seção 2.4 desenvolve uma técnica de análise de sensiblidade para o modelo proposto na seção 2.3.2. Na seção 2.5 realiza-se o ajuste do modelo proposto aos dados de insuficiência renal. Para concluir, a seção 2.6 relata as conclusões obtidas e propostas para pesquisas futuras.

2.2 A distribuição Kumaraswamy e sua generalização

Na literatura existem muitas distribuições de probabilidade clássicas, como, por exemplo, Normal, Log-normal e Beta, entre outras. Kumaraswamy (1980) propõe uma nova distribuição com dois parâmetros, com diversas aplicações na área hidrológica, conhecida como distribuição Kumaraswamy. Para simplificar, a ditribuição Kumaraswamy será chamada de K ao longo do texto. A função de densidade de probabilidade (fdp) e a função de distribuição acumulada (fda) Kumaraswamy são definidas, respectivamente por:

$$g(t;\alpha,\beta) = \alpha\beta t^{\alpha-1}(1-t^{\alpha})^{\beta-1}, \quad 0 < t < 1, \tag{1}$$

e

$$G(t;\alpha,\beta) = 1-(1-t^{\alpha})^{\beta}, \quad 0 < t < 1, \tag{2}$$

em que $\alpha > 0$ e $\beta > 0$ são parâmetros de forma. Algumas propriedades da forma da distribuição Kumaraswamy são: ser unimodal quando $\alpha > 1$ e $\beta > 1$, uniantimodal quando $\alpha < 1$ e $\beta < 1$,

crescente quando $\alpha > 1$ e $\beta \leq 1$, decrescente quando $\alpha \leq 1$ e $\beta > 1$ e constante para $\alpha = \beta = 1$. Essas propriedades são as mesmas da distribuição Beta, mas a Kumaraswamy tem algumas vantagens no que se refere à tratabilidade matemática.

Jones (2009) explora as características da distribuição Kumaraswamy e mostra algumas semelhanças e vantagens em relação a distribuição Beta. Alguns relatos são referentes à constante normalizadora, que na equação (1) tem forma muito simples, à função de distribuição (2) e às funções quantílicas que não envolvem funções especiais, fórmula explícita para L-momentos e simples formulação dos momentos das estatísticas de ordem.

Não obstante, a distribuição Beta mostra algumas vantagens: apresenta formulação simples da função geradora de momentos e dos momentos, distribuições simétricas de um parâmetro como casos particulares, simples estimação dos momentos e mais formas de generalização de distribuições por processos físicos.

Cordeiro e Castro (2010) propõem uma nova classe de distribuições, chamada Kumaraswamy generalizada, baseada nos trabalhos de Eugene et al. (2002) e Jones (2009). Neste trabalho essa distribuição será denotada por K-G. A formulação da distribuição K-G considera uma fda $G(t)$ arbitrária, de forma que sua fda $F(t)$ e fdp $f(t)$ são, respectivamente

$$F(t; a, b) = 1 - [1 - G(t)^a]^b, \tag{3}$$

e

$$f(t; a, b) = abg(t)G(t)^{a-1}[1 - G(t)^a]^{b-1}, \tag{4}$$

em que $g(t) = dG(t)/dt$, $a > 0$ e $b > 0$ são dois parâmetros adicionais, cujo papel é introduzir assimetria e flexibilizar os pesos das caudas. Se T é uma variável aleatória com fdp (4), pode-se escrever $T \sim \mathrm{K} - \mathrm{G}(a, b)$.

Dada a simplicidade das funções (4) e (3), a distribuição K-G pode ser utilizada mesmo que existam observações censuradas nos dados. A interpretação física dessa distribuição supõe um sistema composto de b componetes independentes sendo que cada componente é composto de a subcomponentes independentes. Supõe-se que o sistema falhe se algum dos b componentes falhem e que cada componente falhe se todos os subcomponentes falhem. Sejam $T_{i1}, \ldots, T_{ia}$ os tempos de falha dos subcomponentes dentro do i-ésimo componente, $i = 1, \ldots, b$,

com comum função distribuição acumulada. Seja T_i o tempo de vida do i-ésimo componente, $i = 1, \ldots, b$ e seja T o tempo de vida de todo o sistema. A fda de T é dada por:

$$\begin{aligned} P(T \leq t) &= 1 - P(T_1 > t, \ldots, T_b > t) = 1 - [P(T_1 > t)]^b \\ & 1 - [1 - P(T_1 \leq t)]^b = 1 - [1 - P(T_{11} \leq t, \ldots, T_{1a} \leq t)]^b \\ & 1 - [1 - P(T_{11} \leq t)^a]^b = 1 - [1 - G(t)^a]^b, \end{aligned}$$

resultando nas expressões (4) e (3) da distribuição K-G que representam a ditribuição do tempo de vida de todo o sistema.

Assim como foram discutidas algumas semelhanças e diferenças entre as distribuições Kumaraswamy e Beta, a distribuição K-G também pode ser comparada com a distribuição beta generalizada proposta por Eugene et al. (2002). Ao considerar $b = 1$ na função (4), claramente obtem-se a fdp da função beta generalizada. Contudo, como afirmam Cordeiro e Castro (2010) a grande vantagem da distribuição K-G é não envolver funções especiais na sua função densidade de probabilidade.

As funções K-G são construídas ao associar o nome, a fdp e a fda de cada distribuição contínua. Por exemplo, pode-se obter a distribuição K-normal colocando-se a fdp e a fda da distribuição normal na equação (4). Analogamente, constroem-se as distribuições K-Weibull, K-gama, K-Gumbel, K-Gaussiana inversa, entre outras funções de interesse do pesquisador, seguindo o mesmo procedimento.

Uma distribuição muito popular, extensimavente utilizada na década passada, é a distribuição Weibull proposta originalmente por W. Weibull em estudos relacionados ao tempo de falha devido à fadiga de metais (Colosimo; Giolo, 2006).

Esse é talvez o modelo de distribuição mais amplamente utilizado em análise de tempos de vida. A sua popularidade em aplicações práticas deve-se ao fato de que ela apresenta uma grande variedade de formas, todas com uma propriedade básica: a sua função taxa de falha é monótona, isto é, ela pode ser crescente, decrescente ou constante, e tem a distribuição exponencial como caso particular quando o parâmetro de forma é um.

Devido à grande importância da distribuição Weibull, Cordeiro et al. (2010) propõem a distribuição Kumaraswamy Weibull, denotada por K-Weibull, e fornecem uma descrição detalhada de algumas de suas propriedades matemáticas. Neste trabalho, dar-se-á

ênfase à ditribuição K-Weibull que é obtida ao considerar a fdp da distribuição Weibull,

$$G(t;\lambda,c) = 1 - \exp[-(\lambda t)^c], \tag{5}$$

em que $c > 0$ é o parâmetro de forma e $\lambda > 0$ é o parâmetro de escala.

Logo a fdp e a fda da distribuição K-Weibull são representadas por,

$$f(t;\lambda,c,a,b) = abc\lambda^c t^{c-1} \exp[-(\lambda t)^c]\{1 - \exp[-(\lambda t)^c]\}^{a-1}\{1 - \{1 - \exp[-(\lambda t)^c]\}^a\}^{b-1} \tag{6}$$

e

$$F(t;\lambda,c,a,b) = 1 - \{1 - \{1 - \exp[-(\lambda t)^c]\}^a\}^b, \tag{7}$$

respectivamente.

A função de sobrevivência e função de taxa de falha associadas à distribuição K-Weibull são, respectivamente,

$$S(t;\lambda,c,a,b) = \{1 - \{1 - \exp[-(\lambda t)^c]\}^a\}^b$$

e

$$h(t;\lambda,c,a,b) = \frac{abc\lambda^c t^{c-1} \exp[-(\lambda t)^c]\{1 - \exp[-(\lambda t)^c]\}^a}{\{1 - \{1 - \exp[-(\lambda t)^c]\}^a\}}.$$

Alguns casos particulares da equação (6) são representados ao considerar $a = b = 1$ resultando na distribuição Weibull. Quando $b = 1$ tem-se a distrinbuição Weibull exponenciada, quando $c = b = 1$ a distribuição Exponencial exponenciada, entre outros importantes sub-modelos. Contudo, a flexibilidade da distribuição K-Webull destaca-se ao compará-la aos seus respectivos sub-modelos em aplicações com dados reais. Maiores discussões e informações encontram-se em Cordeiro et al. (2010).

2.3 Modelo bivariado

Em estudos com duas ou mais variáveis de interesse, os modelos multivariados são mais apropriados do que os univariados, uma vez que conseguem captar uma possível associação entre variáveis resposta. Um caso particular dos modelos multivariados são os modelos bivariados. Para modelar essa situação existem algumas distribuições bivariadas que vêm

sendo estudadas na literatura. Por exemplo, a distribuição exponencial bivariada estudada por Gumbel (1960), Freund (1961), Marshall e Olkin (1967) e Block e Basu (1974) e a distribuição Weibull bivariada estudada por Marshall e Olkin (1967), Ryu (1993) e Johnson et al. (1999).

Gumbel (1960) propôs algumas distribuições com marginais exponenciais, mas não discutiu em que situações esses modelos se aplicam. Freund (1961) apresentou uma extensão exponencial específica para um sistema com dois componentes, em que o sistema pode continuar funcionando mesmo depois da falha de um dos componentes.

Entre outras distribuições exponenciais bivariadas propostas, a que recebe maior destaque é a distribuição Exponencial Bivariada (EBV) de Marshall e Olkin (1967). Essa distribuição tem propriedades com interpretações físicas simples e úteis e satisfaz as propriedades de falta de memória marginal e conjunta. Dando continuidade ao trabalho de Marshall e Olkin (1967), Moeschberger (1974) deriva um modelo Weibull para riscos competitivos a partir da distribuição Weibull bivariada de Marshall e Olkin.

Entretanto, a distribuição EBV apresenta a desvantagem de não ser apropriada em situações em que dois componentes não falharam simultaneamente e em algumas situações em que não são razoáveis as propriedades de falta de memória marginal e conjunta. Devido a esses fatores, Ryu (1993) propôs um modelo exponencial bivariado adequado para as situações não contempladas pela EBV, o qual também possui propriedades com interpretações simples. Esse modelo é uma extensão da EBV, conhecido como EEBV (Extensão da Distribuição Exponencial Bivariada).

Alguns exemplos de aplicações das distribuições Exponenciais e Weibull bivariadas podem ser vistos, por exemplo, em: Johnson et al. (1999), cujos estudos fazem uma revisão de técnicas de obtenção de distribuições bivariadas, dando ênfase à distribuição Weibull bivariada. Também Oliveira (2001) estuda a distribuição EEBV sob a abordagem Bayes-empírica e formula um modelo para testes acelerados, em que os tempos até as falhas seguem essa distribuição, assumindo uma relação de potência inversa entre os tempos e a voltagem. A autora utiliza a análise clássica da distribuição EEBV para trabalhar com dados bivariados com e sem a presença de censura. Nessa linha de pesquisa, Tarumoto (2001) desenvolve um modelo de riscos competitivos baseado no modelo proposto por Ryu (1993).

Dada a importância dos modelos bivariados, Cordeiro et al.(2010) propõem a

generalização da distribuição K-Weibull definida em (6) para o caso bivariado que será apresentada na seção 2.3.1.

2.3.1 A distribuição Kumaraswamy Weibull bivariada

Recentemente, Cordeiro et al. (2010) consideraram a extensão da distribuição Kumaraswamy generalizada para o caso bivariado de maneira simples ao considerar que $G(t_1, t_2)$ representa uma fda definida no suporte $(0, \infty) \times (0, \infty)$, com fdp conjunta $g(t_1, t_2)$, respectivas marginais com fdp $g(t_k)$ e fda marginais $G(t_k)$, $k = 1, 2$. Seguindo a mesma idéia utilizada para descrever a distribuição K-G, a distribuição Kumaraswamy generalizada bivariada, K-G bivariada, é expressa por,

$$F(t_1, t_2) = 1 - [1 - G(t_1, t_2)^a]^b, \tag{8}$$

para $a > 0$ e $b > 0$ representando os parâmetros adicionais, em que $G(t_1, t_2)$ pode ser uma fda conjunta arbitrária, como, por exemplo, a função distribuição Weibull bivariada definida por Hougaard (1986) ou a função distribuição Weibull bivariada introduzida por Lu e Bhattacharyya (1990), entre outras possibilidades mencionadas.

Neste trabalho é considerada a distribuição Weibull bivariada definida por Hougaard (1986) devido à sua aplicabilidade e popularidade, evidenciadas na literatura. Dessa maneira define-se $G(t_1, t_2)$ por:

$$G(t_1, t_2) = \exp\Big\{ - \Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}} \Big]^{\alpha} \Big\} - \exp[-(\lambda_1 t_1)^{c_1}] - \exp[-(\lambda_2 t_2)^{c_2}] + 1 \tag{9}$$

e sua respectiva fdp por:

$$\begin{aligned} g(t_1, t_2) &= c_1 \lambda_1 (\lambda_1 t_1)^{\frac{c_1}{\alpha} - 1} c_2 \lambda_2 (\lambda_2 t_2)^{\frac{c_2}{\alpha} - 1} \Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}} \Big]^{\alpha - 2} \\ &\quad \times \Big\{ \Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}} \Big]^{\alpha} + \frac{1}{\alpha} - 1 \Big\} \exp\Big\{ - \Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}} \Big]^{\alpha} \Big\}, \end{aligned}$$

em que $\lambda_k > 0$ parâmetro de escala e $c_k > 0$ parâmetro de forma, para $k = 1, 2$. O parâmetro $0 < \alpha \leq 1$ mede a associação entre T_1 e T_2, e quando esse parâmetro é igual a um significa que as variáveis T_1 e T_2 são independentes. As fdp marginais e as fda marginais relacionadas à distribuição Weibull bivariada definida acima são, respectivamente,

$$g(t_k) = c_k \lambda_k^{c_k} t_k^{c_k - 1} \exp[-(\lambda_k t_k)^{c_k}] \quad \text{e} \quad G(t_k) = 1 - \exp[-(\lambda_k t_k)^{c_k}],$$

para $k = 1, 2$. Ao inserir (9) em (8) tem-se um tipo da distribuição K-Weibull bivariada, com fda expressa por,

$$\begin{aligned} F(t_1, t_2) &= 1 - \Big[1 - \Big\{ \exp\Big\{ - \Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}}\Big]^{\alpha}\Big\} \\ &\quad - \exp[-(\lambda_1 t_1)^{c_1}] - \exp[-(\lambda_2 t_2)^{c_2}] + 1\Big\}^{a}\Big]^{b}. \end{aligned} \tag{10}$$

As fdp marginais $f(t_k)$ e as fda marginais $F(t_k)$, para $k = 1, 2$, da distribuição K-Weibull bivariada (10) são, respectivamente,

$$f(t_k) = abg(t_k)G(t_k^{a-1})[1 - G(t_k)^a]^{b-1} \tag{11}$$

e

$$F(t_k) = 1 - [1 - G(t_k)^a]^b. \tag{12}$$

A função de sobrevivência associada a (10) é expressa utilizando a definição de função de sobrevivência bivariada dada por:

$$S(t_1, t_2) = 1 - F(t_1) - F(t_2) + F(t_1, t_2), \tag{13}$$

em que as funções $F(t_1)$, $F(t_2)$ e $F(t_1, t_2)$ são definidas em (12) e (10), respectivamente. Dessa maneira, pode-se obter a fdp bivariada da distribuição K-Weibull bivariada ao derivar duas vezes em relação aos tempos a função distribuição (10), isto é, $f(t_1, t_2) = \partial^2 F(t_1, t_2)/\partial t_1 \partial t_2$, ou usando a função de sobrevivência bivariada (13), $f(t_1, t_2) = (-1)^2 \partial^2 S(t_1, t_2)/\partial t_1 \partial t_2$. Consequentemente, a fdp da distribuição K-Weibull bivariada reduz-se a

$$f(t_1, t_2) = \frac{abG^{a-2}(t_1, t_2)[A(t_1, t_2) + B(t_1, t_2) + C(t_1, t_2)]}{[1 - G^a(t_1, t_2)]^{1-b}},$$

em que

$$A(t_1, t_2) = -\frac{a(b-1)G^a(t_1, t_2)}{1 - G^a(t_1, t_2)} \frac{\partial G(t_1, t_2)}{\partial t_1} \frac{\partial G(t_1, t_2)}{\partial t_2},$$

$$B(t_1, t_2) = (a-1)\frac{\partial G(t_1, t_2)}{\partial t_1} \frac{\partial G(t_1, t_2)}{\partial t_2},$$

$$C(t_1, t_2) = G(t_1, t_2)g(t_1, t_2),$$

$$\frac{\partial G(t_1,t_2)}{\partial t_1} = -\Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}}\Big]^{\alpha-1} t_1^{\frac{c_1}{\alpha}-1}\lambda_1^{\frac{c_1}{\alpha}} c_1 \exp\Big\{-\Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}}\Big]^{\alpha}\Big\} + \lambda_1^{c_1} t_1^{c_1-1} c_1 \exp[-(\lambda_1 t_1)^{c_1}]$$

e

$$\frac{\partial G(t_1,t_2)}{\partial t_2} = -\Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}}\Big]^{\alpha-1} t_2^{\frac{c_2}{\alpha}-1}\lambda_2^{\frac{c_2}{\alpha}} c_2 \exp\Big\{-\Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}}\Big]^{\alpha}\Big\} + \lambda_2^{c_2} t_2^{c_2-1} c_2 \exp[-(\lambda_2 t_2)^{c_2}].$$

Definidas a fdp, fda e função de sobrevivência conjunta da distribuição K-Weibull bivariada, propõe-se, na próxima seção, uma extensão dessa distribuição para incluir covariáveis, dando origem ao modelo de regressão Kumaraswamy Weibull bivariado.

2.3.2 Modelo de regressão Kumaraswamy Weibull bivariado

Na prática, a grande maioria dos estudos envolve covariáveis que podem estar relacionadas com o tempo de sobrevivência. Em estudos clínicos, por exemplo, o tratamento ao qual o paciente é submetido deve ser considerado como uma covariável, pois esse pode influenciar no tempo de sobrevivência do paciente. Na indústria, o tempo de sobrevivência de um determinado equipamento pode ser influenciado pelo nível de voltagem ao qual é submetido. Tais covariáveis explicam parte da heterogeidade presente na população. Assim, sua inclusão na análise estatística dos dados é de extrema relevância. Por isso, o objetivo deste capítulo é estender a distribuição K-Weibull bivariada pela inclusão de um vetor de covariáveis $\boldsymbol{x}$.

Modelos de regressão podem ser formulados seguindo diversos caminhos. Em análise de sobrevivência, considera-se a classe do modelo de regressão paramétrico e a classe do modelo de regressão de Cox. No entanto, nesta seção, é considerada uma reparametrição da distribuição K-Weibull bivariado para se obter um modelo de regressão.

Sejam T_1 e T_2 variáveis aleatórias com distribuição K-Weibull bivariada definida em (10) dada por:

$$F(t_1,t_2) = 1 - \Big[1 - \Big\{\exp\Big\{-\Big[(\lambda_1 t_1)^{\frac{c_1}{\alpha}} + (\lambda_2 t_2)^{\frac{c_2}{\alpha}}\Big]^{\alpha}\Big\} - \exp[-(\lambda_1 t_1)^{c_1}] - \exp[-(\lambda_2 t_2)^{c_2}] + 1\Big\}^{a}\Big]^{b},$$

em que $t_k > 0$, $a > 0$ e $b > 0$ parâmetros adicionais, $\lambda_k > 0$ parâmetro de escala, $c_k > 0$ parâmetro de forma e $0 < \alpha \leq 1$ parâmetro de associação entre T_1 e T_2, para $k = 1, 2$. Neste

trabalho, propõem-se as reparametrizações $\lambda_k = \exp\{-\mu_k\}$ e $c_k = 1/\sigma_k$, em que $-\infty < \mu_k < \infty$ e $0 < \sigma_k < \infty$. Ao considerar o vetor de variáveis regressoras $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)^T$, usualmente considera-se $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros desconhecidos associado às covariáveis. Assim, a reparametrização pode ser escrita por: $\lambda_k = \exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\} = \exp\{-(\beta_{0k}x_0 + \beta_{1k}x_1 + \ldots + \beta_{pk}x_p)\}$ e $c_k = 1/\sigma_k$, para $k = 1, 2$. O modelo de regressão Kumaraswamy Weibull bivariado é definido por:

$$\begin{aligned} F(t_1, t_2|\boldsymbol{x}) &= 1 - \Big[1 - \Big\{ \exp\Big\{ - \Big[(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_1\}t_1)^{\frac{1}{\sigma_1\alpha}} + (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_2\}t_2)^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha}\Big\} \\ &\quad - \exp[-(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_1\}t_1)^{\frac{1}{\sigma_1}}] - \exp[-(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_2\}t_2)^{\frac{1}{\sigma_2}}] + 1\Big\}^a\Big]^b, \end{aligned} \quad (14)$$

com respectivas fdp marginais $f(t_k|\boldsymbol{x})$ e fda margianis $F(t_k|\boldsymbol{x})$, dadas por:

$$\begin{aligned} f(t_k|\boldsymbol{x}) &= ab\frac{1}{\sigma_k}(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\})^{\frac{1}{\sigma_k}} t_k^{\frac{1}{\sigma_k}-1} \exp\Big[- (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\}t_k)^{\frac{1}{\sigma_k}}\Big] \\ &\quad \Big\{1 - \exp[-(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\}t_k)^{\frac{1}{\sigma_k}}]\Big\}^{a-1}\Big[1 - \Big\{1 - \exp[-(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\}t_k)^{\frac{1}{\sigma_k}}]\Big\}^a\Big]^{b-1} \end{aligned}$$

e

$$F(t_k|\boldsymbol{x}) = 1 - \Big[1 - \Big\{1 - \exp\Big[- (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\}t_k)^{\frac{1}{\sigma_k}}\Big]\Big\}^a\Big]^b,$$

em que $k = 1, 2$.

Ao utilizar a definição de sobrevivência bivariada, a função de sobrevivência associada ao modelo (14) é expressa por:

$$\begin{aligned} S(t_1, t_2|\boldsymbol{x}) &= \Big[1 - \Big\{1 - \exp\Big[- (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_1\}t_1)^{\frac{1}{\sigma_1}}\Big]\Big\}^a\Big]^b \\ &\quad + \Big[1 - \Big\{1 - \exp\Big[- (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_2\}t_2)^{\frac{1}{\sigma_2}}\Big]\Big\}^a\Big]^b \\ &\quad - \Big[1 - \Big\{ \exp\Big\{ - \Big[(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_1\}t_1)^{\frac{1}{\sigma_1\alpha}} + (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_2\}t_2)^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha}\Big\} \\ &\quad - \exp[-(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_1\}t_1)^{\frac{1}{\sigma_1}}] - \exp[-(\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_2\}t_2)^{\frac{1}{\sigma_2}}] + 1\Big\}^a\Big]^b, \end{aligned} \quad (15)$$

com funções de sobrevivência marginais dada por:

$$S_k(t_k|\boldsymbol{x}) = \lim_{t_k \to 0} S(t_1, t_2|\boldsymbol{x}) = \Big[1 - \Big\{1 - \exp\Big[- (\exp\{-\boldsymbol{x}^T\boldsymbol{\beta}_k\}t_k)^{\frac{1}{\sigma_k}}\Big]\Big\}^a\Big]^b. \quad (16)$$

Após a definição do modelo, uma importante etapa é descrever a metodologia utilizada para estimar os parâmetros do modelo proposto. Na próxima seção, encontra-se o procedimento de inferência para modelo de regressão Kumaraswamy Weibull bivariado.

2.3.2.1 Inferência para o modelo de regressão Kumaraswamy Weibull bivariado

Seja uma amostra aleatória observada $(t_{1k}, \delta_{1k}, \boldsymbol{x}_1), \ldots, (t_{nk}, \delta_{nk}, \boldsymbol{x}_n)$, em que t_{ik} representa o tempo de falha ou tempo de censura do k-ésimo evento de interesse do i-ésimo indivíduo, δ_{ik} a respectiva variável indicadora de censura e $\boldsymbol{x}_i$ o vetor de variáveis regressoras associado ao i-ésimo indivíduo, em que $k = 1, 2$ e $i = 1, \ldots, n$.

O método de estimação de máxima verossimilhança para o caso de dados bivariados descrito neste trabalho é baseado nos trabalhos de Lawless (2003) e He e Lawless (2005), em que se considera que t_{i1} e t_{i2} podem ser observações de falha ou censura. Devido a esse fator, a função de verossimilhança será construida com base em quatro possibilidades de combinações de falhas e censuras. Ou seja, um indivíduo pode falhar nos dois eventos de interese, ou um indivíduo pode falhar no primeiro evento e representar uma observação censurada no segundo, ou representar uma observação censurada no primeiro evento e falhar no segundo, ou, ainda, representar uma observação censurada em ambos eventos. Ao utilizar essas informações dos indivíduos, o logaritmo da função de verossimilhança para o modelo de regressão Kumaraswamy Weibull bivariado é, obtido ao considerar a função de verossimilhança descrita por Lawless (2003) para dados bivariados, é expresso por:

$$\begin{aligned} l(\boldsymbol{\Psi}) &= \sum_{i=1}^{n} \left\{ \delta_{i1}\delta_{i2} f(t_{i1}, t_{i2}|\boldsymbol{x}_i) + \delta_{i1}(1-\delta_{i2}) \left[\frac{-\partial S(t_{i1}, t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}} \right] + \right. \\ &\quad \left. (1-\delta_{i1})\delta_{i2} \left[\frac{-\partial S(t_{i1}, t_{i2}|\boldsymbol{x}_i)}{\partial t_{i2}} \right] + (1-\delta_{i1})(1-\delta_{i2}) S(t_{i1}, t_{i2}|\boldsymbol{x}_i) \right\} \end{aligned} \tag{17}$$

em que $S(t_{i1}, t_{i2}|\boldsymbol{x}_i)$ é a função de sbrevivência definida em (15), fdp é obtida por $f(t_1, t_2) = (-1)^2 \partial^2 S(t_1, t_2)/\partial t_1 \partial t_2$ ou por meio da função distribuição acumulada expressa em (14) e, $\boldsymbol{\Psi} = (a, b, \alpha, \boldsymbol{\beta}_k^T, \sigma_k^T)^T$ é o vetor de parâmetros desconhecidos, dado que $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$, para $k = 1, 2$.

Para dar continuidade ao processo de estimação é necessário considerar as restrições existentes em alguns parâmetros. Como já definido ao longo do texto, o parâmetro que mede a associação entre os tempos de falha é definido no espaço paramétrico $0 < \alpha \leq 1$, os parâmetros adicionais são definidos por $a > 0$ e $b > 0$ e o parâmetro de escala é definido por $\sigma_k > 0$, para $k = 1, 2$. Dessa forma, para maximizar o logaritmo da função de verossimilhança definido em (17) é necessário considerar o método da função barreira adaptada (LANGE,

1999), que é uma combinação do método barreira logaritmo com o algoritmo EM.

A obtenção do logaritmo da função de verossimilhança do modelo de regressão Kumaraswamy Weibull bivariado utilizando o método da função barreira adaptada, considera o vetor de parâmetros $\boldsymbol{\Psi}$ sob 6 restrições de inequações lineares $\boldsymbol{u}_j^T\boldsymbol{\Psi} - c_j \geq 0$, em que $\boldsymbol{u}_j, j = 1, 2, \ldots, 6$ são vetores de dimensão 9×1 e c_j são escalares assumindo valores 0 ou 1 dependendo da restrição de interesse. O vetor $\boldsymbol{u}_j$ é escrito por $(1, 0, 0, 0, 0, 0, 0, 0, 0)$, assumindo valor 1 na posição que o parâmetro de interesse encontra-se.

O logaritmo da função de verossimilhança sujeita as restrições lineares é representado por

$$l_R(\boldsymbol{\Psi}, \vartheta) = l(\boldsymbol{\Psi}) + \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T\boldsymbol{\Psi} - c_j), \tag{18}$$

em que o parâmetro de ajuste é uma constante positiva, $\vartheta > 0$, $\boldsymbol{u}_j^T\boldsymbol{\Psi} - c_j$ é o conjunto de restrições de inequações lineares para $j = 1, 2, \ldots, q$, $\boldsymbol{\Psi} = (a, b, \alpha, \boldsymbol{\beta}_k^T, \sigma_k^T)^T$, e $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$.

O estimador de máxima verossimilhança para os parâmetros $\hat{\boldsymbol{\Psi}} = (\hat{a}, \hat{b}, \hat{\alpha}, \hat{\boldsymbol{\beta}}_k^T, \hat{\sigma}_k^T)^T$, do modelo de regressão Kumaraswamy Weibull bivariado, pode ser obtido numericamente ao maximizar o logarítmo da função de verossimilhança definido em (18). Neste trabalho o *software* R (R DEVELOPMENT CORE TEAM, 2009) foi utilizado para obter $\hat{\boldsymbol{\Psi}}$ por meio da função *constrOptim*.

Sob certas condições de regularidade, o vetor de parâmetros $\boldsymbol{\Psi}^*$ possui distribuição assintótica de $\sqrt{n}(\hat{\boldsymbol{\Psi}}^* - \boldsymbol{\Psi}^*)$ dada por $N_{p+1}(0, I^\star(\boldsymbol{\Psi}^*)^{-1})$, em que $I^\star(\boldsymbol{\Psi}^*)$ é a matriz de informação esperada ao considerar o vetor de parâmetros $\boldsymbol{\Psi}^* = \boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ de dimensão $(p + 1)$. A matriz de covariância assintótica $I^\star(\boldsymbol{\Psi}^*)^{-1}$ pode ser aproximada pela inversa da matriz de informação observada $\ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}^*) = -\left\{\frac{\partial^2 l_R(\boldsymbol{\Psi}, \vartheta)}{\partial \boldsymbol{\Psi}^* \partial \boldsymbol{\Psi}^{*T}}\right\}$, avaliada em $\boldsymbol{\Psi}^* = \hat{\boldsymbol{\Psi}}^*$. Assim, os procedimentos de inferência estatística para o vetor de parâmetros $\boldsymbol{\Psi}^*$ podem ser baseados na aproximação para a normal $N(0, -\ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}^*)^{-1})$. Intervalos de confiança assintóticos $100(1 - \alpha^*)\%$ para cada parâmetro $\boldsymbol{\Psi}_v^*$ são dado por

$$ACI_v = \left(\hat{\boldsymbol{\Psi}}_v^* - z_{\alpha^*/2}\sqrt{-\widehat{\ddot{L}_R}^{v,v}}, \hat{\boldsymbol{\Psi}}_v^* + z_{\alpha^*/2}\sqrt{-\widehat{\ddot{L}_R}^{v,v}}\right),$$

em que $-\widehat{\ddot{L}_R}^{v,v}$ denota os elementos da v-ésima diagonal da matriz de informação observada

estimada $-\widehat{\ddot{\boldsymbol{L}}}_R(\boldsymbol{\Psi}^*)^{-1}$ e $z_{\alpha^*/2}$ é o $(1-\alpha^*/2)$ - ésimo quantil da distribuição normal padrão, para $v = 1, \ldots, 2p+2$. Estimados os parâmetros do modelo, outro procedimento de interesse é a realização de testes de hipóteses. Considerando-se $\boldsymbol{\Psi}_1^*$ e $\boldsymbol{\Psi}_2^*$ subconjuntos de $\boldsymbol{\Psi}^*$, uma hipótese de interesse pode ser testar $H_0 : \boldsymbol{\Psi}_1^* = \boldsymbol{\Psi}_{10}^*$ contra $H_1 : \boldsymbol{\Psi}_1^* \neq \boldsymbol{\Psi}_{10}^*$. Considerando-se $\hat{\boldsymbol{\Psi}}^*{}_0$ o estimador de máxima verossimilhança dos parâmetros com e sem restrições, sob H_0, define-se a estatística da razão de verossimilhanças por:

$$LR = 2\left\{\ell(\hat{\boldsymbol{\Psi}}^*) - \ell(\hat{\boldsymbol{\Psi}}^*{}_0)\right\}.$$

Sob H_0 e algumas condições de regularidade, LR converge para uma distribuição qui-quadrada com graus de liberdade igual a dimensão do subconjunto $\boldsymbol{\Psi}_1^*$ de interesse (Cox e Hinkley, 1974).

2.4 Análise de sensibilidade

Uma importante etapa na análise de dados é verificar o quanto as estimativas obtidas a partir do modelo proposto são resistentes a pequenas perturbações nos dados ou no modelo. Se o modelo ajustado não apresentar uma boa descrição dos dados observados, podem ser obtidas conclusões errôneas.

Devido a isso, destaca-se a importância de se realizar um estudo sobre a robustez dos resultados obtidos, considerando-se vários aspectos que envolvem a formulação do modelo e as estimativas dos seus parâmetros.

Essa análise de sensibilidade é obtida ao utilizar a metodologia de Influência Global, Local e Local Total. A primeira é caracterizada pela deleção de casos, em que o i-ésimo indivíduo é retirado da análise. Cook (1977) propõe uma medida de influência conhecida como Distância de Cook para avaliar a influência da deleção de um caso nas estimativas dos parâmetros em modelos de regressão linear. A Distância de Cook mede a influência de cada observação na análise estatística. Outra medida utilizada em Influência Global é o Afastamento da Verossimilhança, que mede o quanto cada observação pode alterar a verossimilhança.

A metodologia de Influência Local proposta por Cook (1986) tem o objetivo de avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo. A escolha da forma de perturbar o modelo proposto entre as inúmeras formas existentes,

deve levar em consideração aspectos da análise que se deseja monitorar. Obviamente, deve-se pensar em esquemas de perturbação que sejam interpretáveis. Lesaffre e Verbeke (1998) desenvolvem a medida de Influência Local Total.

Diversos autores têm estudado essas medidas. Por exemplo, Xie e Wei (2007) propõem medidas de diagnósticos baseadas na deleção de casos para o modelo de regressão log-Birnbaum-Saunders; Escobar e Meeker (1992) adaptaram métodos de Influência Local para os modelo de regressão paramétricos com dados censurados; Ortega et al.(2003) consideram Influência Local em modelos de regressão log-gama generalizada; Ortega et al. (2006) desenvolvem uma análise de Influência Local no modelo de regressão Weibull - exponenciado; Fachini et al.(2008) utilizam Influência Local nos modelos de riscos múltiplos.

No contexto de máxima verossimilhança com restrição nos parâmetros, Kwan e Fung (1998) utilizam Influência Local para análise fatorial sujeito a restrições; Gu e Fung (2001) consideram a abordagem de Influência Local na estrutura da verossimilhança restrita geral e afirmam que a curvatura encontrada para verossimilhança restrita pode ser usada seguindo a mesma abordagem indicada por Cook (1986). Paula e Cysneiros (2009) discutem, em trabalhos mais recentes, a avaliação de Influência Local e "leverage" sujeito a equações de restrições lineares nos parâmetros com extensão a inequações de restrições para modelos lineares generalizados.

A seguir, apresentam-se as medidas de análise de Influência Global e Local propostas por Cook (1977, 1986), sob o enfoque da metodologia de verossimilhança sujeita a restrições nos parâmetros.

2.4.1 Influência Global sob verossimilhança restrita

A primeira metodologia de análise de sensibilidade descrita para avaliar o modelo de regressão Kumaraswamy Weibull bivariado é o procedimento de deleção de casos que avalia o efeito da i-ésima observação nas estimativas, o que permite observar o quanto uma deleção pode alterar os resultados.

Ao considerar que a i-ésima observação foi retirada da amostra, o logaritmo da função de verossimilhança com restrição nos parâmetros é denotado por $l_{R(i)}(\boldsymbol{\Psi}, \vartheta)$. Seja $\hat{\boldsymbol{\Psi}}_{(i)}$ o estimador de máxima verossimilhança sujeito às restrições obtido a partir de $l_{R(i)}(\boldsymbol{\Psi}, \vartheta)$,

a influência da i-ésima observação no estimador de máxima verossimilhança é avaliada por meio da diferença $\hat{\boldsymbol{\Psi}}_{(i)} - \hat{\boldsymbol{\Psi}}$. Quando essa diferença $\hat{\boldsymbol{\Psi}}_{(i)} - \hat{\boldsymbol{\Psi}}$ é relativamente grande, pode-se considerar a observação como influente. Os casos considerados influentes devem ser analisados, pois essa metodologia identifica pontos que podem comprometer ou alterar a formulação do modelo e possíveis pontos extremos ou discrepantes que podem ser resultado de erro na coleta ou administração dos dados.

Na literatura, as medidas de Influência Global mais utilizadas são denominadas Distância de Cook Generalizada e Afastamento da Verossimilhança. A Distância de Cook é definida como a norma padronizada de $\hat{\boldsymbol{\Psi}}_{(i)} - \hat{\boldsymbol{\Psi}}$ e é descrita por:

$$GD_i(\boldsymbol{\Psi}) = (\hat{\boldsymbol{\Psi}}_{(i)} - \hat{\boldsymbol{\Psi}})^T \boldsymbol{M} (\hat{\boldsymbol{\Psi}}_{(i)} - \hat{\boldsymbol{\Psi}}),$$

em que podem ser consideradas várias escolhas de $\boldsymbol{M}$, segundo Cook e Weisberg (1982). Entretanto, as escolhas mais utilizadas entre os pesquisadores é considerar $\boldsymbol{M} = -\ddot{L}(\hat{\boldsymbol{\Psi}})$ ou $\boldsymbol{M} = [-\ddot{L}(\hat{\boldsymbol{\Psi}})]^{-1}$.

A medida Afastamento da Verossimilhança é expressa em função de $l_R(\hat{\boldsymbol{\Psi}}, \vartheta)$ e $l_{R(i)}(\hat{\boldsymbol{\Psi}}, \vartheta)$ e é expressa por:

$$LD_i(\boldsymbol{\Psi}) = 2[l_R(\hat{\boldsymbol{\Psi}}, \vartheta) - l_{R(i)}(\hat{\boldsymbol{\Psi}}, \vartheta)].$$

2.4.2 Influência Local

A metodologia de Influência Local, introduzida primeiramente por Cook (1986), propõe o estudo de pequenas perturbações no modelo que podem causar variações nos resultados do modelo estatístico proposto. Essa metodologia é baseada no Afastamento da Verossimilhança.

Adequando a idéia de Cook (1986) ao modelo de regressão Kumaraswamy Weibull bivariado, para um conjunto de dados observados, considera-se $l(\boldsymbol{\Psi})$ o logaritmo da função de verossimilhança do modelo postulado, em que $\boldsymbol{\Psi}$ é um vetor de dimensão $(2p + 5)$, de parâmetros desconhecidos. Uma perturbação é introduzida no modelo através de um vetor $\boldsymbol{w}$, $m \times 1$, que é restrito a algum subconjunto aberto Ω de $\mathbf{R}^m$. Seja $l(\boldsymbol{\Psi}|\boldsymbol{w})$ o logaritmo da função de verossimilhança do modelo perturbado.

Assume-se que $l(\boldsymbol{\Psi}|\boldsymbol{w})$ é duas vezes continuamente diferenciável em $(\boldsymbol{\Psi}, \boldsymbol{w})^T$ e que o modelo postulado está encaixado no modelo perturbado, ou seja, supõe-se que existe $\boldsymbol{w}_0 \, \epsilon \, \Omega$ tal que $l(\boldsymbol{\Psi}|\boldsymbol{w}_0) = l(\boldsymbol{\Psi})$ para todo $\boldsymbol{\Psi} \, \epsilon \, \mathbf{R}^{(2p+5)}$. Em geral, a dimensão m do vetor de perturbação está relacionada com a dimensão do vetor $\boldsymbol{\Psi}$ ou com o tamanho da amostra, dependendo do esquema de perturbação.

Considere $\hat{\boldsymbol{\Psi}}$ o estimador de máxima verossimilhança de $\boldsymbol{\Psi}$, obtido ao maximizar $l(\boldsymbol{\Psi})$, e $\hat{\boldsymbol{\Psi}}_{\boldsymbol{w}}$ o estimador de $\boldsymbol{\Psi}$ sob $l(\boldsymbol{\Psi}|\boldsymbol{w})$. Uma comparação direta entre $\hat{\boldsymbol{\Psi}}$ e $\hat{\boldsymbol{\Psi}}_{\boldsymbol{w}}$ pode não ser simples devido a diversos fatores, tais como diferença de escala, unidade de medida, erro de estimativa, necessidade de definir um vetor arbitrário de perturbação, entre outros. Então, para avaliar a influência da variação de $\boldsymbol{w}$ em toda parte de Ω, Cook (1986) sugere, inicialmente, considerar o Afastamento da Verossimilhança

$$LD(\boldsymbol{w}) = 2[l(\hat{\boldsymbol{\Psi}}) - l(\hat{\boldsymbol{\Psi}}_{\boldsymbol{w}})] \tag{19}$$

em torno de $\boldsymbol{w}_0$.

Se $LD(\boldsymbol{w})$ for grande, significa que $l(\boldsymbol{\Psi})$ é fortemente curvado para $\hat{\boldsymbol{\Psi}}$, ou seja, $\boldsymbol{\Psi}$ é estimado com grande precisão. Entretanto, $LD(\boldsymbol{w})$ será pequeno se $l(\boldsymbol{\Psi})$ é relativamente plano em $\hat{\boldsymbol{\Psi}}$. Considerando essa idéia, o método consiste em inspecionar a superfície geométrica formada por valores do vetor $(m+1) \times 1$:

$$\boldsymbol{\gamma}(\boldsymbol{w}) = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ LD(\boldsymbol{w}) \end{pmatrix} \tag{20}$$

com $\boldsymbol{w}$ variando em Ω. O comportamento dessa função é avaliado em uma vizinhança do ponto mínimo local $\boldsymbol{w}_0$.

Cook (1986) sugere usar a curvatura normal (20) ao redor de $\boldsymbol{w}_0$ em uma direção unitária $\boldsymbol{d}$, $\| \, \boldsymbol{d} \, \| = 1$ do espaço Ω como uma medida de diagnóstico.

Quando há restrições nos parâmetros do modelo, a verossimilhança e, consequentemente, o Afastamento da Verossimilhança são definidos em um espaço paramétrico restrito. Dessa forma, é necessário obter algumas mudanças algébricas para definir a verossimilhança sujeita as restrições nos parâmetros, como enunciado na seção (2.3.2.1). A seguir, é

descrita a metodologia de Influência Local sob verossimilhança restrita com base nos trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009).

2.4.3 Influência Local sob verossimilhança restrita

Por meio da teoria descrita na seção (2.3.2.1), para maximizar a função $l(\boldsymbol{\Psi})$ sob q restrições de inequações lineares, $\boldsymbol{u}_j^t\boldsymbol{\Psi} - c_j \geq 0$, e equações lineares, $\boldsymbol{v}_j^t\boldsymbol{\Psi} = d_j$, $j = 1, \ldots, q$, é necessário reescrever a função $l(\boldsymbol{\Psi})$ como o logaritmo da função de verossimilhança sujeita as restrições lineares

$$l_R(\boldsymbol{\Psi}, \vartheta) = l(\boldsymbol{\Psi}) + \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^t\boldsymbol{\Psi} - c_j). \tag{21}$$

Ao considerar um vetor de perturbações $\boldsymbol{w}_{m\times 1}$, o logaritmo da função de verossimilhança perturbada sujeita as restrições lineares é definido por

$$\mathrm{l_R}(\boldsymbol{\Psi}, \vartheta|\boldsymbol{w}) = \mathrm{l}(\boldsymbol{\Psi}|\boldsymbol{w}) + \vartheta \sum_{\mathrm{j=1}}^{\mathrm{q}} (\boldsymbol{u}_\mathrm{j}^\mathrm{t}\boldsymbol{\Psi} - \mathrm{c_j}), \tag{22}$$

em que a constante $\vartheta > 0$ é o multiplicador do termo barreira. É importante observar na equação (22) que as restrições não são alteradas pelo esquema de perturbação, o que garante solução no subespaço paramétrico.

Seja $\hat{\boldsymbol{\Psi}}$ o estimador de máxima verossimilhança de $\boldsymbol{\Psi}$, obtido ao maximizar (21), e $\hat{\boldsymbol{\Psi}}_{\boldsymbol{w}}$ o estimador de $\boldsymbol{\Psi}$ sob (22). Para algum $\boldsymbol{w} \in \Omega$, o Afastamento da Verossimilhança sujeita as restrições lineares, nesse caso, é dado por

$$LD(\boldsymbol{w}) = 2[l_R(\hat{\boldsymbol{\Psi}}, \vartheta) - l_R(\hat{\boldsymbol{\Psi}}_{\boldsymbol{w}}, \vartheta)], \tag{23}$$

que é uma extensão da equação (19).

Cook (1986) sugere, como uma medida de diagnóstico, estudar $LD(\boldsymbol{w})$ em uma vizinhança de $\boldsymbol{w}_0$ para cada direção $\boldsymbol{d}$. Desta forma, $\boldsymbol{w}$ pode ser representado por meio de

$$\boldsymbol{w}(o) = \boldsymbol{w}_0 + o\boldsymbol{d},$$

em que $o \in \boldsymbol{R}^1$. O gráfico de $LD(\boldsymbol{w}_0 + o\boldsymbol{d})$ contra o, é chamado de linha projetada. Cada linha pode ser investigada por meio da curvatura normal $C_{\boldsymbol{d}}$ em torno de $o = 0$.

Gu e Fung (2001) mostram que, utilizando a expansão de Taylor de segunda ordem de $LD(\boldsymbol{w}_0 + a\boldsymbol{d})$ em torno de $\boldsymbol{w}_0$ e realizando alguns cálculos, obtém-se a seguinte expressão

$$LD(\boldsymbol{w}_0 + o\boldsymbol{d}) \approx -o^2 \boldsymbol{d}^T \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta)^{-1} \boldsymbol{\Delta} \boldsymbol{d}$$

e que a matriz $\boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta)^{-1} \boldsymbol{\Delta}$ pode ser usada no mesmo enfoque introduzido por Cook (1986). Assim, a curvatura normal na direção $\boldsymbol{d}$ é dada por

$$C_{\boldsymbol{d}} = 2\|\boldsymbol{d}^T \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta)^{-1} \boldsymbol{\Delta} \boldsymbol{d}\|,$$

em que

$$\boldsymbol{\Delta} = \frac{\partial^2 l_R(\boldsymbol{\Psi}, \vartheta | \boldsymbol{w})}{\partial \boldsymbol{\Psi} \partial \boldsymbol{w}^T}$$

e

$$\ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta) = -\frac{\partial^2 l_R(\boldsymbol{\Psi}, \vartheta)}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T},$$

ambas matrizes avaliadas em $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$, sendo que $\boldsymbol{\Delta}$ depende do esquema de perturbação utilizado e $\ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta)$ é a matriz de informação do modelo postulado.

Uma importante informação é saber a direção que produz a maior Influência Local na estimativa dos parâmetros. Essa direção é dada por $\boldsymbol{d}_{max}$, sendo esse o autovetor normalizado correspondente ao maior autovalor $C_{\boldsymbol{d}_{max}}$ da matriz $F = \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta)^{-1} \boldsymbol{\Delta}$. O gráfico do autovetor $\boldsymbol{d}_{max}$ contra a ordem das observações pode identificar as observações mais influentes para o esquema de perturbação considerado.

Lesaffre e Verbeke (1998) sugerem considerar, também, a direção do i-ésimo indivíduo que corresponderia a $\boldsymbol{d}_i = (0, \ldots, 1, \ldots, 0)^T$, tal que o i-ésimo elemento é um. Sendo assim, a curvatura normal chamada Influência Local Total do i-ésimo indivíduo é representada por:

$$C_i = 2|\Delta_i^T \ddot{\boldsymbol{L}}_R(\boldsymbol{\Psi}, \vartheta)^{-1} \Delta_i|.$$

O gráfico de C_i contra a ordem das observações pode ser usado como diagnóstico em Influência Local. Os possíveis pontos, tal que $C_i \geq 2\bar{C}$ em que $\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$, merecem uma atenção especial.

De acordo com o modelo proposto neste capítulo, apresenta-se a seguir três esquemas de perturbações mais comumente utilizados: ponderação de casos, perturbação na variável resposta e perturbação na variável explicativa. Para cada um dos esquemas de perturbação considerados é necessário obter a matriz $\boldsymbol{\Delta}$, com componentes $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_a, \boldsymbol{\Delta}_b, \boldsymbol{\Delta}_\alpha, \boldsymbol{\Delta}_{\boldsymbol{\beta}_k}, \boldsymbol{\Delta}_{\sigma_k})^T$, definida por:

$$\Delta = \Big(\Delta_{vi}\Big)_{\big[(2p+5)\times n\big]} = \left(\frac{\partial^2 l_R(\boldsymbol{\Psi}, \vartheta|\boldsymbol{w})}{\partial \boldsymbol{\Psi}_v \partial w_i}\right)_{\big[(2p+5)\times n\big]},$$

em que $v = 1, 2, \ldots, 2p+5$ e $i = 1, 2, \ldots, n$, considerando o modelo definido em (14) e o logaritmo da função de verossimilhança sujeita as restrições lineares (22). Neste trabalho as matrizes referentes a cada esquema de perturbação são obtidas numericamente.

2.4.3.1 Perturbação de casos

Ao considerar o vetor de perturbação $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^T$, o logaritmo da função de verossimilhança perturbada sujeita as restrições lineares é expresso por:

$$\begin{aligned} l(\boldsymbol{\Psi}) \;=\; & \sum_{i=1}^{n} w_i \left\{ \delta_{i1}\delta_{i2} f(t_{i1}, t_{i2}|\boldsymbol{x}_i) + \delta_{i1}(1-\delta_{i2}) \left[\frac{-\partial S(t_{i1}, t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}}\right] \right\} + \\ & \sum_{i=1}^{n} w_i \left\{ (1-\delta_{i1})\delta_{i2} \left[\frac{-\partial S(t_{i1}, t_{i2}|\boldsymbol{x}_i)}{\partial t_{i2}}\right] + (1-\delta_{i1})(1-\delta_{i2}) S(t_{i1}, t_{i2}|\boldsymbol{x}_i) \right\} + \\ & \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T \boldsymbol{\Psi} - c_j), \end{aligned}$$

em que o vetor correspondente à não perturbação é o vetor $\boldsymbol{w}_0 = (1, \ldots, 1)^T$, n-dimensional. Esse esquema de perturbação pode ser interpretado como uma flexibilização da deleção de casos, de forma que a influência conjunta das observações pode ser investigada.

2.4.3.2 Perturbação da variável resposta

Nesse esquema de perturbação considera-se que cada variável resposta t_{i1} e t_{i2} é perturbada como $t_{i1}^* = t_{i1} + S_{t_1} w_i$ e $t_{i2}^* = t_{i2} + S_{t_2} w_i$, em que S_{t_k} são fatores de escala que podem ser a estimativa do desvio padrão da variável T_k, $k = 1, 2$ e $w_i \in \mathbf{R}$. O logaritmo da

verossimilhança perturbada sujeita as restrições lineares é dado por:

$$
\begin{aligned}
l(\boldsymbol{\Psi}) &= \sum_{i=1}^{n}\left\{\delta_{i1}\delta_{i2}f(t^*_{i1},t^*_{i2}|\boldsymbol{x}_i)+\delta_{i1}(1-\delta_{i2})\left[\frac{-\partial S(t^*_{i1},t^*_{i2}|\boldsymbol{x}_i)}{\partial t^*_{i1}}\right]\right\}+ \\
&\sum_{i=1}^{n}\left\{(1-\delta_{i1})\delta_{i2}\left[\frac{-\partial S(t^*_{i1},t^*_{i2}|\boldsymbol{x}_i)}{\partial t^*_{i2}}\right]+(1-\delta_{i1})(1-\delta_{i2})S(t^*_{i1},t^*_{i2}|\boldsymbol{x}_i)\right\}+ \\
&\vartheta\sum_{j=1}^{q}(\boldsymbol{u}_j^T\boldsymbol{\Psi}-c_j),
\end{aligned}
$$

em que $t^*_{ik}=t_{ik}+S_{t_k}w_i$ e o vetor de não perturbação $\boldsymbol{w}_0=(0,\ldots,0)^T$. Nesse caso, podem ser consideradas três possibilidades de perturbação, ou seja, perturbar apenas a variável resposta t_{i1} ou apenas t_{i2} ou t_{i1} e t_{i2} conjuntamente. A idéia de perturbar o modelo seguindo esse esquema de perturbação pode ser interpretada como uma forma de detectar pontos atípicos, com erro ou simplesmente mal modelados.

2.4.3.3 Perturbação de uma covariável

Nesse esquema, a finalidade é avaliar a sensibilidade do modelo a pequenas perturbações em uma particular variável explicativa contínua, X_l. Considere uma perturbação aditiva para a variável explicativa, $x_{ilw}=x_{il}+S_{xl}w_i$, em que S_{xl} é um fator de escala que pode ser a estimativa do desvio padrão de X_l e $w_i\in\mathbf{R}$. O logaritmo da verossimilhança perturbada sujeita as restrições lineares é dado por:

$$
\begin{aligned}
l(\boldsymbol{\Psi}) &= \sum_{i=1}^{n}\left\{\delta_{i1}\delta_{i2}f(t_{i1},t_{i2}|\boldsymbol{x}^*_i)+\delta_{i1}(1-\delta_{i2})\left[\frac{-\partial S(t_{i1},t_{i2}|\boldsymbol{x}^*_i)}{\partial t_{i1}}\right]\right\}+ \\
&\sum_{i=1}^{n}\left\{(1-\delta_{i1})\delta_{i2}\left[\frac{-\partial S(t_{i1},t_{i2}|\boldsymbol{x}^*_i)}{\partial t_{i2}}\right]+(1-\delta_{i1})(1-\delta_{i2})S(t_{i1},t_{i2}|\boldsymbol{x}^*_i)\right\}+ \\
&\vartheta\sum_{j=1}^{q}(\boldsymbol{u}_j^T\boldsymbol{\Psi}-c_j),
\end{aligned}
$$

em que $\boldsymbol{x}_i^{*T}\boldsymbol{\beta}_j=\beta_{0k}+\beta_{1k}x_{i1}+\beta_{2k}x_{i2}+\ldots+\beta_{lk}(x_{il}+S_{xl}w_i)+\ldots+\beta_{pk}x_{ip}$ e o vetor de não perturbação $\boldsymbol{w}_0=(0,\ldots,0)^T$. Para esse esquema de perturbação a idéia de perturbar o modelo tem a mesma interpretação do esquema de perturbação da variável resposta.

2.5 Aplicação

Nesta seção é realizada uma análise estatística dos dados de pacientes com insuficiência renal (MCGILCHRIST; AISBETT, 1991) utilizando o modelo de regressão Kumaraswamy Weibull bivariado desenvolvido neste capítulo. O conjunto de dados corresponde a observações de tempo até a ocorrência de dois eventos distintos de infecção em pacientes com insuficiência renal. Os pacientes utilizam máquinas de diálise portáteis e considera-se a ocorrência das infecções no ponto em que um cateter é inserido. Quando uma infecção ocorre o cateter deve ser removido até que a infecção seja curada para que o cateter seja reinserido. Os tempos entre a inserção do cateter e a ocorrência de uma infecção são registrados. Dessa maneira podem ocorrer diversas infecções nos pacientes, mas apenas duas infecções são registradas nesse conjunto de dados. Existem situações em que o cateter é removido sem que a infecção tenha ocorrido; neste caso, ocorrem as censuras à direita. As variáveis resposta de interesse são os tempos entre as inserções do cateter e a ocorrência da infecção. Nesta análise é considerada a covariável sexo referente a 38 pacientes. Para cada paciente i, $i = 1, 2, \ldots, 38$, as variáveis associadas são descritas por:

- t_{i1}: tempo até a ocorrência da infecção 1, em semanas;
- t_{i2}: tempo até a ocorrência da infecção 2, em semanas;
- δ_{i1}: indicador de censura do evento 1;
- δ_{i2}: indicador de censura do evento 2;
- x_{i1}: sexo do paciente (0: masculino, 1: feminino).

2.5.1 Análise Descritiva

Realizou-se uma análise exploratória dos dados considerando principalmente as variáveis referentes aos tempos até a ocorrência dos dois eventos de interesse. Foi calculado o coeficiente de correlação de τ de Kendall, resultando em $\tau = 0,03$, indicando uma ausência de correlação entre os tempos dos eventos de interesse. Apesar da não correlação, esse conjunto de dados será utilizado para exemplificar a aplicação do modelo de regressão Kumaraswamy

Weibull bivariado. O parâmetro referente à associação entre as respostas contido no modelo deverá detectar essa independência, sem prejuizo do ajuste do modelo aos dados. As estimativas de sobrevivência de Kaplan-Meier e o ajuste marginal do modelo com distribuição K-Weibull para ambos os eventos 1 e 2 são apresentadas na Figura 1, verificando-se que é adequado supor distribuição K-Weibull para os tempos de sobrevivência.

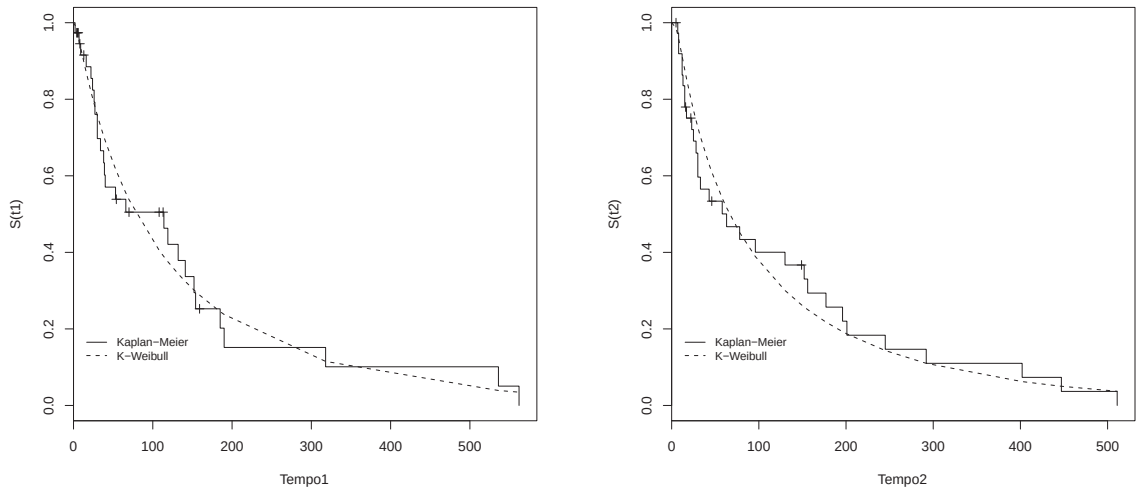

Figura 1 - Curvas de sobrevivência estimadas marginalmente usando a distribuição K-Weibull para comparar com as estimativas não paramétricas obtidas pelo método de Kaplan-Meier para os dados de insuficiência renal

2.5.2 Ajuste do modelo de regressão Kumaraswamy Weibull bivariado

Nesta seção, dar-se-á início a análise dos dados de insuficiência renal considerando o modelo de regressão Kumaraswamy Weibull bivariado. Constatada a adequabilidade marginal da distribuição K-Weibull para os tempos 1 e 2, é coerente, para o caso de dados bivariados, considerar a fda Weibull bivariada definida por Hougaard (1986), expressa

como:

$$\begin{aligned} G(t_{i1}, t_{i2}|\boldsymbol{x}_i) &= \exp\Big\{ - \Big[(\exp\{-\beta_{01} - \beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}} + (\exp\{-\beta_{02} - \beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha}\Big\} \\ &\quad - \exp[-(\exp\{-\beta_{01} - \beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1}}] \\ &\quad - \exp[-(\exp\{-\beta_{02} - \beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2}}] + 1, \end{aligned} \tag{24}$$

tendo essa forma reparametrizada para considerar a presença da covariável sexo na análise, $i = 1, \ldots 38, k = 1, 2$, β_{pk} são os parâmetros de regressão, σ_k são os parâmetros de escala e o parâmetro α mede a associação entre T_1 e T_2.

Ao considerar a fda conjunta definida em (24) e a fda conjunta (8) descrita na seção 2.3.1, tem-se a fda do modelo de regressão Kumaraswamy Weibull bivariado dada por:

$$\begin{aligned} F(t_{i1}, t_{i2}|\boldsymbol{x}_i) &= 1 - \Big[1 - \Big\{ \exp\Big\{ - \Big[(\exp\{-\beta_{01} - \beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}} + \\ &\quad (\exp\{-\beta_{02} - \beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha}\Big\} - \\ &\quad \exp[-(\exp\{-\beta_{01} - \beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1}}] - \\ &\quad \exp[-(\exp\{-\beta_{02} - \beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2}}] + 1\Big\}^{a}\Big]^{b}, \end{aligned} \tag{25}$$

com respectivas funções densidades de probabilidade marginais $f(t_{ik}|\boldsymbol{x}_i)$ e funções de distribuições margianis $F(t_{ik}|\boldsymbol{x}_i)$ dadas por:

$$\begin{aligned} f(t_{ik}|\boldsymbol{x}_i) &= ab\frac{1}{\sigma_k}(\exp\{-\boldsymbol{x}_i^T\boldsymbol{\beta}_k\})^{\frac{1}{\sigma_k}} t_{ik}^{\frac{1}{\sigma_k}-1} \exp\Big[- (\exp\{-\boldsymbol{x}_i^T\boldsymbol{\beta}_k\}t_{ik})^{\frac{1}{\sigma_k}}\Big] \\ &\quad \Big\{1 - \exp[-(\exp\{-\boldsymbol{x}_i^T\boldsymbol{\beta}_k\}t_{ik})^{\frac{1}{\sigma_k}}]\Big\}^{a-1}\Big[1 - \Big\{1 - \exp[-(\exp\{-\boldsymbol{x}_i^T\boldsymbol{\beta}_k\}t_{ik})^{\frac{1}{\sigma_k}}]\Big\}^{a}\Big]^{b-1} \end{aligned}$$

e

$$F(t_{ik}|\boldsymbol{x}_i) = 1 - \Big[1 - \Big\{1 - \exp\Big[- (\exp\{-\boldsymbol{x}_i^T\boldsymbol{\beta}_k\}t_{ik})^{\frac{1}{\sigma_k}}\Big]\Big\}^{a}\Big]^{b},$$

em que $k = 1, 2$. A função de sobrevivência para o modelo de regressão Kumaraswamy Weibull

bivariado é expressa por:

$$\begin{aligned} S(t_{i1},t_{i2}|\boldsymbol{x}_i) &= \Big[1-\Big\{1-\exp\Big[-(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1}}\Big]\Big\}^a\Big]^b \\ &+\Big[1-\Big\{1-\exp\Big[-(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2}}\Big]\Big\}^a\Big]^b \\ &-\Big[1-\Big\{\exp\Big\{-\Big[(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}} \\ &+(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^\alpha\Big\} \\ &-\exp[-(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1}}] \\ &-\exp[-(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2}}]+1\Big\}^a\Big]^b. \end{aligned} \tag{26}$$

Ao considerar os dados de insuficiência renal, o logaritmo da função de verossimilhança sujeita a restrição nos parâmetros para o modelo de regressão Kumaraswamy Weibull bivariado está definido na equação (18) e é composto por:

$$-\frac{\partial S(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}} = f(t_{i1}|\boldsymbol{x}_i) - \frac{[1-G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)]^b bG^a(t_{i1},t_{i2}|\boldsymbol{x}_i)a[\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)/\partial t_{i1}]}{G(t_{i1},t_{i2}|\boldsymbol{x}_i)[1-G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)]},$$

$$-\frac{\partial S(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i2}} = f(t_{i2}|\boldsymbol{x}_i) - \frac{[1-G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)]^b bG^a(t_{i1},t_{i2}|\boldsymbol{x}_i)a[\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)/\partial t_{i2}]}{G(t_{i1},t_{i2}|\boldsymbol{x}_i)[1-G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)]},$$

$$f(t_{i1},t_{i2}|\boldsymbol{x}_i) = \frac{abG^{a-2}(t_{i1},t_{i2}|\boldsymbol{x}_i)\Big[A(t_{i1},t_{i2}|\boldsymbol{x}_i)+B(t_{i1},t_{i2}|\boldsymbol{x}_i)+C(t_{i1},t_{i2}|\boldsymbol{x}_i)\Big]}{[1-G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)]^{1-b}},$$

em que

$$A(t_{i1},t_{i2}|\boldsymbol{x}_i) = -\frac{a(b-1)G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)}{1-G^a(t_{i1},t_{i2}|\boldsymbol{x}_i)}\frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}}\frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i2}},$$

$$B(t_{i1},t_{i2}|\boldsymbol{x}_i) = (a-1)\frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}}\frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i2}},$$

$$C(t_{i1},t_{i2}|\boldsymbol{x}_i) = G(t_{i1},t_{i2}|\boldsymbol{x}_i)g(t_{i1},t_{i2}|\boldsymbol{x}_i),$$

$$\begin{aligned} \frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}} &= -\Big[(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}}+(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha-1}t_{i1}^{\frac{1}{\sigma_1\alpha}-1} \\ &\exp\{-\beta_{01}-\beta_{11}x_{i1}\}^{\frac{1}{\sigma_1\alpha}}\frac{1}{\sigma_1}\exp\Big\{-\Big[(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}} \\ &+(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^\alpha\Big\}+\exp\{-\beta_{01}-\beta_{11}x_{i1}\}^{\frac{1}{\sigma_1}} \\ &t_{i1}^{\frac{1}{\sigma_1}-1}\frac{1}{\sigma_1}\exp[-(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1}}], \end{aligned}$$

$$\begin{aligned}\frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i2}} &= -\Big[(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}}+(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha-1}t_{i2}^{\frac{1}{\sigma_2\alpha}-1}\\ &\exp\{-\beta_{02}-\beta_{12}x_{i2}\}^{\frac{1}{\sigma_2\alpha}}\frac{1}{\sigma_2}\exp\Big\{-\Big[(\exp\{-\beta_{01}-\beta_{11}x_{i1}\}t_{i1})^{\frac{1}{\sigma_1\alpha}}\\ &+(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2\alpha}}\Big]^{\alpha}\Big\}+\exp\{-\beta_{02}-\beta_{12}x_{i2}\}^{\frac{1}{\sigma_2}}\\ &t_{i2}^{\frac{1}{\sigma_2}-1}\frac{1}{\sigma_2}\exp[-(\exp\{-\beta_{02}-\beta_{12}x_{i2}\}t_{i2})^{\frac{1}{\sigma_2}}],\end{aligned}$$

e

$$\frac{\partial G(t_{i1},t_{i2}|\boldsymbol{x}_i)}{\partial t_{i1}\partial t_{i2}}=g(t_{i1},t_{i2}|\boldsymbol{x}_i).$$

Tabela 1 - Estimativa de máxima verossimilhança para o modelo de regressão Kumaraswamy Weibull bivariado

Parâmetro	Estimativa	Erro Padrão	valor-p
β_{01}	2,5204	1,6159	0,1188
β_{11}	0,7075	0,4981	0,1555
β_{02}	1,7274	1,2462	0,1657
β_{12}	1,9164	0,3659	<0,0001
σ_1	1,6826	0,8100	-
σ_2	1,4891	0,6823	-
a	2,9146	3,1558	-
b	0,5782	0,1739	-
α	1,0000	0,3743	-

As estimativas dos parâmetros do modelo de regressão Kumaraswamy Weibull bivariado, os respectivos erros padrões e significâncias encontram-se na Tabela 1. Dos resultados, conclui-se que pacientes do sexo masculino têm a ocorrência da infecção 2 acelerada em relação aos pacientes do sexo feminino. Esse fato pode ser confirmando visualmente por meio da Figura 2. A estimativa do parâmetro de associação indica que os tempos de vida são independentes, o que confirma o resultado obtido na seção 2.5.1 por meio do coeficiente de correlação τ de Kendall.

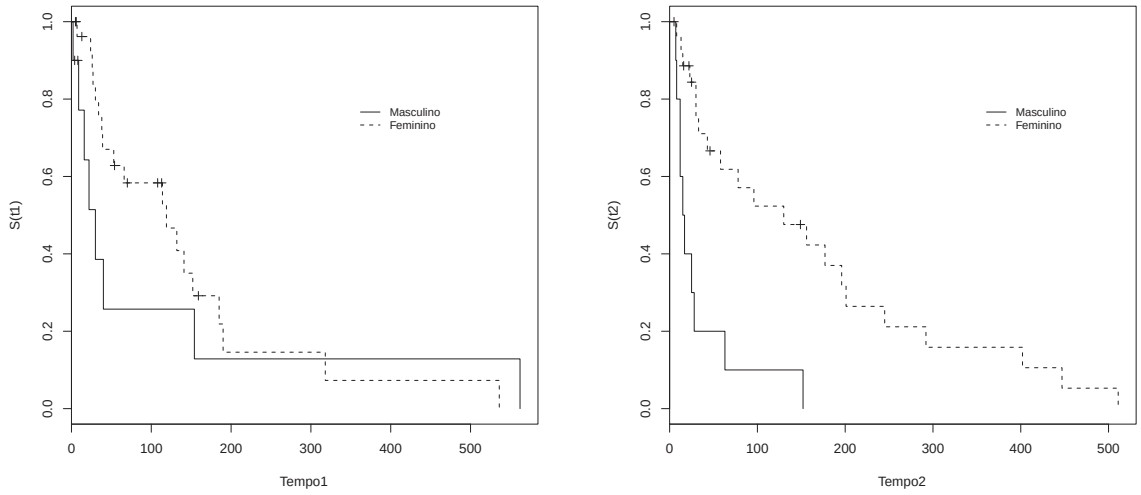

Figura 2 - Curvas de sobrevivência estimadas marginalmente por Kaplan-Meier para os dados de de insuficiência renal para cada evento por meio da covariável sexo

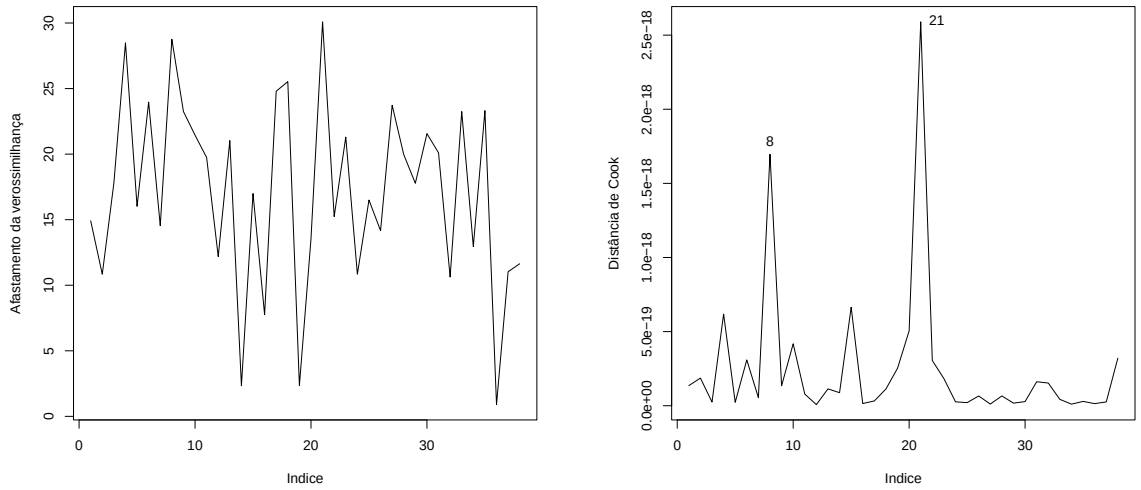

Figura 3 - Gráficos de medidas de Influência Global para o modelo de regressão Kumaraswamy Weibull bivariado para os dados de insuficiência renal. (a) Afastamento da Verossimilhança (b) Distância de Cook

2.5.3 Análise de Influência Global

Após a modelagem, é importante verificar se existem possíveis observações influenciando o ajuste do modelo de regressão Kumaraswamy Weibull bivariado. Para investigar esse fato, as medidas de Influência Global, isto é, o Afastamento da Verossimilhança ($LD_i(\boldsymbol{\Psi})$) e a Distância de Cook Generalizada ($GD_i(\boldsymbol{\Psi})$), foram calculadas como definidas na seção 2.4.1.

A observação #21 é a que mais se destaca das demais, como mostra a Figura 3. Esse indivíduo caracteriza-se por apresentar o maior tempo até a ocorrência da infecção 1 e um tempo mediano até a ocorrência da infecção 2.

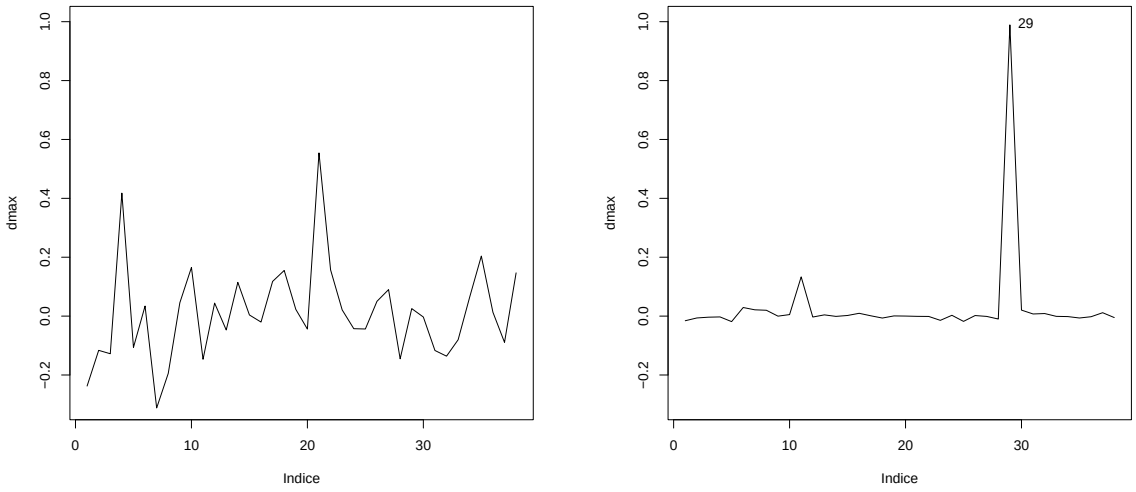

Figura 4 - Gráfico de medidas de influência do modelo de regressão Kumaraswamy Weibull bivariado considerando o esquema de perturbação de casos e da variável resposta 1 para os dados de insuficiência renal. (a) Influência Local $\boldsymbol{d}_{max}$ casos (b) Influência Local $\boldsymbol{d}_{max}(t_1)$

2.5.4 Análise de Influência Local

Ao considerar a teoria de Influência Local sob verossimilhança restrita desenvolvida na seção 2.4.3 para o modelo de regressão Kumaraswamy Weibull bivariado considerando os dados de insuficiência renal, foram realizados os esquemas de ponderação de

casos, perturbação nos tempos até a ocorrência da infecção 1, perturbação nos tempos até a ocorrência da infecção 2 e perturbação conjunta em ambos os tempos. O esquema de perturbação da covariável não foi considerado nesta análise, pois a covariável sexo é categórica.

Para os esquemas de perturbação citados, foram calculados os vetores $\boldsymbol{d}_{max}$, $\boldsymbol{d}_{max}(t_1)$, $\boldsymbol{d}_{max}(t_2)$ e $\boldsymbol{d}_{max}(t_1t_2)$, correspondentes as direções da maior curvatura, e os autovalores das curvaturas máximas, que são dados por: $C_{\boldsymbol{d}_{max}}(\boldsymbol{\Psi}) = 2,83$, $C_{\boldsymbol{d}_{max}}(t_1) = 1636,60$, $C_{\boldsymbol{d}_{max}}(t_2) = 209,34$ e $C_{\boldsymbol{d}_{max}}(t_1t_2) = 1677,50$, respectivamente. Considerando as medidas em questão, gráficos de influência foram construidos e encontram-se nas Figuras 4 a 7. Nesses gráficos nota-se que a observação #29, destacou-se com maior evidência, que representa um indivíduo com menor tempo até a infecção 1 e baixo tempo até a infecção 2.

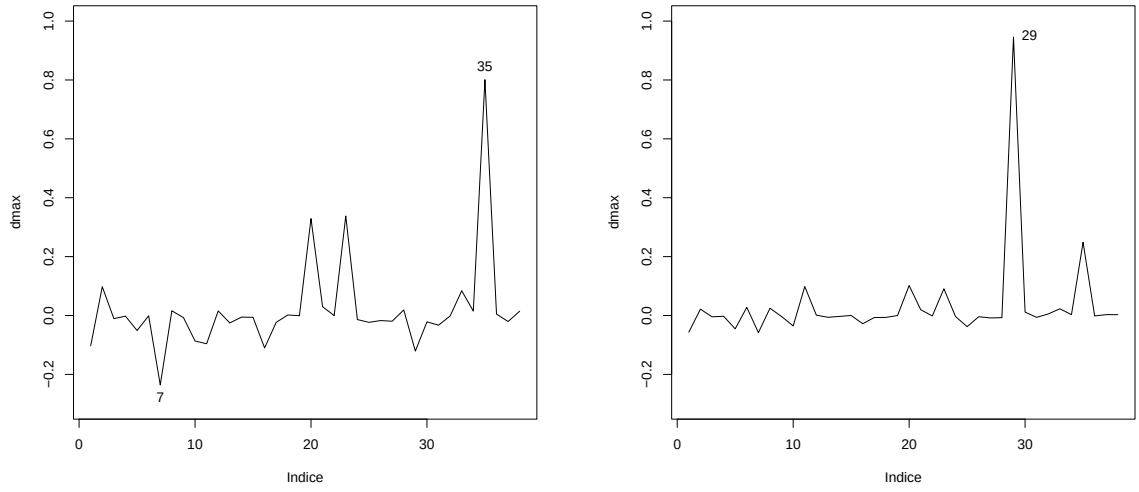

Figura 5 - Gráfico de medidas de influência do modelo de regressão Kumaraswamy Weibull bivariado considerando o esquema de perturbação da variável resposta 2 e a perturbação conjunta de ambas as variáveis respostas para os dados de insuficiência renal. (a) Influência Local $\boldsymbol{d}_{max}(t_2)$ (b) Influência Local $\boldsymbol{d}_{max}(t_1t_2)$

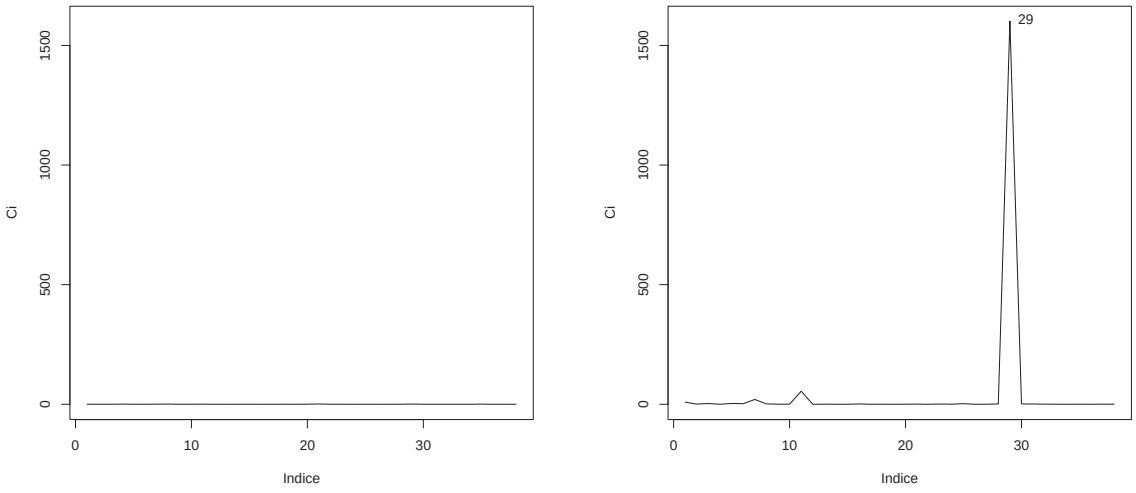

Figura 6 - Gráfico de medidas de influência do modelo de regressão Kumaraswamy Weibull bivariado considerando o esquema de perturbação de casos e da variável resposta 1 para os dados de insuficiência renal. (a) Influência Local Total C_i casos (b) Influência Local Total $C_{t_{i1}}$

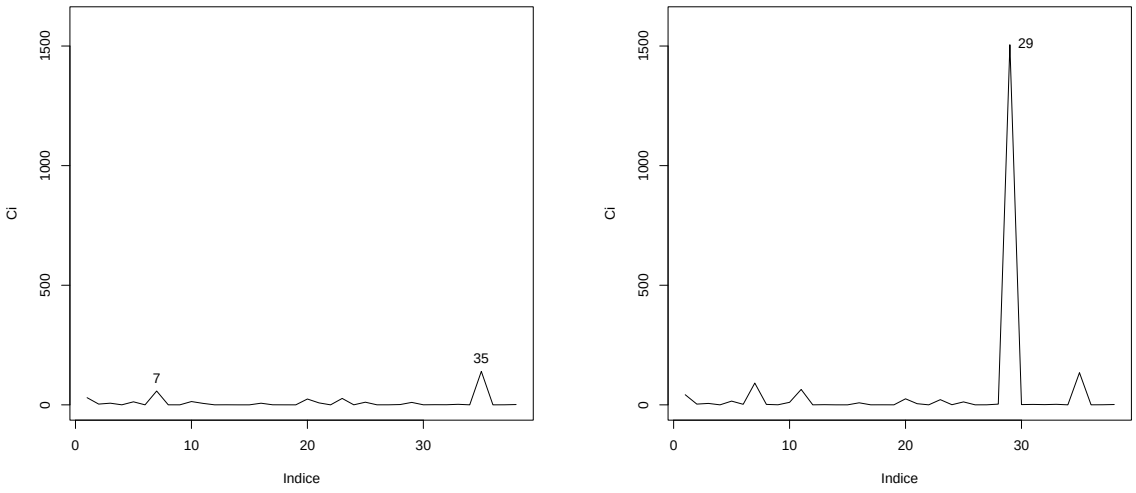

Figura 7 - Gráfico de medidas de influência do modelo de regressão Kumaraswamy Weibull bivariado considerando o esquema de perturbação da variável resposta 2 e a perturbação conjunta de ambas as variáveis respostas para os dados de insuficiência renal. (a) Influência Local Total $C_{t_{i2}}$ (b) Influência Local Total $C_{t_{i1}t_{i2}}$

2.5.5 Impacto das observações influentes

A análise de diagnóstico para os dados de insuficiência renal considerando o modelo de regressão Kumaraswamy Weibull bivariado, permite detectar as observações #21 e #29 como destacadas das demais, representando possíveis pontos influentes. A observação #21 é referente ao indivíduo do sexo masculino com maior tempo até a ocorrência da infecção 1 e um tempo mediano até a ocorrência da infecção 2 e a observação #29 é referente ao indivíduo do sexo masculino com menor tempo até a ocorrência da infecção 1 e baixo tempo até a ocorrência da infecção 2. Ambas as observações representam casos peculiares, mas não apresentam indícios de erro na coleta ou transcrição dos dados e, portanto, devem ser mantidas no conjunto de observações.

O impacto dessas observações sobre o modelo deve ser analisado para se avaliar sua sensibilidade e suas estimativas. Para realizar esta análise, novas estimativas para os parâmetros do modelo foram obtidas a partir de subamostras referentes à retirada dessas observações, individualmente e em grupo.

Considerando as subamostras obtidas a partir da exclusão individual e conjunta dos valores possivelmente influentes, as estimativas para os parâmetros, os respectivos p-valores e as mudanças relativas de cada parâmetro encontram-se na Tabela 2. A mudança relativa foi definida como $\boldsymbol{RC}_{\boldsymbol{\Psi}_j} = [(\hat{\boldsymbol{\Psi}}_j - \hat{\boldsymbol{\Psi}}_{j(I)})/\hat{\boldsymbol{\Psi}}_j]$, sendo (I) o índice referente as observações excluídas da amostra.

Na Tabela 2 pode-se observar mudanças expressivas nos valores estimados para os parâmetros, assim como mudanças no conjunto de parâmetros que se mostraram significativos no modelo. Apesar dessa sensibilidade no modelo, a inclusão ou exclusão dos pontos identificados não implica em mudanças na interpretação dos resultados, uma vez que não houve alteração no sinal do coeficiente referente à variável sexo, que indica que homens tendem a ter uma infecção de forma mais precoce.

Tabela 2 - Mudança relativa [RC], estimativas dos parâmetros, e correspondentes (p-valor)

Sub-amostra	$I-\{completo\}$	$I-\{21\}$	$I-\{29\}$	$I-\{21,29\}$
	[-]	[1,692]	[6,9591]	[7,8614]
β_{01}	2,5204	2,4777	2,3450	2,3222
	(0,1188)	(0,0846)	(0,1649)	(0,1258)
	[-]	[-67,7746]	[25,0070]	[-32,8636]
β_{11}	0,7075	1,1871	0,5306	0,9401
	(0,1555)	(0,0291)	(0,2141)	(0,0458)
	[-]	[-10,9615]	[29,7890]	[20,1219]
β_{02}	1,7274	1,9168	1,2128	1,3798
	(0,1657)	(0,1098)	(0,4268)	(0,3567)
	[-]	[-1,9063]	[-0,8373]	[-2,5333]
β_{12}	1,9164	1,9529	1,9324	1,9650
	(0,0000)	(0,0000)	(0,0000)	(0,0000)
	[-]	[9,1365]	[-4,1007]	[3,6082]
σ_1	1,6826	1,5289	1,7516	1,6219
	(-)	(-)	(-)	(-)
	[-]	[0,3287]	[-14,7761]	[-15,3607]
σ_2	1,4891	1,4842	1,7091	1,7178
	(-)	(-)	(-)	(-)
	[-]	[6,1817]	[-41,5784]	[-31,8750]
a	2,9146	2,7344	4,1264	3,8436
	(-)	(-)	(-)	(-)
	[-]	[-31,6078]	[5,2288]	[-21,1361]
b	0,5782	0,7609	0,5479	0,7004
	(-)	(-)	(-)	(-)
	[-]	[0,0000]	[0,0000]	[0,0000]
α	1,0000	1,0000	1,0000	1,0000
	(-)	(-)	(-)	(-)

2.5.6 Qualidade de ajuste

Com o objetivo de verificar a qualidade do ajuste do modelo proposto neste capítulo, foram construídos gráficos que representam as funções de sobrevivência estimada pelo método de Kaplan-Meier e as funções de sobrevivência marginais estimadas para o modelo de regressão Kumaraswamy Weibull bivariado. A Figura 8 ilustra as funções citadas sem considerar covariáveis e a Figura 9 ilustra a incorporação da covariável sexo nas funções citadas.

Por meio das Figuras 8 e 9 verifica-se um bom ajuste do modelo proposto, pois se observa que a curva do modelo de regressão Kumaraswamy Weibull bivariado acompanha o gráfico da função de sobrevivência estimada por Kaplan-Meier.

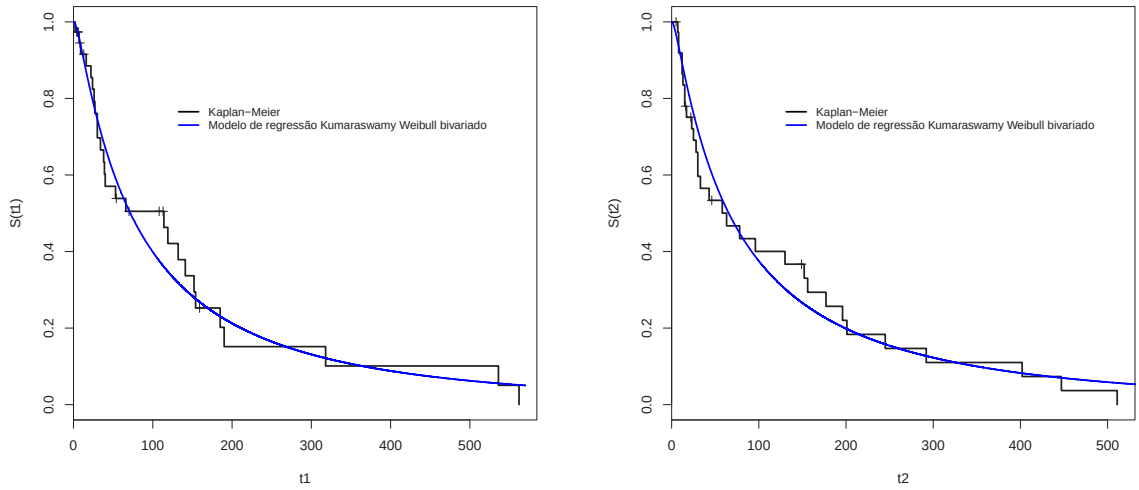

Figura 8 - Curvas de sobrevivência de Kaplan-Meier e função de sobrevivência estimada para os dados de pacientes com insuficiência renal. (a) função de sobrevivência marginal $S_1(t_1)$, e (b) função de sobrevivência marginal $S_2(t_2)$

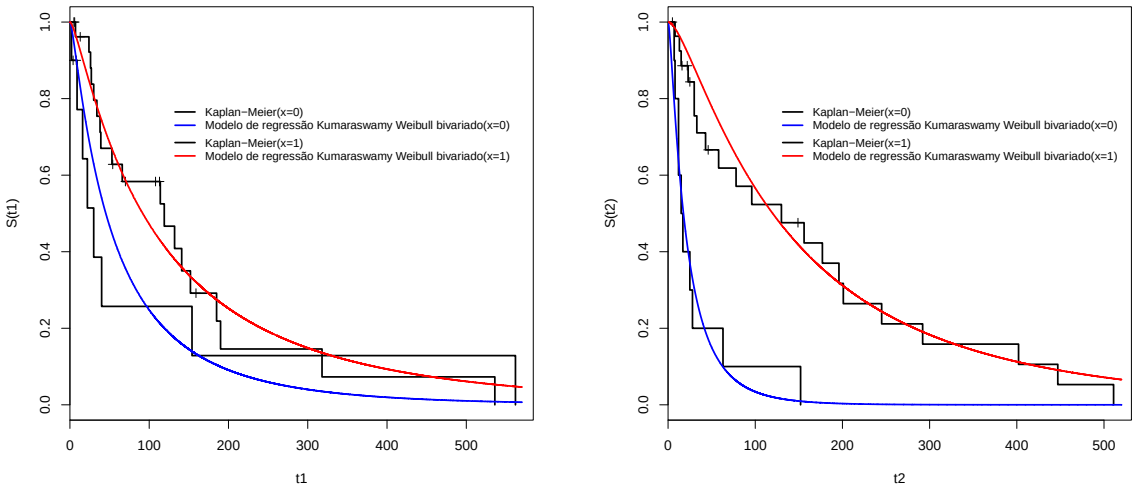

Figura 9 - Curvas de sobrevivência de Kaplan-Meier estratificadas por sexo e função de sobrevivência estimada para dados de pacientes com insuficiência renal. (a) função de sobrevivência marginal $S_1(t_1|\boldsymbol{x})$, e (b) função de sobrevivência marginal $S_2(t_2|\boldsymbol{x})$

2.6 Conclusões

Neste trabalho, o modelo de regressão Kumaraswamy Weibull bivariado foi proposto como extensão da distribuição Kumaraswamy Weibull tipo I. Seus parâmetros foram estimados utilizando uma metodologia de máxima verossimilhança sujeita a restrições lineares nos parâmetros e uma análise de sensibilidade foi conduzida para avaliar a robustez dos resultados obtidos.

A metodologia de estimação citada foi conduzida com o apoio do *software* R (R DEVELOPMENT CORE TEAM, 2009). Diversas funções implementadas no mesmo, assim como novas funções com os cálculos das derivadas parciais da função de verossimilhança, foram incluídas na programação utilizada. O processo de otimização mostrou-se sensível à escolha dos valores iniciais utilizados nos algorítmos. A forma adotada para a escolha desses valores iniciais para os parâmetros de regressão foi a obtenção de estimativas considerando modelos marginais para cada variável resposta. Para os outros parâmetros foram atribuídos os valores

$a = 1$ e $b = 1$, representando assim, como caso particular, a distribuição Weibull bivariada. Para o parâmetro α foi utilizada como referência a interpretação da associação indicada pelo valor calculado para o coeficiente τ - Kendall.

Utilizando as técnicas de Influência Global, Influência Local e Influência Local Total alguns pontos foram identificados como possivelmente influentes. Uma atenção especial foi dedicada a esses pontos para detectar possíveis erros na administração do conjunto de dados, mas essa possibilidade foi descartada. A robustez dos resultados obtidos foi verificada após reestimar os parâmetros do modelo considerando a exclusão individual e conjunta dos pontos identificados. Apesar de pequenas modificações nas estimativas, os resultados e suas interpretações não foram influenciados pelas observações apontadas como possíveis pontos influentes pelas técnicas de diagnóstico consideradas.

No conjunto de dados analisado não foi possível explorar, em sua totalidade, a capacidade do modelo de captar, nos parâmetros de regressão, o efeito da associação entre as variáveis resposta, pois essas puderam ser consideradas independentes. Essa indepêndencia entre as variáveis foi verificada inicialmente na análise exploratória dos dados e confirmada com a utilização do modelo que teve seu parâmetro α, referente a associação entre as variáveis, estimado com valor igual a um, indicando independência.

As funções e técnicas apresentadas neste capítulo permitem a utilização do modelo de regressão Kumaraswamy Weibull bivariado para conjuntos de dados de sobrevivência com duas variáveis resposta e covariáveis, bem como a realização de uma análise de sensibilidade para verificar a adequação das suposições adotadas e validar os resultados obtidos.

2.6.1 Propostas para trabalhos futuros

Como possíveis trabalhos futuros podem-se considerar os seguintes temas de pesquisa:

1. Propor um modelo de mistura como uma extensão do modelo de regressão Kumaraswamy Weibull bivariado.

2. Considerar outras distribuições de probabilidade bivariadas, originando novos modelos de regressão Kumaraswamy para dados bivariados.

Referências

BLOCK, H.W, BASU, A.P. A continuos bivariate exponencial extension **Journal of American Statistical Association** Alexandria, v.69, p. 1031-1037, 1974.

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2nd ed. Pacific Grove: Thomson Learning, 2002. 660 p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Edgard Blücher, 2006. 392 p.

CORDEIRO, G.M; CASTRO, M. A new family of generalized distributions. **Journal of Statistical Computation and Simulation**, New York, v.0, p. 1-17, 2010.

CORDEIRO, G.M; ORTEGA, E.M.M., NADARAJAH, S. The Kumaraswamy Weibull distributions with application to failure data. **Journal of Franklin Institute**, New York, v.347, p. 1399-1429, 2010.

COOK, R.D. Detection of influential observations in linear regression. **Technometrics**, Alexandria, v. 19, p. 15-18, 1977.

COOK, R. D.; WEISBERG, S. **Residuals and influence in regression**. New York: Chapman and Hill, 1982. 230 p.

COOK, R.D. Assement of local influence (with discussion). **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 48, n. 2, p. 133-169, 1986.

COOK, R. D.; PEÑA, D.; WEISBERG, S. The likelihood displacement: a unifying principle for influence. **Communications in Statistics: Part Theory and Methods**, New York, v. 17, n. 3, p. 623-640, 1988.

COX, D.R.; HINKLEY, D.V. **Theoretical Statistics**. Londom: Chapman and Hall, 1974. 511 p.

COX, D.R.; OAKES, D. **Analysis of survival data**. London: Chapman and Hall, 1984. 201 p.

ESCOBAR, L.A.; MEEKER, W.Q. Assessing influence in regression analysis with censored data. **Biometrics**, Washington, v. 48, n.2, p. 507-528, 1992.

EUGENE, N.; LEE, C.; FAMOYE, F. Beta-normal distribution and its applications. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 31, p. 497-512, 2002.

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.

FREUND, J. E. Bivariate Extension of the Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 56, p. 971-977, 1961.

GOMES, E. M. C. **Análise de Sensibilidade e resíduos em modelos de regressão com respostas bivariadas por meio de cópulas.** 2007. 103p. Dissertação (Mestre em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

GU, H; FUNG, W.K. Local influence for the restricted likelihood with applications. **Sankhya: The Indian Journal of Statistical**, Indian, v. 63, pt. 2, p. 250-259, 2001.

GUMBEL, E. J. Bivariate Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 55, p. 698-707, 1960.

HE, W.; LAWLESS, J. F. Bivariate location-scale models for regression analysis, with applications to lifetime data. **Journal of the Royal Statistical Society**, London, v. 67, n. 1, p. 63-78, 2005.

HOUGAARD, P. A class of multivariate failure time distributions. **Biometrika**, Great Britain, v. 73, p. 671-678, 1986.

HOUGAARD, P. Fitting a multivariate failure time distribution. **IEEE Transactions on Reliability**, New York, v. 38, p. 444-448, 1989.

JOHNSON, R. A.; EVANS, J. W.; GREEN, D. W. Some bivariate distributions for modeling the strength properties of lumber. **United States Department of Agriculture**, Washington, FPL-RP-575, 1999.

JONES, M.C. Kumaraswamy´s distribution: a beta-type distribution with some tractability advantages. **Statistical Methodology**, London, v. 6, p. 70-81, 2009.

KALBFLEISCH, J.D.; PRENTICE, R.L. **The statistical analysis of failure time data**. 2nd ed. New York: John Wiley, 2002. 439 p.

Kumaraswamy, P. A generalized probability density function for doublebounded random processes. **Journal of Hydrology**, Amsterdam, v. 46, p. 79-88, 1980.

KWAN, C. W; FUNG, W. K. Assessing local influence for specific restricted likelihood: application to factor analysis. **Psychometrika**, New York, v. 63, n. 1, p. 35-46, 1998.

LANGE, K. **Numerical analysis for statisticians**. New York: Springer, 1999. 356 p.

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

LEE. E. T. **Statistical models and for survival data analysis**, 2nd ed., New York: Wiley, 1992. 482 p.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, Washington, v. 54, n. 2, p. 570-582, 1998.

LU, J.C.; BHATTACHARYYA, G.K. Some new constructions of bivariate Weibull models. **Annals of the Institute of Statistical Mathematics**, Heidelberg, v.42, p. 543-559, 1990.

MARSHALL, A. W.; OLKIN, I. A Multivariate Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 62, p. 30-44, 1967.

MCGILCHRIST, C.A.; AISBETT, C.W. Regression with Frailty in Survival Analysis. **Biometrics**, Washington, v. 47, p. 461-466, 1991.

OLIVEIRA, L. P. **Estudo da extensão do modelo bivariado exponencial de Marshall e Olkin para dados de confiabilidade.** 2001. 166p. Dissertação (Mestrado em Estatística)- Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 2001.

ORTEGA, E. M. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. **Computational Statistics and Data Analysis**, New York, 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 Operation Reserch Transactions**, Catalunya, v. 30, n. 2, p. 171-192, 2006.

ORTEGA, E. M. M. ; CANCHO, V. G.; PAULA, G. A. Generalized log-gamma regression models with cure fraction. **Lifetime Data Analysis**, Boston, v. 15, p. 79-106, 2009a.

PAULA, G.; CYSNEIROS, F. J. A. Local influence under parameter constraints. **Communications in Statistics: Theory and Methods**, New York, v.88, p. 1-23, 2009.

R Development Core Team (2009). **R**: A language and environment for statistical computing. Disponível em:¡http://www.R-project.org¿. Acesso em: 17 maio 2011.

RYU, K. An Extention of Marshall and Olkin's Bivariate Exponential Distribution. **Journal of the American Statistical Association**, Alexandria, v. 88, p. 1458-1465, 1993.

TARUMOTO, M. H. **Um modelo Weibull bivariado para riscos competitivos.** 2001. 154p. Tese (Doutorado em Matemática Aplicada)- Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 2001.

XIE, F.; WEI, B. Diagnostics analysis for log-Birnbaum-Saunders regression models. **Computational Statistics and Data Analysis**, Amsterdam, v. 51, p.4692-4706, 2007.

ZHU, H.; ZHANG, H. A diagnostic procedure based on local influence. **Biometrika**, Cambridge, v. 91, n. 3, p. 579-589, 2004.

3 MODELO DE REGRESSÃO COM FRAÇÃO DE CURA POR MEIO DE CÓPULAS

Resumo

O modelo proposto neste capítulo é uma extensão dos trabalhos de Chatterjee e Shih (2001) e Wienke et al. (2003). O objetivo deste modelo é descrever a correlação entre os tempos bivariados por meio de cópulas, uma possível fração de indivíduos curados individual e conjuntamente, e captar a presença dessas quantidades nos parâmetros de regressão. Os parâmetros do modelo foram estimados por meio do método da função barreira adaptada (LANGE, 1999), que é uma combinação do método barreira logaritmo com o algoritmo EM. Uma análise de sensibilidade considerando-se as metodologias de Influência Global, Influência Local e Influência Local Total de um indivíduo foi implementada. Como ilustração, um conjunto de dados de retinopotia diabética foi analisado sob o modelo de regressão com fração de cura para dados bivariados por meio de cópulas.

Palavras-chave: Fração de cura; Verossimilhança sujeita a restricão nos parâmetros; Modelos de regressão; Dados bivariados e censurados; Cópulas Arquimedianas; Análise de sensibilidade

Abstract

The model proposed in this chapter is an extension the work of Chatterjee e Shih (2001) e Wienke et al. (2003). The objective of this model is to describe the bivariate correlation between the times through copulas, a possible fraction of cured individuals jointly and marginally, and the capture this effects in the regression parameters. The model parameters were estimated by adapted barrier function method (LANGE, 1999), which is a combination of the logarithm barrier method with the EM algorithm. A sensitivity analysis considering the methodology Global Influence, Local Influence and Total Local Influence of an individual was implemented. As an illustration, a data set of diabetic retinopathy was analyzed under the regression model with a cured fraction for bivariate data through copulas.

Keywords: Cured fraction; Archimedean Copulas; Regression models; Bivariate data and

censored; Likelihood subject to restriction on the parameters; Sensitivity analysis

3.1 Introdução

Em estudos de análise de sobrevivência, é necessário, muitas vezes, considerar como variável(is) resposta o tempo transcorrido até a ocorrência de um ou mais eventos de interesse. Estudos clínicos, como The Diabetic Retinopathy Study, avaliam pacientes diabéticos submetidos a terapia de laser em um dos olhos escolhido aleatoriamente e consideram como evento de interesse principal o tempo até a perda da acuidade visual para ambos os olhos. Nesse caso, os eventos de interesse no estudo são o tempo até a perda da acuidade visual para o olho com tratamento e o tempo até a perda da acuidade visual para o olho sem tratamento, podendo-se supor que existe uma associação entre os tempos até a perda da visão em um determinado paciente.

Vários enfoques têm sido considerados para modelar dados de sobrevivência bivariados. Neste capítulo, dar-se-á atenção para a metodologia de cópulas, proposta inicialmente por Clayton (1978) a partir de um modelo de associação bivariado para análise de sobrevivência. Na literatura sobre teoria de cópulas existe uma variedade de famílias, como, por exemplo, as classes de cópulas de Marshall-Olkin, as Elípticas e as Arquimedianas. Neste trabalho será considerada a família de cópulas Arquimedianas, relacionada aos modelos de sobrevivência bivariados, mais especificamente a cópula de Clayton, que tem apresentado bons resultados em trabalhos publicados, por exemplo, Wienke at al (2006) e Gomes (2007).

No contexto de cópulas, alguns trabalhos podem ser consultados, como, por exemplo, Hougaard (1989) e Shih e Louis (1995) para estudar procedimentos de estimação; Tibaldi (2004) e Núñez (2005) na abordagem de modelos marginais de riscos proporcionais; e He e Lawless (2005) e Gomes (2007) para modelos marginais de locação-escala.

Nos modelos citados, pressupõe-se que todos os indivíduos em estudo irão desenvolver o evento de interesse definido no início do experimento. Durante o estudo, alguns indivíduos poderão vir a falhar ou ser censurados. Às vezes, acontece que para uma proporção de indivíduos o evento de interesse não ocorrerá. Esses indivíduos, freqüentemente, são conhecidos como imunes, curados ou não suscetíveis. Nesse caso, considerar os modelos de

sobrevivência usuais, que assumem que a função de sobrevivência converge para zero quando a variável tempo tende a infinito (função de sobrevivência própria), podem não ser adequados. Para modelar esse tipo de dados, modelos com fração de cura são mais apropriados.

A modelagem de fração de cura para o caso univariado pode ser abordada seguindo a metodologia introduzida por Berkson e Gage, (1952) que considera a construção de uma função de sobrevivência populacional na forma de mistura. O modelo proposto por Berkson e Gage (1952) foi estendido para o caso bivariado por Chatterjee e Shih (2001), com base na teoria de cópulas e considerando o procedimento de estimação dos parâmetros em duas etapas. Wienke et al. (2003) consideram o procedimento de estimação em uma etapa e o modelo de fragilidade gama correlacionado. Mais recentemente, Wienke et al. (2006) aplicam três modelos de fragilidade correlacionados. Para dar continuidade nesses modelos, o objetivo principal deste trabalho é propor a inclusão de variáveis regressoras nos modelos citados anteriormente, dando origem ao modelo de regressão bivariado com fração de cura. Para estimar os parâmetros do modelo proposto utilizou-se o método de máxima verossimilhança sujeito às restrições nos parâmetros. O método implementado utiliza a função barreira adaptada, que é uma combinação do método barreira com o algoritmo EM. Para maiores detalhes sobre esse método ver, por exemplo, Lange (1999).

Quando se ajusta um modelo a um conjunto de dados, é importante estudar a robustez dos resultados obtidos com relação à presença de observações extremas ou observações influentes que podem causar alterações nos resultados das estimativas dos parâmetros do modelo. Para detectar observações influentes nas estimativas dos parâmetros, podem ser consideradas metodologias de Influência Global, Local e Local Total. A primeira metodologia é baseada na deleção de casos, proposta por Cook (1977), enquanto a segunda metodologia, proposta por Cook (1986) é baseada em pequenas perturbações nos dados ou no modelo. A terceira metodologia é desenvolvida por Lesaffre e Verbeke (1998), a medida de Influência Local Total.

Este trabalho pesquisa a utilização das medidas de influência citadas considerando restrições no espaço paramétrico associado ao modelo, com base nos trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009) que utilizam Influência Local na estrutura de verossimilhança restrita.

O trabalho está organizado da seguinte forma: na seção 3.2 é apresentada uma revisão do modelo com fração de cura para o caso univariado. Na seção 3.3 descreve-se o modelo bivariado considerando a metodologia de cópulas. Na seção 3.4 é proposto o modelo de regressão bivariado com fração de cura. A seção 3.5 abrange uma descrição da análise de sensibilidade do modelo proposto baseada nas Teorias de Influência Global, Local e Local Total sob o enfoque da verossimilhança sujeita a restrições nos parâmetros. Na seção 3.6 é feita a aplicação do modelo proposto. Para finalizar, a seção 3.7 relata as principais conclusões e o direcionamento da continuidade deste trabalho.

3.2 Fração de cura univariada seguindo abordagem de Berkson e Gage

Os modelos tradicionais para análise de sobrevivência, partem do pressuposto que os indivíduos experimentarão o evento de interesse delimitado no estudo, sendo que alguns deles, ao longo do estudo, poderão vir a falhar ou ser censurados.

Entretanto, há situações que, para uma proporção de indivíduos, o evento de interesse não ocorrerá. Esses indivíduos são, frequentemente, conhecidos como imunes, curados ou não-suscetíveis. Algumas pesquisas, por exemplo, têm o interesse em analizar a recorrência de doenças. Nesse caso, muitos indivíduos nunca desenvolverão a recorrência, portanto, existe uma fração de indivíduos curados no estudo.

Na literatura, a existência de uma proporção de indivíduos curados é caracterizada pelo fato de que a função de sobrevivência não converge para zero quando o tempo aumenta, conhecida como função de sobrevivência imprópria. A presença de uma proporção de indivíduos curados em determinado conjunto de dados pode ser identificada por meio de um gráfico da função de sobrevivência empírica estimada pelo método de Kaplan-Meier (LAWLESS, 2003), que deve apresentar a cauda à direita em um nível constante acima de zero por um período considerado suficientemente grande, como ilustra a Figura 10. Maiores detalhes encontram-se em Maller e Zhou (1996).

Ao considerar a modelagem de fração de cura seguindo a metodologia introduzida por Berkson e Gage (1952), a função de sobrevivência populacional é construída na forma de mistura e conhecida como função de sobrevivência populacional imprópria.

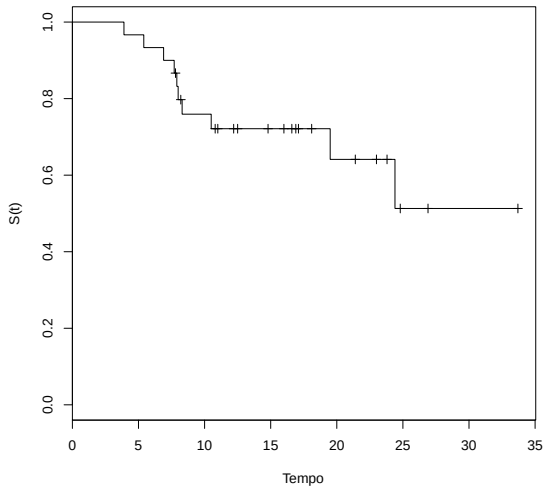

Figura 10 - Gráfico de função de sobrevivência imprópria estimada pelo método de Kaplan-Meier

Modelos com fração de cura dividem a população em duas sub-populações: indivíduos curados com probabilidade $(1-\phi)$ e indivíduos suscetíveis ao evento de interesse com função de sobrevivência própria com probabilidade ϕ, $\phi \in (0,1)$. Define-se a função de sobrevivência populacional em forma de mistura representada por:

$$S_{pop}(t) = (1-\phi) + \phi S(t), \tag{27}$$

que possui as seguintes propriedades: $\lim_{t \longrightarrow \infty} S_{pop}(t) = (1-\phi)$ e $\lim_{t \longrightarrow 0} S_{pop}(t) = 1$.

Maller e Zhou (1996) consideram uma variável aleatória Bernoulli associada a cada indivíduo i para indicar se o i-ésimo indivíduo é suscetível ou curado, isto é,

$$V_i = \begin{cases} 1 & \text{se o indivíduo é suscetível} \\ 0 & \text{se o indivíduo é curado.} \end{cases}$$

Associa-se, portanto, a cada variável aleatória V_i uma probabilidade $P(V_i = 1) = \phi$, que é a probabilidade de um indivíduo ser suscetível ao evento de interesse do estudo, e uma probabilidade $P(V_i = 0) = 1 - \phi$, que é a probabilidade de um indivíduo ser curado.

Nas próximas seções, são expostos o modelo bivariado e a extensão do modelo com fração de cura seguindo a abordagem de Berkson e Gage (1952) para o caso bivariado.

3.3 Modelo Bivariado

No capítulo 2 foram abordadas a utilidade e aplicabilidade de modelos bivariados e discutida a modelagem de dados bivariados considerando-se distribuições bivariadas, amplamente estudadas e formuladas ao longo do tempo.

Uma outra abordagem para modelar dados bivariados é a metodologia de cópulas. Segundo Nelsen (1986), cópulas fornecem um meio de relacionar funções de distribuições multivariadas a partir de suas funções de distribuição marginais. Os estudos que utilizam a abordagem de cópulas iniciam-se com Clayton (1978), cujo modelo de associação bivariado para análise de sobrevivência, apesar de não mencionar o conceito de cópulas, utiliza-o implicitamente.

Alguns pesquisadores estudaram diferentes procedimentos de estimação dos parâmetros ao usar cópulas. Hougaard (1989), por exemplo, considera, em um primeiro passo, a estimação dos parâmetros de cada marginal usando o estimador de Nelson para a função taxa de falha acumulada, ignorando a dependência entre os tempos de falha. Em um segundo passo, adota um modelo de cópula em que o parâmetro de dependência é estimado considerando as distribuições marginais fixas. Similarmente, Shih e Louis (1995) também consideram o procedimento de duas etapas para estimar os parâmetros de modelos paramétricos e semi-paramétricos.

No contexto de cópulas, Tibaldi (2004) dá ênfase à cópula de Plackett e utiliza os modelos marginais de riscos proporcionais e Nuñez (2005) considera cópulas Arquimedianas para modelar a estrutura de dependência entre os tempos de falha por meio da abordagem bayesiana. Entretanto, He e Lawless (2005) utilizam os modelos marginais de locação-escala e Gomes (2007), seguindo essa abordagem, aplica uma análise de sensibilidade e resíduos para verificar o ajuste de modelos.

Neste capítulo, também são utilizados modelos marginais de locação-escala e a metodologia de cópulas para modelar a estrutura de dependência entre os dois tempos de falha de cada indivíduo. Na Seção 3.3.1 é apresentada uma revisão de Cópulas e na Seção 3.3.2 encontra-se o modelo de regressão bivariado obtido por meio da cópula de Clayton. Essas seções são essenciais para a compreensão da estrutura do modelo de regressão bivariado com fração de cura desenvolvido na seção 3.4.

3.3.1 Cópula

Cópulas são funções que ligam funções de distribuição multivariadas com suas funções marginais, ou, ainda, são funções de distribuição multivariadas cujas marginais são uniformes no intervalo $[0, 1]$. A definição de cópulas por meio de um dos principais teoremas de cópulas, o Teorema de Sklar, é dada a seguir.

Teorema 1.1 Seja F uma função de distribuição k-dimensional com marginais $F_1, \ldots, F_k$. Então, existe uma cópula C_α tal que para todo $(t_1, t_2, \ldots, t_k)$ em $\mathbb{R}^k$,

$$F(t_1, t_2, \ldots, t_k) = C_\alpha(F_1(t_1), \ldots, F_k(t_k)).$$

Se $F_1, \ldots, F_k$ são todas contínuas, então C_α é única, em que α é o parâmetro de associação. Caso contrário, C_α é determinada unicamente na imagem $F_1 \times \ldots \times$ imagem F_k. Inversamente, se C_α é uma cópula e $F_1, \ldots, F_k$ são funções de distribuição, então a função F definida é uma função de distribuição k-dimensional com marginais $F_1, \ldots, F_k$.

Algumas propriedades importantes de cópulas são:

(i) $T_1, T_2, \ldots, T_k$ são independentes se e somente se $C_\alpha(F_1(T_1), \ldots, F_k(T_k)) = \prod_{j=1}^{k} F_j(T_j)$;

(ii) a estrutura de dependência de cópulas é invariante a transformações contínuas e crescente das marginais, ou seja, se $(T_1, T_2, \ldots, T_k)$ tem cópula C_α e $G_1, G_2, \ldots, G_k$ são funções contínuas e crescentes, então $(G(T_1), G(T_2), \ldots, G(T_k))$ também tem cópula C_α;

(iii) elas são uniformemente contínuas e têm derivadas parciais de primeira ordem limitadas.

Na literatura existe uma variedade de famílias de cópulas, mas neste trabalho dar-se á ênfase à família de cópulas Arquimedianas relacionada aos modelos de sobrevivência bivariados e aos modelos multiplicativos de fragilidade.

A definição de cópulas Arquimedianas é dada da seguinte forma: C_α é considerada uma cópula Arquimediana se existir uma função convexa ψ: $[0, \infty] \longmapsto [0, 1]$, com $\psi(0) = 1, \psi^{'} < 0, \psi^{''} > 0$, tal que a cópula C_α é expressa como

$$C_\alpha(u_1, u_2) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2)),$$

para todo $(u_1, u_2) \in [0, 1]^2$ e $\alpha \in \mathcal{A}$, sendo $\mathcal{A}$ o espaço paramétrico que α irá assumir dependendo da cópula. A função de distribuição bivariada com marginais contínuas é considerada

uma cópula Arquimediana se sua cópula geradora é uma cópula Arquimediana. Logo, a função ψ^{-1} é chamada de função geradora. Maiores detalhes podem ser encontrados em Núñez (2005).

Na família de cópulas Arquimedianas existem três cópulas muito utilizadas na literatura de análise de sobrevivência: a cópula de Clayton, a cópula de Frank e o modelo de fragilidade estável positiva. Essas cópulas têm, em comum, o fato de que o gerador ψ^{-1} é uma função derivável que tem um único parâmetro $\alpha \in \mathcal{A}$ e a possibilidade de obter facilmente a medida de associação τ de Kendall, o que as torna bastante úteis.

Neste trabalho, será considerada a cópula de Clayton com duas variáveis resposta, ou seja, $k = 2$. Essa cópula também é conhecida como modelo de fragilidade gama compartilhado, pois pressupõe que todos os indivíduos compartilham dos mesmos fatores não observáveis.

A função de sobrevivência conjunta de acordo com a família de Clayton (1978) tem a seguinte forma:

$$S(t_1, t_2) = [S(t_1)^{-\alpha} + S(t_2)^{-\alpha} - 1]^{-1/\alpha}, \tag{28}$$

em que $\psi(u) = (1+\alpha u)^{-1/\alpha}$ corresponde à transformação de Laplace da distribuição gama com ambos parâmetros iguais a α^{-1}. $S(t_1)$ e $S(t_2)$ são as funções de sobrevivência marginais de T_1 e T_2, respectivamente. T_1 e T_2 são positivamente associados quando $\alpha > 0$ e são independentes quando $\alpha \longrightarrow 0$ obtendo-se $S(t_1, t_2) = S(t_1)S(t_2)$. O coeficiente τ de Kendall para essa família associado ao parâmetro α é dado por:

$$\tau_\alpha = \frac{\alpha}{\alpha + 2}. \tag{29}$$

3.3.2 Modelo de Regressão Bivariado

Para se verificar como variáveis explicativas podem afetar o tempo de sobrevivência, são considerados dois eventos de interesse, $k = 1, 2$. Considera-se $\boldsymbol{x} = (x_0, x_1, x_2, \ldots, x_p)^T$ o vetor de covariáveis associado às variáveis respostas transformadas $Y_k = \log(T_k)$.

O modelo de locação-escala é representado por:

$$\mathrm{Y_k} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{k}} + \sigma_{\mathrm{k}}\mathrm{Z_k}, \tag{30}$$

em que $Y_k = \log(T_k)$ é o logaritmo do tempo de falha do k-ésimo evento de interesse e, tem distribuição pertencente à família de distribuições que se caracteriza pelo fato de ter um parâmetro de locação $\boldsymbol{x}^T\boldsymbol{\beta}_k$ e um parâmetro de escala $\sigma_k > 0$. Considera-se $\boldsymbol{x}$ o vetor de covariáveis, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})^T$ o vetor de parâmetros desconhecidos e Z_k o erro aleatório. Para maiores detalhes, ver He e Lawless (2005), Gomes (2007) e Barriga et al. (2010).

Uma característica importante desse modelo é ser log-linear para T_k. Logo, é um modelo de regressão linear para Y_k. A densidade de Y_k, dado $\boldsymbol{x}$, é representada por $f(y_k|\boldsymbol{x})$ e a correspondente função de sobrevivência por $S(y_k|\boldsymbol{x})$.

Ao especificar o modelo de locação-escala para Y_k, a função de sobrevivência para o modelo de regressão log-linear bivariado com p covariáveis é representada por meio da cópula de Clayton:

$$S(y_1, y_2|\boldsymbol{x}) = [S(y_1|\boldsymbol{x})^{-\alpha} + S(y_2|\boldsymbol{x})^{-\alpha} - 1]^{-1/\alpha}, \tag{31}$$

como definida na equação (28). Quando $\alpha \longrightarrow 0$ tem-se independência entre Y_1 e Y_2, obtendo-se modelos de regressão log-linear univariados.

Algumas importantes características permitem grande flexibilidade na modelagem, como, por exemplo, a possibilidade de considerar que as funções marginais referentes a (31) podem ser quaisquer funções de sobrevivência, não necessariamente iguais, e que o vetor de covariáveis correspondente a Y_1 e Y_2 pode ser o mesmo ou diferente para ambas as respostas.

Uma vez definido o modelo matemático, a próxima etapa é baseada nos procedimentos de estimação. Esses procedimentos têm o objetivo de obter estimativas para os parâmetros do modelo a partir de uma amostra. O método de máxima verossimilhança para o modelo de regressão bivariado será descrito baseado nos trabalhos de Lawless (2003) e He e Lawless (2005).

3.3.3 Inferência para o modelo bivariado

No caso de modelos bivariados, utiliza-se o método de máxima verossimilhança considerando-se uma modificação na função de máxima verossimilhança usual. Como existem duas respostas, a função de verossimilhança será formada por quatro possíveis combinações de censuras nos dados. Essas combinações representam observações referente à falha dos dois

eventos de interesse, falha no evento 1 e censura no evento 2, censura no evento 1 e falha no evento 2 e censura em ambos os eventos de interesse.

Em uma amostra observada $(y_{1k}, \delta_{1k}, \boldsymbol{x}_{1k}), \ldots, (y_{nk}, \delta_{nk}, \boldsymbol{x}_{nk})$, considera-se y_{ik} o logaritmo do tempo do k-ésimo evento de interesse do i-ésimo indivíduo, δ_{ik} a respectiva variável indicadora de censura e $\boldsymbol{x}_{ik}$ o vetor de covariáveis do k-ésimo evento de interesse associado ao i-ésimo indivíduo, em que $k = 1, 2$ e $i = 1, 2, \ldots, n$. Ao utilizar a função de verossimilhança descrita em Lawless (2003), tem-se para o modelo bivariado descrito em (30) a seguinte função de verossimilhança:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2)^{\delta_{i1}\delta_{i2}} \times \left[\frac{-\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial y_{i1}}\right]^{\delta_{i1}(1-\delta_{i2})} \times \\ &\quad \left[\frac{-\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial y_{i2}}\right]^{(1-\delta_{i1})\delta_{i2}} \times S(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2)^{(1-\delta_{i1})(1-\delta_{i2})}, \end{aligned} \tag{32}$$

em que $S(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2)$ é a função de sobrevivência bivariada obtida por meio de uma cópula, $f(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{(-1)^2\partial^2 S(y_{i1}, y_{i2}|\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial y_{i1}\partial y_{i2}}$ é a função densidade conjunta de (y_{i1}, y_{i2}), $\boldsymbol{\theta} = (\alpha, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ é o vetor de parâmetros desconhecidos, sendo que $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k)^T$ e $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ para $k = 1, 2$.

Os estimadores de máxima verossimilhança para o vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$ são os valores de $\boldsymbol{\theta}$ que maximizam L($\boldsymbol{\theta}$), ou de forma equivalente, o logaritmo da função de verossimilhança $l(\boldsymbol{\theta})$. Assim, eles são encontrados resolvendo-se o seguinte sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

Para fazer inferências dos parâmetros, intervalos de confiança e testes de hipósteses. Utiliza-se as distribuições assintótica dos estimadores de máxima verossimilhança.

Como indicado na seção 3.2, existem situações que se caracterizam por apresentar uma significativa fração de curados, ou seja, indivíduos que não apresentam o evento de interesse mesmo após um longo período de acompanhamento. Diante dessa particularidade, o objetivo principal deste capítulo é, considerando a presença de uma proporção de indivíduos curados em cada evento de interesse no modelo de regressão bivariado apresentado na seção 3.3.2, propor o modelo de regressão bivariado com fração de cura.

3.4 Modelo de regressão com fração de cura para dados bivariados por meio de cópulas

A idéia da metodologia de fração de cura univariada (descrita na seção 3.2) é utilizada para identificar as sub-populações no modelo de regressão bivariado. Considera-se uma variável indicadora V_k associada a cada indivíduo i para indicar se o k-ésimo evento de interesse do i-ésimo indivíduo é suscetível ou curado, $k = 1, 2$, isto é,

$$V_k = \begin{cases} 1 & \text{se o indivíduo no } k\text{-ésimo evento é suscetível} \\ 0 & \text{se o indivíduo no } k\text{-ésimo evento é curado.} \end{cases}$$

Como consequência, são definidas as seguintes probabilidades de suscetibilidade:

(i) Seja $\phi_{11} = P(V_1 = 1, V_2 = 1)$ indicando a probabilidade de o indivíduo ser suscetível aos dois eventos de interesse;

(ii) Seja $\phi_{10} = P(V_1 = 1, V_2 = 0)$, indicando a probabilidade de o indivíduo ser suscetível ao evento 1 e não ser suscetível ao evento 2;

(ii) Seja $\phi_{01} = P(V_1 = 0, V_2 = 1)$ indicando a probabilidade de o indivíduo não ser suscetível ao evento 1 e ser suscetível ao evento 2;

(iv) Seja $\phi_{00} = P(V_1 = 0, V_2 = 0)$ indicando a probabilidade de o indivíduo não ser suscetível a nenhum dos eventos.

Ao considerar a idéia da função de sobrevivência populacional construída na forma de mistura, Wienke et al. (2006) escrevem a função de sobrevivência bivariada populacional na forma de mistura considerando os tempos de sobrevivência. Neste capítulo, propõe-se uma extensão dessa função de sobrevivência bivariada populacional pela inclusão de um vetor de covariáveis. Seja $\boldsymbol{x} = (x_0, x_1, x_2, \ldots, x_p)^T$ o vetor de covariáveis associado às variáveis resposta $Y_k = \log(T_k)$, $k = 1, 2$ tem-se o modelo de locação-escala dado por (30). Para especificar a estrutura de dependência entre o logaritmo dos tempos de falha de dois eventos suscetíveis é considerado o modelo de Clayton definido pela equação (31). Sendo assim, a função de sobrevivência bivariada populacional seguindo Wienke et al. (2006) é expressa pela equação:

$$S_{pop}(y_1, y_2|\boldsymbol{x}) = \phi_{11} S(y_1, y_2|\boldsymbol{x}) + \phi_{10} S(y_1|\boldsymbol{x}) + \phi_{01} S(y_2|\boldsymbol{x}) + \phi_{00}, \tag{33}$$

em que $S(y_1, y_2|\boldsymbol{x})$ é a função de sobrevivência bivariada representada por meio de cópula, $S(y_k)$, $k = 1, 2$, é uma função de sobrevivência marginal e são consideradas as seguintes restrições $0 < \phi_{11} < 1$, $0 < \phi_{10} < 1$, $0 < \phi_{01} < 1$, $0 < \phi_{00} < 1$ e $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$. É importante notar que se pode considerar $\phi_{10} = \phi_{01}$, ou ainda, $\phi_{10} \neq \phi_{01}$, dependendo do comportamento dos dados em estudo. A identificação desse comportamento pode ser observada por meio de um gráfico da função de sobrevivência empírica estimada pelo método de Kaplan-Meier para o tempo de cada evento de interesse. Se as funções de sobrevivência empírica apresentarem a cauda à direita em um mesmo nível constante, tem-se $\phi_{10} = \phi_{01}$; caso contrário, $\phi_{10} \neq \phi_{01}$.

Da equação (33), a fração de cura conjunta é determinada por:

$$\lim_{y_1 \to \infty, y_2 \to \infty} S(y_1, y_2|\boldsymbol{x}) = \phi_{00} \tag{34}$$

e as funções de sobrevivência marginais são

$$S_1(y_1|\boldsymbol{x}) = \lim_{y_2 \to -\infty} S_{pop}(y_1, y_2|\boldsymbol{x}) = \phi_{11} S(y_1|\boldsymbol{x}) + \phi_{10} S(y_1|\boldsymbol{x}) + \phi_{01} + \phi_{00}, \tag{35}$$

com probabilidade de cura $\phi_{01} + \phi_{00}$ e

$$S_2(y_2|\boldsymbol{x}) = \lim_{y_1 \to -\infty} S_{pop}(y_1, y_2|\boldsymbol{x}) = \phi_{11} S(y_2|\boldsymbol{x}) + \phi_{10} + \phi_{01} S(y_2|\boldsymbol{x}) + \phi_{00}, \tag{36}$$

com probabilidade de cura $\phi_{10} + \phi_{00}$.

O parâmetro $\mu_i = \boldsymbol{x}_i^T \beta$ é a locação de y_i. O vetor de parâmetros de locação $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ é representado por um modelo linear $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, em que $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ é uma matriz conhecida. O modelo de regressão bivariado com fração de cura usando as equações (30), (31) e (33) gera novas possibilidades para ajustar diferentes tipos de dados. Para ilustrar a proporção de indivíduos curados em um modelo bivariado, as Figuras 11a, 11b e 11c mostram algumas possibilidades da forma da função de sobrevivência bivariada populacional (33) para valores selecionados dos parâmetros, considerando-se a distribuição do valor extremo com $\mu_1 = 3.4$, $\sigma_1 = 0.94$, $\mu_2 = 3.05$, $\sigma_2 = 1.15$ e $\alpha = 0.5$ para 0% de cura, 30% de cura e 50% de cura, respectivamente. Definido o modelo de regressão bivariado com fração de cura, a próxima etapa é baseada nos procedimentos de estimação. Neste caso específico, é preciso pensar em um procedimento de estimação que considere as restrições dos parâmetros do modelo definido em (33). Na próxima seção será discutido o método de máxima verossimilhança sujeito às restrições lineares nos parâmetros.

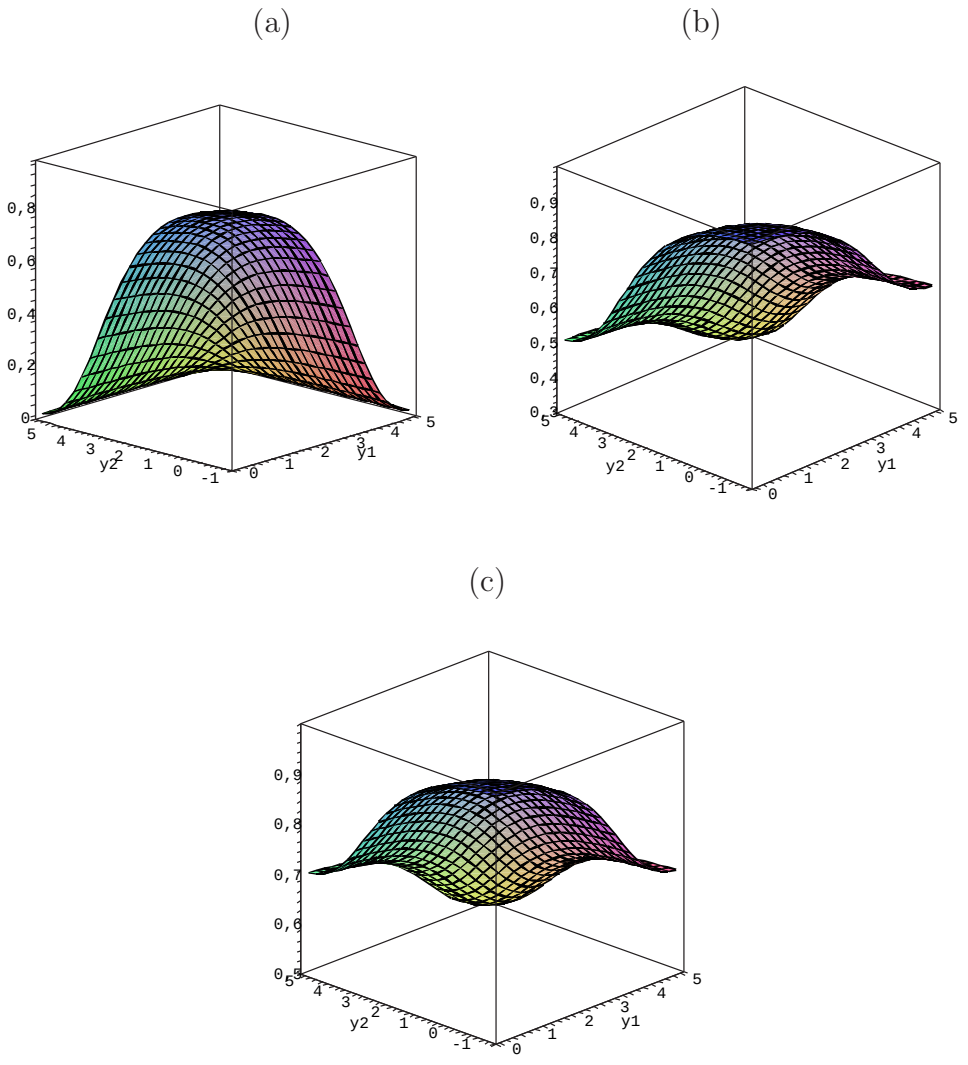

Figura 11 - Função de sobrevivência bivariada populacional para alguns valores dos parâmetros. (a) 0% de cura. (b) 30% de cura. (a) 50% de cura

3.4.1 Inferência para o modelo de regressão com fração de cura para dados bivariados por meio de cópulas

Considera-se uma amostra observada com variáveis $(y_{1k}, \delta_{1k}, \boldsymbol{x}_1), \ldots, (y_{nk}, \delta_{nk}, \boldsymbol{x}_n)$, sendo y_{ik} o logaritmo do tempo do k-ésimo evento de interesse do i-ésimo indivíduo, δ_{ik} a respectiva variável indicadora de censura e $\boldsymbol{x}_i$ o vetor de covariáveis associado ao i-ésimo indivíduo. São consideradas as mesmas covariáveis associadas

a ambas variáveis resposta, em que $k = 1, 2$ e $i = 1, 2, \ldots, n$. Dado que, y_{i1} e y_{i2} podem ser cada um deles censurado ou não, uma observação pode assumir quatro possibilidades. Sendo assim, a função de verossimilhança para o modelo de regressão bivariado com fração de cura é obtida ao utilizar a expressão definida em (32), em que a função de sobrevivência bivariada assume a forma dada em (33). Dessa forma, a contribuição do i-ésimo indivíduo para a função de verossimilhança é dada por:

(i) se $\delta_{i1} = 1$ e $\delta_{i2} = 1$, então a contribuição para a verossimilhança é dada por

$$(-1)^2 \partial^2 S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)/\partial y_{i1}\partial y_{i2} = \phi_{11} f(y_{i1}, y_{i2}|\boldsymbol{x}_i);$$

(ii) se $\delta_{i1} = 1$ e $\delta_{i2} = 0$, então a contribuição para a verossimilhança é dada por

$$\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)/\partial y_{i1} = \phi_{11}[\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)/\partial y_{i1}] + \phi_{10}[\partial S(y_{i1}|\boldsymbol{x}_i)/\partial y_{i1}];$$

(iii) se $\delta_{i1} = 0$ e $\delta_{i2} = 1$, então a contribuição para a verossimilhança é dada por

$$\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)/\partial y_{i2} = \phi_{11}\Big[\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)/\partial y_{i2}\Big] + \phi_{01}\Big[\partial S(y_{i2}|\boldsymbol{x}_i)/\partial y_{i2}\Big];$$

(iv) se $\delta_{i1} = 0$ e $\delta_{i2} = 0$, então a contribuição para a verossimilhança é dada por

$$S_{pop}(y_1, y_2|\boldsymbol{x}_i) = \phi_{11} S(y_{i1}, y_{i2}|\boldsymbol{x}_i) + \phi_{10} S(y_{i1}|\boldsymbol{x}_i) + \phi_{01} S(y_{i2}|\boldsymbol{x}_i) + \phi_{00}.$$

O logaritmo da função de verossimilhança para o modelo de regressão bivariado com fração de cura é representado pela seguinte equação:

$$\begin{aligned} l(\boldsymbol{\theta}) \;=\; & \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \log[\phi_{11} f(y_{i1}, y_{i2}|\boldsymbol{x}_i)] + \\ & \sum_{i=1}^{n} \delta_{i1}(1-\delta_{i2}) \log\left\{\phi_{11}\left[-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}}\right] + \phi_{10}\left[-\frac{\partial S(y_{i1}|\boldsymbol{x}_i)}{\partial y_{i1}}\right]\right\} + \\ & \sum_{i=1}^{n} (1-\delta_{i1})\delta_{i2} \log\left\{\phi_{11}\left[-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}}\right] + \phi_{01}\left[-\frac{\partial S(y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}}\right]\right\} + \\ & \sum_{i=1}^{n} (1-\delta_{i1})(1-\delta_{i2}) \log[\phi_{11} S(y_{i1}, y_{i2}|\boldsymbol{x}_i) + \phi_{10} S(y_{i1}|\boldsymbol{x}_i) + \phi_{01} S(y_{i2}|\boldsymbol{x}_i) + \phi_{00}], \end{aligned} \tag{37}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\phi}, \alpha, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, $\boldsymbol{\phi} = (\phi_{11}, \phi_{10}, \phi_{01}, \phi_{00})^T$, $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k)^T$, $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ e $\phi_{11} f(y_{i1}, y_{i2}|\boldsymbol{x}_i) = (-1)^2 \partial^2 S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)/\partial y_{i1}\partial y_{i2}$. O logaritmo da função de verossimilhança

definido em (37) possui as seguintes restrições nos parâmetros $\sigma_k > 0$, $\alpha > 0$, $0 < \phi_{11} < 1$, $0 < \phi_{10} < 1$, $0 < \phi_{01} < 1$, $0 < \phi_{00} < 1$ e $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$, para $k = 1, 2$.

O interesse é estimar o vetor de parâmetros $\boldsymbol{\theta}$ sob 11 restrições de inequações lineares $\boldsymbol{u}_j^T\boldsymbol{\theta} - c_j \geq 0$, em que $\boldsymbol{u}_j, j = 1, 2, \ldots, 11$ são 11×1 vetores e c_j são escalares assumindo valores 0 ou 1 dependendo da restrição de interesse e equações lineares $\boldsymbol{v}_j^T\boldsymbol{\theta} = d_j$, em que $\boldsymbol{v}_j, j = 1$ é o vetor de dimensão 11×1 e d_j é um escalar que assume valor 1. Os vetores $\boldsymbol{u}_j$ e $\boldsymbol{v}_j$ são do tipo $(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, assumindo valor 1 na posição que o parâmetro de interesse encontra-se.

Para utilizar um método de estimação em um passo que maximize o logaritmo da função de verossimilhança $l(\boldsymbol{\theta})$ definido em (37) e sujeito às restrições nos parâmetros, é necessário considerar o método da função barreira adaptada (LANGE, 1999), que é uma combinação do método barreira logaritmo com o algoritmo EM. Sendo assim, o logaritmo da função de verossimilhança sujeito às restrições lineares é representado por

$$l_R(\boldsymbol{\theta}, \vartheta) = l(\boldsymbol{\theta}) + \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T\boldsymbol{\theta} - c_j), \tag{38}$$

em que o parâmetro de ajuste é uma constante positiva, $\vartheta > 0$, $\boldsymbol{u}_j^T\boldsymbol{\theta} - c_j$ é o conjunto de restrições de inequações lineares, que pode reduzir-se em, $\boldsymbol{v}_j^T\boldsymbol{\theta} = d_j$ que é o conjunto de restrições de equações lineares para $j = 1, 2, \ldots, q$, $\boldsymbol{\theta} = (\boldsymbol{\phi}, \alpha, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, $\boldsymbol{\phi} = (\phi_{11}, \phi_{10}, \phi_{01}, \phi_{00})^T$, $\boldsymbol{\theta}_k^T = (\boldsymbol{\beta}_k^T, \sigma_k)^T$ e $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})^T$.

Neste trabalho, o *software* R (R DEVELOPMENT CORE TEAM, 2009) foi utilizado para obter as estimativas de máxima verossimilhança com restrição nos parâmetros por meio da função *constrOptim*. Para maiores detalhes sobre o método da função Barreira adaptada ver, por exemplo, Lange (1999).

Para realizar as inferências dos parâmetros de locação são consideradas propriedades assintóticas. Estimativas da matriz de covariância para os estimadores de máxima verossimilhança $\hat{\boldsymbol{\beta}}_k$ podem ser obtidas utilizando a matriz Hessiana. Nesse caso, os estimadores de máxima verossimilhança têm distribuição assintótica normal com média $\boldsymbol{\beta}_k$ e matriz de covariância dada por $\boldsymbol{I}^{-1}(\boldsymbol{\beta}_k)$ com $\boldsymbol{I}(\boldsymbol{\beta}_k) = E[\ddot{\boldsymbol{L}}_R(\boldsymbol{\beta}_k)]$, em que $\ddot{\boldsymbol{L}}_R(\boldsymbol{\beta}_k) = -\left\{\frac{\partial^2 l_R(\boldsymbol{\theta}, \vartheta)}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_k^T}\right\}$.

O cálculo de $\boldsymbol{I}(\boldsymbol{\beta}_k)$ fica comprometido devido à presença de observações censuradas, então pode-se utilizar, alternativamente, a matriz $[\ddot{\boldsymbol{L}}(\boldsymbol{\beta}_k)]$ avaliada em $\boldsymbol{\beta}_k = \hat{\boldsymbol{\beta}}_k$,

denominada matriz de informação observada, que é um estimador consistente de $\boldsymbol{I}(\boldsymbol{\beta}_k)$. A diagonal principal da matriz $\ddot{\boldsymbol{L}}^{-1}(\boldsymbol{\beta}_k)$ é utilizada como uma estimativa para a variância dos estimadores.

Assim, um intervalo de confiança de $(1-\alpha)100\%$ para um parâmetro β_{lk}, em que $l = 0, 1, 2, \ldots, p$ e $k = 1, 2$, é expresso por:

$$\hat{\beta_{lk}} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\beta_{lk}})}.$$

Para realizar testes de hipóteses sobre os parâmetros utiliza-se, por exemplo, a estatística definida por:

$$z = \frac{\hat{\beta_{lk}} - \beta_{l0}}{\sqrt{\widehat{Var}(\hat{\beta_{lk}})}},$$

em que $Z \sim N(0,1)$ e β_{l0} é o verdadeiro valor do parâmetro.

Neste trabalho não será descrito o procedimento inferêncial para os parâmetros com restrição, devido à sua complexidade. O tema poderá ser abordado em pesquisas futuras.

3.5 Análise de sensibilidade

Quando se ajusta um modelo a um conjunto de dados, é imprescindível que as estimativas obtidas a partir do modelo proposto sejam resistentes a pequenas perturbações nas observações. Se o modelo ajustado não apresentar uma boa descrição dos dados observados, poderá conduzir a conclusões errôneas.

Por isso, é importante que se faça um estudo sobre a robustez dos resultados obtidos, considerando-se vários aspectos que envolvem a formulação do modelo e as estimativas dos seus parâmetros. A análise de diagnóstico deve consistir de métodos que avaliem o grau de sensibilidade das inferências a pequenas perturbações nos dados ou mesmo no modelo proposto.

Uma primeira metodologia proposta por Cook (1977) considera a deleção do i-ésimo indivíduo e verifica o quanto sua ausência influencia nos resultados da estimação dos parâmetros do modelo. Essa metodologia é conhecida como Influência Global, ou deleção de casos. Para maiores detalhes, ver Cook et al. (1988) e Xie e Wei (2007).

Contudo, tal análise considera a deleção de todas as informações referentes a um indivíduo, o que dificulta a verificação da influencia do indivíduo em algum aspecto específico

do modelo. Para corrigir esse déficit, Cook (1986) propõe a medida de Influência Local com o objetivo de verificar se o modelo é relativamente estável sob pequenas perturbações. Os esquemas de perturbação devem levar em consideração os aspectos da análise que se deseja monitorar e devem ser interpretáveis. Uma extensão da metodologia proposta por Cook (1986) é a metodologia de Influência Local Total proposta por Lesaffre e Verbeke (1998).

Devido à importância dessa metodologia, verifica-se, quando se considera a linha de pesquisa de modelos com fração de cura segundo a abordagem de Berkson e Gage (1952) e modelos bivariados construídos por meio de cópulas, aumento considerável de trabalhos que propõem uma análise de sensibilidade. Por exemplo, Gomes (2007) considera a metodologia de influência em modelos de regressão bivariados por meio de cópulas; Ortega et al. (2009b) propõem uma análise de Influência Local e resíduo para o modelo de mistura log-gama generalizado com covariáveis; Barriga et al. (2010) consideram um modelo de regressão bivariado por meio de cópula para dados de sobrevivência pareados e aplicam uma análise de Influência Local e resíduo. No entanto, tais pesquisas consideram o método de máxima verossimilhança como o método de estimação. Como detalhado anteriormente, este trabalho utiliza o método de estimação baseado em uma função de verossimilhança sujeita a restrições nos parâmetros. Por isso, os trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009) serão utilizados como referências para desenvolver uma análise de sensibilidade para o modelo de regressão bivariado com fração de cura.

3.5.1 Influência Global sob verossimilhança restrita

Como enunciado anteriormente, uma importante técnica em análise de sensibilidade é a metodologia de deleção de casos. Essa metodologia é considerada para avaliar o efeito da i-ésima observação nas estimativas e o quanto a deleção de um caso pode alterar os resultados do modelo proposto.

Para a variável aleatória contínua $Y_k = \log(T_k)$ com $k = 1, 2$, o modelo caso deleção para o modelo (30) é representado por:

$$\boldsymbol{y}_{(i)k} = \boldsymbol{x}_{(i)k}^T \boldsymbol{\beta}_k - \sigma_k \boldsymbol{z}_{(i)k}, \qquad i = 1, \ldots, n, \tag{39}$$

em que o subscrito $(i)k$ indica que a i-ésima observação do k-ésimo evento foi retirada da amostra.

O estimador de máxima verossimilhança sujeita a restrições $\hat{\boldsymbol{\theta}}_{(i)}$ é obtido a partir de $l_{R(i)}(\boldsymbol{\theta}, \vartheta)$. A influência da i-ésima observação nos estimadores é avaliada por meio da diferença $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$. Essa diferença $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ é relativamente grande para observações que podem ser consideradas influentes. Nessa situação, a observação considerada deve ser analisada com prudência.

As medidas de Influência Global definidas na literatura são conhecidas como Distância de Cook Generalizada e Afastamento da Verossimilhança. A primeira é definida como a norma padronizada de $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ e é dada por:

$$GD_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \boldsymbol{M} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}),$$

em que podem ser consideradas várias escolhas de $\boldsymbol{M}$, segundo Cook e Weisberg (1982). As escolhas mais utilizadas entre os pesquisadores consideram $\boldsymbol{M} = -\ddot{L}(\hat{\boldsymbol{\theta}})$ ou $\boldsymbol{M} = [-\ddot{L}(\hat{\boldsymbol{\theta}})]^{-1}$.

A segunda medida para avaliar a sensibilidade causada pela i-ésima observação é chamada Afastamento da Verossimilhança e é dada por:

$$LD_i(\boldsymbol{\theta}) = 2[l_R(\hat{\boldsymbol{\theta}}, \vartheta) - l_{R(i)}(\hat{\boldsymbol{\theta}}, \vartheta)].$$

3.5.2 Influência Local sob verossimilhança restrita

Para desenvolver uma metodologia de Influência Local para o modelo de regressão bivariado com fração de cura, serão utilizadas a teoria desenvolvida por Cook (1986) e a metodologia de influêncial local sob verossimilhança restrita descrita detalhadamente na seção 2.4.3 para o modelo de regressão Kumaraswamy Weibull bivariado com base nos trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009). De acordo com a metodologia exposta na seção 2.4.3, para o modelo de regressão bivariado com fração de cura o logaritmo da função de verossimilhança perturbada sujeita as restrições lineares dado um vetor de perturbações $\boldsymbol{w}$, $b \times 1$, é definido por

$$\mathrm{l_R}(\boldsymbol{\theta}, \vartheta|\boldsymbol{w}) = \mathrm{l}(\boldsymbol{\theta}|\boldsymbol{w}) + \vartheta \sum_{\mathrm{j}=1}^{\mathrm{q}} (\boldsymbol{u}_{\mathrm{j}}^{\mathrm{t}} \boldsymbol{\theta} - \mathrm{c_j}) \tag{40}$$

em que a constante $\vartheta > 0$ é o multiplicador do termo barreira. É importante observar na equação (40) que as restrições não são alteradas pelo esquema de perturbação, o que garante solução no subespaço paramétrico.

A seguir, são apresentados três esquemas de perturbações mais comumente utilizados: ponderação de casos, perturbação na variável resposta e perturbação na variável explicativa para o modelo proposto neste capítulo. Para cada um dos esquemas de perturbação é necessário obter a matriz $\boldsymbol{\Delta}$, com componentes $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_\alpha, \boldsymbol{\Delta}_{\boldsymbol{\phi}}, \boldsymbol{\Delta}_{\boldsymbol{\beta}}, \boldsymbol{\Delta}_{\boldsymbol{\sigma}})^T$, definida por:

$$\Delta = \Big(\Delta_{vi}\Big)_{\big[(2p+7)\times n\big]} = \left(\frac{\partial^2 l_R(\boldsymbol{\theta}, \vartheta|\boldsymbol{w})}{\partial\theta_v \partial w_i}\right)_{\big[(2p+7)\times n\big]},$$

em que $v = 1, 2, \ldots, 2p+7$ e $i = 1, 2, \ldots, n$, utilizando-se o modelo definido em (33) e o logaritmo da função de verossimilhança com restrição nos parâmetros (40). Os elementos da matriz são obtidos numericamente.

A idéia de perturbar o modelo seguindo o esquema de perturbação de casos pode ser interpretada como uma flexibilização da deleção de casos, de forma que a influencia conjunta das observações pode ser investigada e os casos de perturbação das respostas ou covariáveis podem ser interpretados como uma forma de detecção de pontos atípicos, com erro ou simplesmente mal modelados.

3.5.2.1 Perturbação de casos

Sob o esquema de perturbação de casos, o logaritmo da função de verossimilhança perturbada sujeita as restrições lineares é representado por:

$$\begin{aligned}
l_R(\boldsymbol{\theta}, \vartheta|\boldsymbol{w}) \;=\; & \sum_{i=1}^{n} w_i \delta_{i1}\delta_{i2} \log[\phi_{11} f(y_{i1}, y_{i2}|\boldsymbol{x}_i)] + \\
& \sum_{i=1}^{n} w_i \delta_{i1}(1-\delta_{i2}) \log\left\{\phi_{11}\left[-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}}\right] + \phi_{10}\left[-\frac{\partial S(y_{i1}|\boldsymbol{x}_i)}{\partial y_{i1}}\right]\right\} + \\
& \sum_{i=1}^{n} w_i (1-\delta_{i1})\delta_{i2} \log\left\{\phi_{11}\left[-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}}\right] + \phi_{01}\left[-\frac{\partial S(y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}}\right]\right\} + \\
& \sum_{i=1}^{n} w_i (1-\delta_{i1})(1-\delta_{i2}) \log[\phi_{11} S(y_{i1}, y_{i2}|\boldsymbol{x}_i) + \phi_{10} S(y_{i1}|\boldsymbol{x}_i) + \phi_{01} S(y_{i2}|\boldsymbol{x}_i) + \phi_{00}] + \\
& \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^t \boldsymbol{\theta} - c_j),
\end{aligned}$$

em que o vetor correspondente à não perturbação é o vetor $\boldsymbol{w}_0 = (1, \ldots, 1)^T$, n-dimensional.

3.5.2.2 Perturbação da resposta

Nesse esquema de perturbação considera-se que cada variável resposta y_{i1} e y_{i2} é perturbada como $y^*_{i1} = y_{i1} + S_{y_1} w_i$ e $y^*_{i2} = y_{i2} + S_{y_2} w_i$, em que S_{y_k} são fatores de escala que podem ser a estimativa do desvio padrão da variável Y_k, $k = 1, 2$ e $w_i \in \mathbf{R}$. O logaritmo da verossimilhança perturbada sujeita as restrições lineares é dado por:

$$
\begin{aligned}
l_R(\boldsymbol{\theta}, \vartheta|\boldsymbol{w}) \;=\; & \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \log[\phi_{11} f(y^*_{i1}, y^*_{i2}|\boldsymbol{x}_i)] + \\
& \sum_{i=1}^{n} \delta_{i1}(1-\delta_{i2}) \log\left\{ \phi_{11}\left[-\frac{\partial S(y^*_{i1}, y^*_{i2}|\boldsymbol{x}_i)}{\partial y^*_{i1}} \right] + \phi_{10}\left[-\frac{\partial S(y^*_{i1}|\boldsymbol{x}_i)}{\partial y^*_{i1}} \right] \right\} + \\
& \sum_{i=1}^{n} (1-\delta_{i1})\delta_{i2} \log\left\{ \phi_{11}\left[-\frac{\partial S(y^*_{i1}, y^*_{i2}|\boldsymbol{x}_i)}{\partial y^*_{i2}} \right] + \phi_{01}\left[-\frac{\partial S(y^*_{i2}|\boldsymbol{x}_i)}{\partial y^*_{i2}} \right] \right\} + \\
& \sum_{i=1}^{n} (1-\delta_{i1})(1-\delta_{i2}) \log[\phi_{11} S(y^*_{i1}, y^*_{i2}|\boldsymbol{x}_i) + \phi_{10} S(y^*_{i1}|\boldsymbol{x}_i) + \phi_{01} S(y^*_{i2}|\boldsymbol{x}_i) + \phi_{00}] + \\
& \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^t \boldsymbol{\theta} - c_j),
\end{aligned}
$$

em que $y^*_{ik} = y_{ik} + S_{y_k} w_i$ e o vetor de não perturbação $\boldsymbol{w}_0 = (0, \ldots, 0)^T$. Nesse caso, podem ser consideradas três possibilidades de perturbação, ou seja, perturbar apenas a variável resposta y_{i1}, ou apenas y_{i2}, ou y_{i1} e y_{i2} conjuntamente.

3.5.2.3 Perturbação da covariável

Considera-se agora uma perturbação aditiva em uma determinada variável explicativa contínua, X_t, que pode ser representada por $x_{itw} = x_{it} + S_{xt} w_i$, em que S_{xt} é um fator de escala que pode ser a estimativa do desvio padrão de X_t e $w_i \in \mathbf{R}$. O logaritmo da

verossimilhança perturbada sujeita as restrições lineares é dado por:

$$
\begin{aligned}
l_R(\boldsymbol{\theta}, \vartheta|\boldsymbol{w}) \;=\; & \sum_{i=1}^{n} \delta_{i1}\delta_{i2} \log[\phi_{11} f(y_{i1}, y_{i2}|\boldsymbol{x}_i^*)] + \\
& \sum_{i=1}^{n} \delta_{i1}(1-\delta_{i2}) \log\left\{ \phi_{11}\left[-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i^*)}{\partial y_{i1}} \right] + \phi_{10}\left[-\frac{\partial S(y_{i1}|\boldsymbol{x}_i^*)}{\partial y_{i1}} \right] \right\} + \\
& \sum_{i=1}^{n} (1-\delta_{i1})\delta_{i2} \log\left\{ \phi_{11}\left[-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i^*)}{\partial y_{i2}} \right] + \phi_{01}\left[-\frac{\partial S(y_{i2}|\boldsymbol{x}_i^*)}{\partial y_{i2}} \right] \right\} + \\
& \sum_{i=1}^{n} (1-\delta_{i1})(1-\delta_{i2}) \log[\phi_{11} S(y_{i1}, y_{i2}|\boldsymbol{x}_i^*) + \phi_{10} S(y_{i1}|\boldsymbol{x}_i^*) + \phi_{01} S(y_{i2}|\boldsymbol{x}_i^*) + \phi_{00}] + \\
& \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^t \boldsymbol{\theta} - c_j),
\end{aligned}
$$

em que $\boldsymbol{x}_i^{*T}\boldsymbol{\beta}_j = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \ldots + \beta_{tk}(x_{it} + S_{xt}w_i) + \ldots + \beta_{pk}x_{ip}$ e o vetor de não perturbação $\boldsymbol{w}_0 = (0, \ldots, 0)^T$.

3.6 Aplicação

A aplicação descrita a seguir tem como objetivo mostrar a utilidade do modelo de regressão bivariado com fração de cura para dados bivariados por meio de cópulas. O conjunto de dados bivariados considerado na aplicação foi analisado por Huster et al. (1989), Liang et al. (1993), Wada e Hotta (2000) e Tarumoto (2001).

Os dados são referentes a um subconjunto de dados de estudo de retinopatia diabética iniciado em 1971. A retinopatia diabética é uma complicação associada com diabetes mellitus e consiste em anormalidades na microcirculação verificadas na retina do olho. É a principal causa de cegueira em pacientes com menos de 60 anos de idade nos Estados Unidos e a principal causa de perda de acuidade visual no Brasil e no mundo.

O objetivo do estudo de Huster et al. (1989) foi testar a efetividade do tratamento de pacientes com cegueira provocada por retinopatia diabética por meio da utilização de fotocoagulação a laser. O estudo propôs-se a verificar se este tipo de tratamento pode retardar o aparecimento da cegueira em pacientes portadores da doença. O objetivo secundários foi determinar se o tempo de sobrevivência para os olhos está relacionado ao tratamento e tipo de diabetes, classificado em dois grupos gerais de acordo com a idade apresentada no início do tratamento: diabetes juvenil e adulto.

Pacientes com retinopatia diabética em ambos os olhos e com acuidade visual menor ou igual a 20/100 para ambos os olhos fizeram parte do estudo. Um olho foi selecionado aleatoriamente para receber o tratamento e o outro foi observado sem tratamento. Os pacientes foram observados por dois períodos de 4 meses completos e foi considerado como falha a ocorrência da acuidade visual menor que 5/200. No total, 1742 pacientes foram acompanhados durante 7 anos. Ao final, 197 pacientes fizeram parte do subconjunto em estudo definido por critério de estudo de retinopatia diabética. Para cada paciente i, $i = 1, 2, \ldots, 197$ as variáveis associadas são:

- t_{i1}: tempo de queda da acuidade visual até o nível definido para o olho com tratamento (evento 1);
- t_{i2}: tempo de queda da acuidade visual até o nível definido para o olho sem tratamento (evento 2);
- δ_{i1}: indicador de censura do evento 1;
- δ_{i2}: indicador de censura do evento 2;
- x_{i1}: tipo de diabetes (0: diabetes juvenil, 1: diabetes adulto).

3.6.1 Análise Descritiva

Realizou-se uma análise exploratória dos dados que considerou como evento de interesse o tempo até a perda da acuidade visual para o olho com tratamento e para o olho sem tratamento. O coeficiente de correlação de τ de Kendall foi calculado para os tempos bivariados, obtendo-se $\tau = 0,39$, o que indicou uma pequena associação positiva entre os eventos 1 e 2 em estudo, fato que justifica o uso do modelo bivariado para captar a associção existente entre as variáveis respostas.

Na Figura 12, são apresentadas as estimativas de sobrevivência de Kaplan-Meier e o ajuste marginal do modelo com distribuição Weibull para ambos os eventos 1 e 2. A análise dessa figura, permite concluir que a distribuição Weibull é adequada para os tempos de sobrevivência, e, consequentemente, usar a distribuição do valor extremo para o logaritmo dos tempos. Observa-se, ainda, a existência de uma significativa fração de indivíduos não

suscetíveis a ambos os eventos de interesse, pois o limite da estimativa de sobrevivência do evento 1 tende a $0,67$, $\lim_{t_1 \to \infty} \hat{S}(t_1) = 0,67$, e o limite da estimativa de sobrevivência do evento 2 tende a $0,39$, $\lim_{t_2 \to \infty} \hat{S}(t_2) = 0,39$, indicando que os dados de retinopatia diabética devem ser analisados por modelos que consigam captar uma fração de indivíduos curados. A seguir, abordada-se a modelagem desses dados.

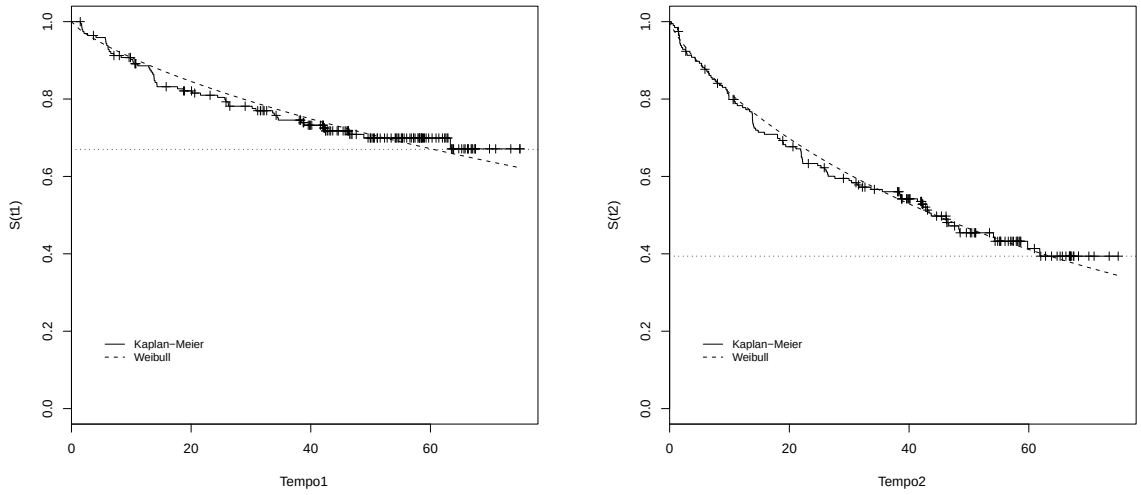

Figura 12 - Curvas de sobrevivência estimadas marginalmente por Kaplan-Meier para os dados de retinopatia

3.6.2 Ajuste do modelo de regressão com fração de cura para dados bivariados por meio de cópulas

Para analisar o conjunto de dados de pacientes portadores de retinopatia diabética é utilizado o modelo de regressão marginal de locação-escala, apresentado na seção 3.3.2, dado por:

$$y_{ik} = \beta_{0k} + \beta_{1k} x_{i1} + \sigma_k z_{ik}, \tag{41}$$

em que $i = 1, \ldots 197, k = 1, 2$, β_{pk} são os parâmetros de locação, σ_k são os parâmetros de escala, e z_{ik} são variáveis aleatórias independentes com função de sobrevivência do valor extremo

definida por:

$$S(z_{ik}) = \exp[-\exp(z_{ik})], \tag{42}$$

e $y_{ik} = \log(t_{ik})$ representa o logaritmo do tempo de falha.

Ao considerar o modelo definido na equação (41), a função de sobrevivência para o modelo de regressão log-linear bivariado por meio da cópula de Clayton, como definida na seção 3.3.2, é representada por:

$$S(y_{i1}, y_{i2}|\boldsymbol{x}_i) = \left\{ [\exp(-\exp(z_{i1}))]^{-\alpha} + [\exp(-\exp(z_{i2}))]^{-\alpha} - 1 \right\}^{-1/\alpha}, \tag{43}$$

em que $z_{ik} = [y_{ik} - (\beta_{0k} + \beta_{1k}x_{i1})]/\sigma_k$ e $k = 1, 2$.

Na seção 3.6.1, foi verificado por meio da Figura 12, um possível problema de fração de indivíduos não suscetíveis a ambos os eventos em estudo. Ao considerar essa informação, obtém-se uma função de sobrevivência bivariada populacional, como introduzida na seção 3.4, para os dados de pacientes com retinopatia diabética e, é dada por:

$$S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) = \phi_{11}S(y_{i1}, y_{i2}|\boldsymbol{x}_i) + \phi_{10}S(y_{i1}|\boldsymbol{x}_i) + \phi_{01}S(y_{i2}|\boldsymbol{x}_i) + \phi_{00}$$

em que $S(y_{i1}, y_{i2}|\boldsymbol{x}_i)$ é definida pela equação (43), $S(y_{ik}|\boldsymbol{x}_i)$ é uma função de sobrevivência do valor extremo dada em (42), e neste trabalho considera-se que $\phi_{10} \neq \phi_{01}$, pois $\lim_{t_1 \to \infty} \hat{S}(t_1) = 0,67$, e $\lim_{t_2 \to \infty} \hat{S}(t_2) = 0,39$, indicando a existência de uma fração de curados para o evento 1 diferente da proporção de curados no evento 2, como mostra a Figura 12.

Nesta etapa do trabalho, o interesse é estimar os parâmetros do modelo de regressão bivariado com fração de cura, sendo considerado o método de máxima verossimilhança sujeito às restrições nos parâmetros visto na seção 3.4.1.

Ao considerar os dados de pacientes portadores de retinopatia, a função de verossimilhança para o modelo de regressão log-linear bivariado com fração de cura é $l_R(\boldsymbol{\theta}, \vartheta)$ dada na expressão (38), em que as funções que compõem a função de verossimilhança ao considerar a cópula de Clayton com suposições marginais de distribuição valor extremo para os logaritmos dos tempos até a ocorrência dos eventos 1 e 2, são:

$$-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}} = \frac{S(y_{i1}, y_{i2}|\boldsymbol{x}_i)S(y_{i1}|\boldsymbol{x}_i)^{-\alpha}\exp(z_{i1})}{\sigma_1 S(y_{i1}, y_{i2}|\boldsymbol{x}_i)^{-\alpha}},$$

$$-\frac{\partial S(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}} = \frac{S(y_{i1}, y_{i2}|\boldsymbol{x}_i)S(y_{i2}|\boldsymbol{x}_i)^{-\alpha}\exp(z_{i2})}{\sigma_2 S(y_{i1}, y_{i2}|\boldsymbol{x}_i)^{-\alpha}},$$

$$-\frac{\partial S(y_{i1}|\boldsymbol{x}_i)}{\partial y_{i1}} = \frac{\exp(z_{i1})S(y_{i1}|\boldsymbol{x}_i)}{\sigma_1},$$

$$-\frac{\partial S(y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}} = \frac{\exp(z_{i2})S(y_{i2}|\boldsymbol{x}_i)}{\sigma_2},$$

$$f(y_{i1}, y_{i2}|\boldsymbol{x}_i) = \frac{S(y_{i1}, y_{i2}|\boldsymbol{x}_i)S(y_{i1}|\boldsymbol{x}_i)^{-\alpha}\exp(z_{i1})S(y_{i2}|\boldsymbol{x}_i)^{-\alpha}\exp(z_{i2})}{\sigma_1 S(y_{i1}, y_{i2}|\boldsymbol{x}_i)^{-2\alpha}\sigma_2}(1+\alpha),$$

sendo que $S(y_{i1}, y_{i2}|\boldsymbol{x}_i)$ é definida na equação (43), $S(y_{i1}|\boldsymbol{x}_i)$ e $S(y_{i2}|\boldsymbol{x}_i)$ são definidas na equação (42) e, o último termo da função de verossimilhança é composto pelas restrições nos parâmetros, $\sigma_k > 0$, $\alpha > 0$, $0 < \phi_{11} < 1$, $0 < \phi_{10} < 1$, $0 < \phi_{01} < 1$, $0 < \phi_{00} < 1$ e $\phi_{11} + \phi_{10} + \phi_{01} + \phi_{00} = 1$.

Na Tabela 3, encontram-se as estimativas para os parâmetros do modelo de regressão log-linear bivariado com fração de cura, bem como, os erros padrões e significâncias. Ao considerar um nível de significância de 5% a covariável x_1 é significativa para o evento 1, indicando que diabetes juvenil acelera a perda da acuidade visual no caso do olho tratado. Isso pode ser verificado por meio da Figura 13 que indica curvas de sobrevivência diferentes para cada tipo de diabetes.

A correlação entre as variáveis resposta, representada pelo parâmetro α, foi estimada conforme a equação (29) resultando em uma estimativa para o τ de Kendall $\tau_\alpha = 0,31$ próxima do τ de Kendall calculado na análise exploratória, indicando uma pequena associação entre os eventos 1 e 2 em estudo.

A estimativa da proporção de indivíduos não suscetíveis aos dois eventos de interesse é de $23,87\%$ representada pelo parâmetro ϕ_{00}, a proporção de indivíduos não suscetíveis ao evento 1 é de $59,41\%$, $\phi_{00} + \phi_{01}$, e a proporção de indivíduos não suscetíveis ao evento 2 é de $29,54\%$, $\phi_{00} + \phi_{10}$.

Tabela 3 - Estimativa de máxima verossimilhança para o modelo de regressão bivariado por meio da cópula de Clayton com fração de cura

Parâmetro	Estimativa	Erro-Padrão	valor-p
β_{01}	3,2937	0,3610	0,0000
β_{11}	0,7482	0,3782	0,0479
β_{02}	3,6970	0,3399	0,0000
β_{12}	-0,3613	0,2969	0,2236
σ_1	0,9371	0,1453	-
σ_2	1,0560	0,1116	-
α	0,8891	0,7069	-
ϕ_{00}	0,2387	0,0802	-
ϕ_{01}	0,3554	0,0801	-
ϕ_{10}	0,0567	0,0501	-
ϕ_{11}	0,3492	0,0754	-

Ao comparar as proporções de curados marginais obtidas por meio do modelo com a probabilidade de cura calculada empiricamente pelas curvas de Kaplan-Meier apresentadas na Figura 12 observa-se que os valores são relativamente próximos, concluindo que existe de fato a presença de uma proporção de indivíduos não suscetíveis a perda da acuidade visual tanto para o olho com tratamento como para o olho sem tratamento.

Através do modelo também é possível mensurar a probabilidade de indivíduos curados conjuntamente em ambos os eventos de interesse. Pode-se notar que o tratamento aplicado a um dos olhos mostrou-se eficiente devido a uma maior proporção de curados observada para os olhos tratados.

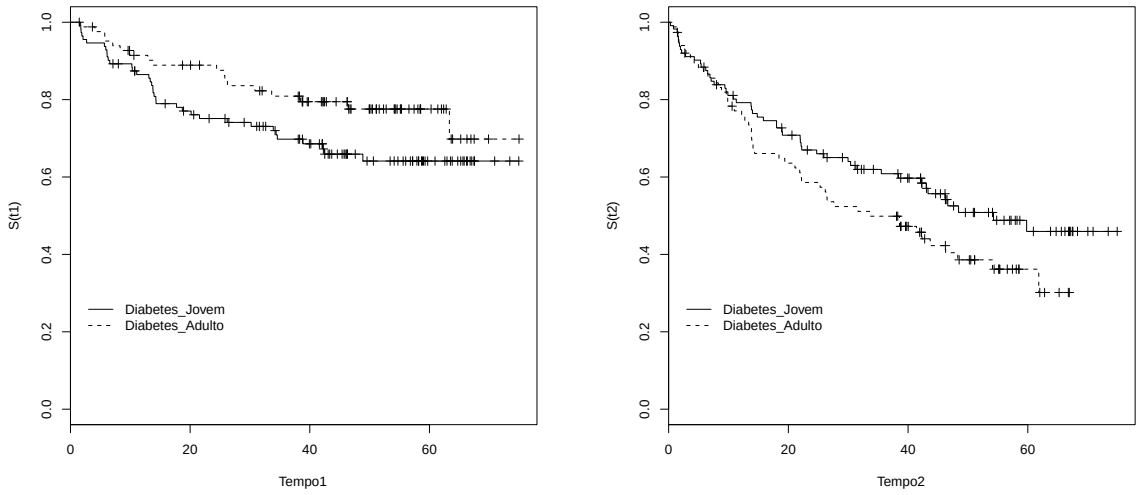

Figura 13 - Curvas de sobrevivência estimadas marginalmente por Kaplan-Meier para os dados de retinopatia diabética para cada evento por meio da covariável tipo de diabetes

3.6.3 Análise de Influência Global

Para detectar possíveis observações influentes no modelo de regressão bivariado com fração de cura para os dados de retinopatia, foram calculadas as medidas de Influência Global, Afastamento da Verossimilhança ($LD_i(\boldsymbol{\theta})$) e a Distância de Cook Generalizada ($GD_i(\boldsymbol{\theta})$).

A Figura 14 mostra que as observações #49 e #95 são as que mais se destacam, indicando que essas podem ser consideradas como possíveis pontos influentes.

Ao pesquisar sobre essas observações nos dados, verifica-se que: a observação #49 refere-se ao indivíduo com tempo de falha igual em ambos os eventos, e a observação #95 refere-se ao indivíduo com pequeno tempo de falha para o evento 1 e um grande tempo de censura para o evento 2. Devido as características discrepantes quanto ao tempo de falha e,ou censura de cada observação, elas foram detectadas por meio da análise de Influência Global como possíveis pontos influentes na análise estatística.

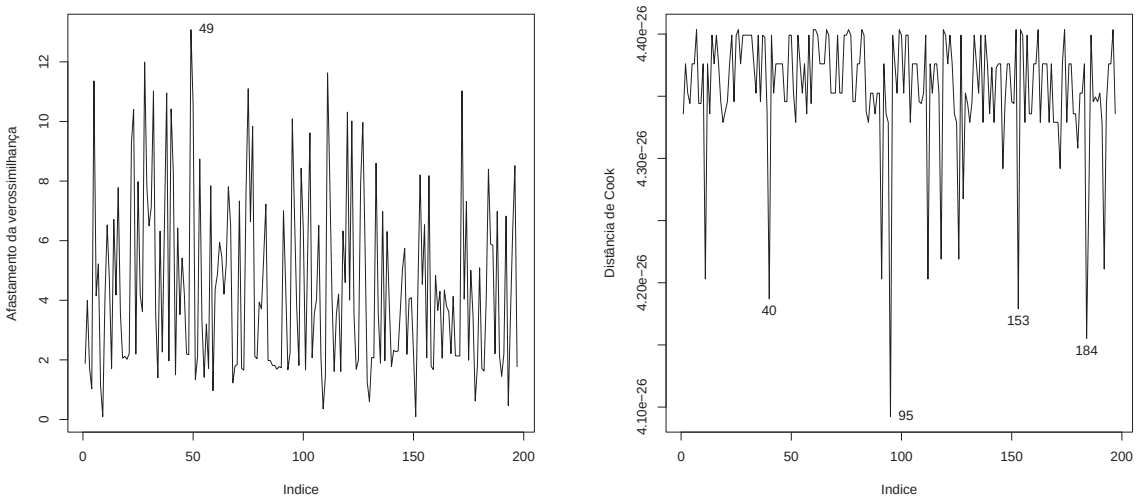

Figura 14 - Gráficos de medidas de Influência Global para o modelo de regressão bivariado com fração de cura para os dados de retinopatia diabética. (a) Afastamento da Verossimilhança (b) Distância de Cook

3.6.4 Análise de Influência Local

A teoria de Influência Local é aplicada aos dados de retinopatia considerando o modelo de regressão bivariado com fração de cura, em que é considerada distribuição Weibull para cada tempo de falha. Tal análise foi realizada considerando os esquemas de ponderação de casos, perturbação nos tempos até a perda da acuidade visual para o olho com tratamento (variável resposta 1), perturbação nos tempos até a perda da acuidade visual para o olho sem tratamento (variável resposta 2), e perturbação conjunta em ambas as variáveis resposta. O esquema de perturbação da covariável não foi considerado nesta análise, pois a covariável tipo de diabetes é categórica.

Para cada esquema considerado, foram calculados os vetores $\boldsymbol{d}_{max}$, $\boldsymbol{d}_{max}(y_1)$, $\boldsymbol{d}_{max}(y_2)$ e $\boldsymbol{d}_{max}(y_1y_2)$ correspondente à direção da maior curvatura, e os autovalores das curvaturas máximas que são dados por: $C_{\boldsymbol{d}_{max}}(\boldsymbol{\theta}) = 2,57$, $C_{\boldsymbol{d}_{max}}(y_1)$ =6,65, $C_{\boldsymbol{d}_{max}}(y_2) = 10,50$ e $C_{\boldsymbol{d}_{max}}(y_1y_2) = 8,88$, respectivamente.

As Figuras 15, 16, 17 e 18 apresentam os gráficos para a medida de Influência

Local $\boldsymbol{d}_{max}$ $\boldsymbol{d}_{max}(y_1)$, $\boldsymbol{d}_{max}(y_2)$ e $\boldsymbol{d}_{max}(y_1y_2)$, e para a medida de Influência Local Total, C_i, $C_{y_{i1}}$, $C_{y_{i2}}$ e $C_{y_{i1}y_{i2}}$ contra o índice das observações. Ao analisar essas figuras, observa-se que a observação #38 foi a que mais se destacou das demais e pode ser considerada como possível ponto influente por apresentar algumas característica interessantes, isto é, essa observação refere-se ao indivíduo com pequeno tempo de falha para o evento 1 e um grande tempo de falha para o evento 2.

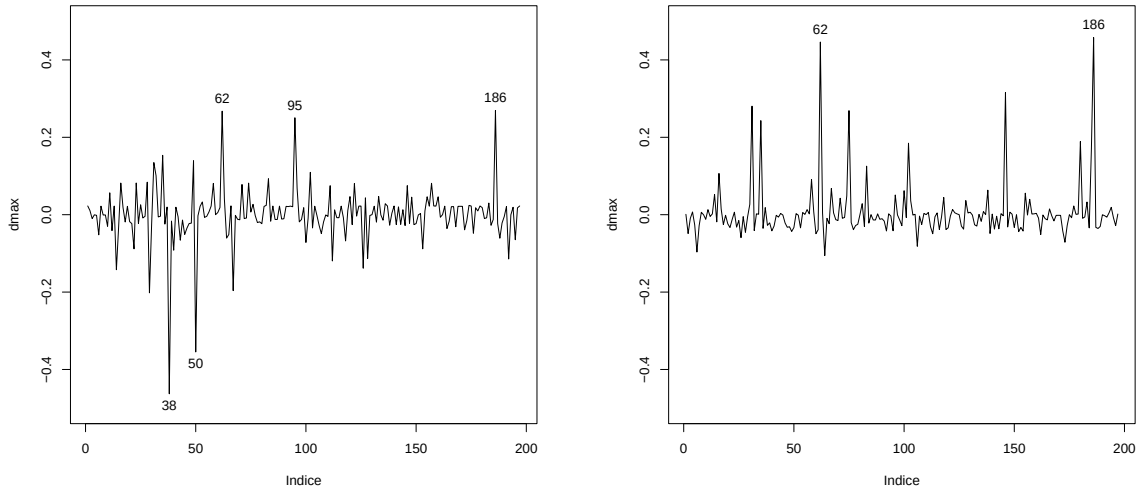

Figura 15 - Gráfico de medidas de influência do modelo de regressão bivariado com fração de cura considerando o esquema de perturbação de casos e da variável resposta 1 para os dados de retinopatia diabética. (a) Influência Local $\boldsymbol{d}_{max}$ casos (b) Influência Local $\boldsymbol{d}_{max}(y_1)$

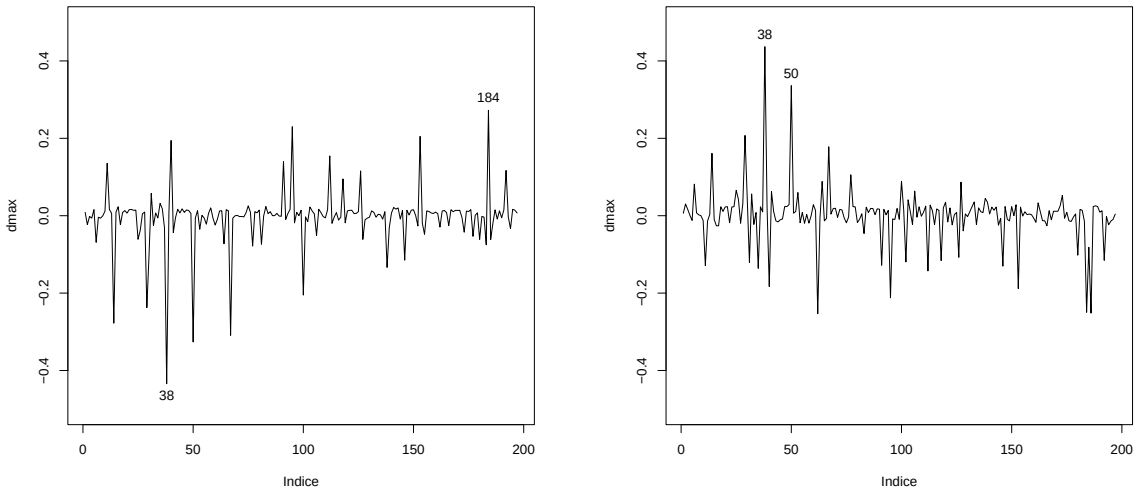

Figura 16 - Gráfico de medidas de influência do modelo de regressão bivariado com fração de cura considerando o esquema de perturbação da variável resposta 2 e a perturbação conjunta de ambas as variáveis resposta para os dados de retinopatia diabética. (a) Influência Local $\boldsymbol{d}_{max}(y_2)$ (b) Influência Local $\boldsymbol{d}_{max}(y_1 y_2)$

3.6.5 Impacto das observações influentes

Ao considerar a análise de sensibilidade para os dados de retinopatia diabética por meio do modelo de regressão bivariado com fração de cura, algumas observações são detectadas como possíveis pontos influentes na modelagem por se destacarem das demais. As observações #38 e #95 representam indivíduos com comportamentos peculiares, mas não apresentam sinais de erro na coleta ou na transcrição dos dados, e portanto devem ser mantidas no banco de dados.

Para avaliar a sensibilidade do modelo e de suas estimativas, será verificado o quanto essas observações podem influenciar no comportamento do modelo. Essa análise considera novas estimativas para os parâmetros do modelo a partir de subamostras referentes a retirada dessas observações individualmente e em grupo.

As estimativas para os parâmetros, os respectivos p-valores e as mudanças relati-

vas de cada parâmetro, considerando cada sub amostra, encontram-se na Tabela 4. A mudança relativa foi definida como $\boldsymbol{RC}_{\theta_j} = [(\hat{\theta}_j - \hat{\theta}_{j(I)})/\hat{\theta}_j]$, sendo (I) o índice referente as observações excluidas da amostra.

Ao analisar a Tabela 4 verifica-se que as conclusões baseadas no modelo de regressão bivariado com fração de cura não sofrem mudanças significativas com a presença ou ausência das observações identificadas como possíveis pontos influentes. Esse fato é muito importante, pois indica que o modelo proposto neste capítulo é robusto nesta aplicação.

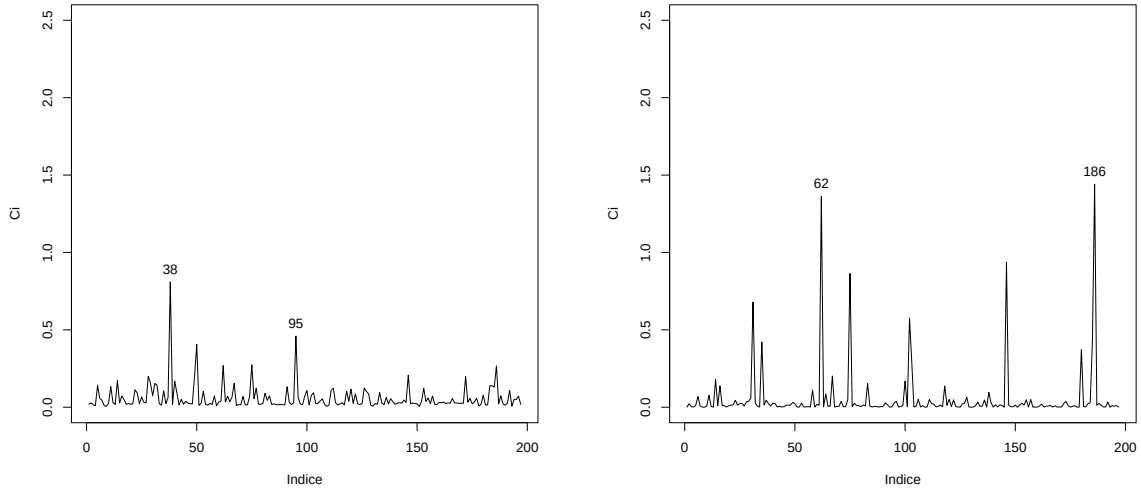

Figura 17 - Gráfico de medidas de influência do modelo de regressão bivariado com fração de cura considerando o esquema de perturbação de casos e da variável resposta 1 para os dados de retinopatia diabética. (a) Influência Local Total C_i casos (b) Influência Local Total $C_{y_{i1}}$

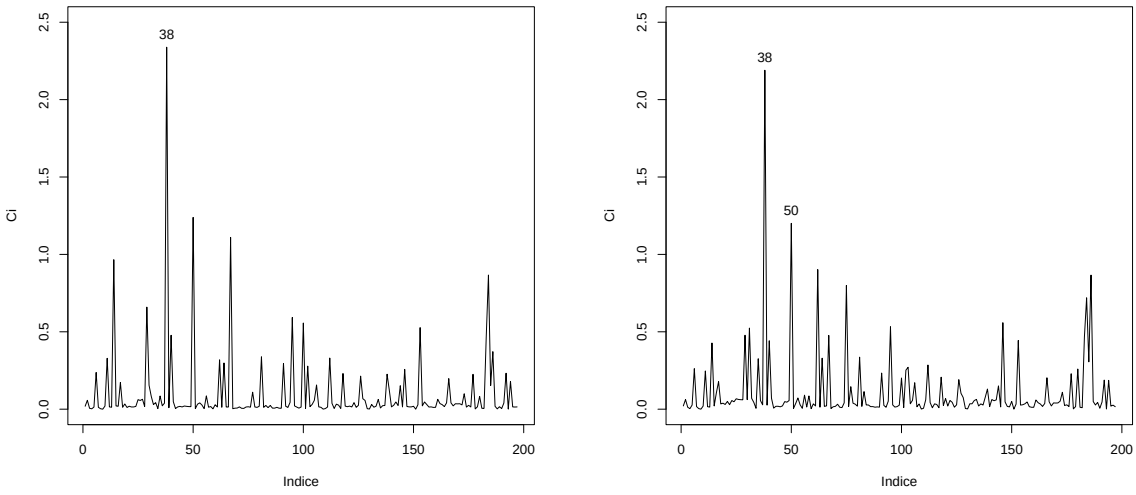

Figura 18 - Gráfico de medidas de influência do modelo de regressão bivariado com fração de cura considerando o esquema de perturbação da variável resposta 2 e a perturbação conjunta de ambas as variáveis respostas para os dados de retinopatia diabética. (a) Influência Local Total $C_{y_{i2}}$ (b) Influência Local Total $C_{y_{i1}y_{i2}}$

Tabela 4 - Mudança relativa [RC], estimativas dos parâmetros, e correspondentes (p-valor)

Sub-amostra	$I-\{completo\}$	$I-\{38\}$	$I-\{95\}$	$I-\{38,95\}$
	[-]	[-3,2499]	[3,0493]	[-1,3390]
β_{01}	3,2937	3,4007	3,1933	3,3378
	(0,0000)	(0,0000)	(0,0000)	(0,0000)
	[-]	[4,6383]	[-10,6162]	[-9,9135]
β_{11}	0,7482	0,7135	0,8277	0,8224
	(0,0479)	(0,0598)	(0,0291)	(0,0324)
	[-]	[2,3072]	[-6,4682]	[0,8178]
β_{02}	3,6970	3,6117	3,9362	3,6668
	(0,0000)	(0,0000)	(0,0000)	(0,0000)
	[-]	[0,3238]	[-36,8812]	[-7,6333]
β_{12}	-0,3613	-0,3601	-0,4946	-0,3889
	(0,2236)	(0,1952)	(0,1159)	(0,2000)
	[-]	[-0,8711]	[3,2380]	[1,5556]
σ_1	0,9371	0,9453	0,9067	0,9225
	(-)	(-)	(-)	(-)
	[-]	[1,8761]	[-3,8556]	[0,8461]
σ_2	1,0560	1,0362	1,0967	1,0471
	(-)	(-)	(-)	(-)
	[-]	[-58,1308]	[33,6971]	[-33,8986]
α	0,8891	1,4060	0,5895	1,1905
	(-)	(-)	(-)	(-)
	[-]	[-0,6156]	[6,6276]	[-2,0088]
ϕ_{00}	0,2387	0,2402	0,2229	0,2435
	(-)	(-)	(-)	(-)
	[-]	[4,4585]	[-9,8612]	[2,6845]
ϕ_{01}	0,3554	0,3395	0,3904	0,3458
	(-)	(-)	(-)	(-)
	[-]	[-46,6489]	[98,2263]	[-9,4287]
ϕ_{10}	0,0567	0,0832	0,0010	0,0621
	(-)	(-)	(-)	(-)
	[-]	[3,4616]	[-10,4528]	[0,1730]
ϕ_{11}	0,3492	0,3371	0,3857	0,3486
	(-)	(-)	(-)	(-)

3.6.6 Qualidade de Ajuste

Para avaliar a qualidade do ajuste do modelo de regressão com fração de cura para os dados de retinopatia, utilizou-se a função de sobrevivência estimada por Kaplan-Meier e o ajuste das funções de sobrevivência marginais do modelo de regressão proposto, que encontram-se na Figura 19. Também foi utilizado como medida de qualidade de ajuste o gráfico da função de sobrevivência estimada por Kaplan-Meier estratificada pela covariável tipo de diabetes e o ajuste das funções de sobrevivência marginais do modelo de regressão com fração de cura estratificado pela covariável tipo de diabetes, como mostra a Figura 20.

Ao analisar as Figuras 19 e 20, observa-se o ganho na qualidade do ajuste do modelo ao modelar a proporção de curados presente nos dados.

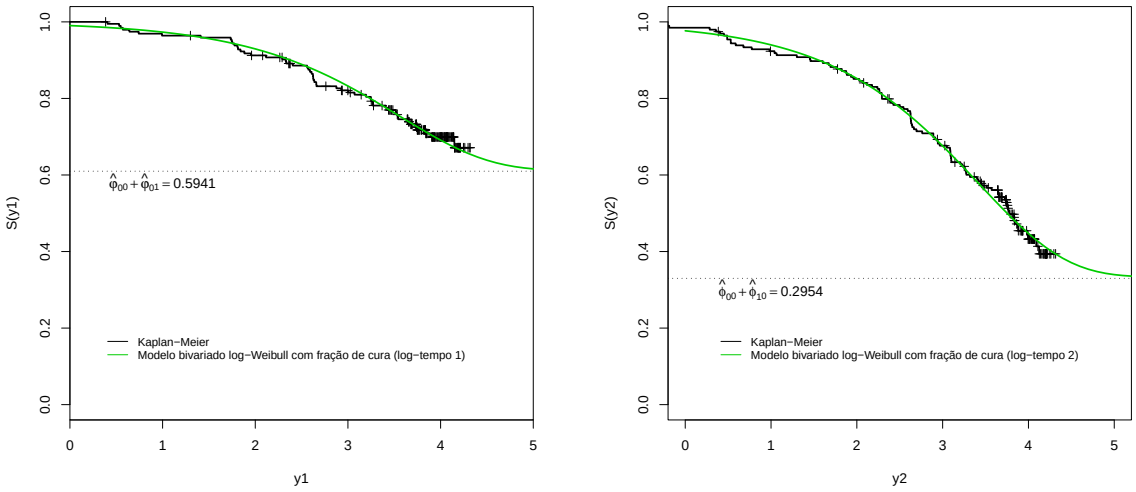

Figura 19 - Curvas de sobrevivência de Kaplan-Meier função de sobrevivência estimada para os dados de retinopatia. (a) função de sobrevivência marginal $S_1(y_1)$, e (b) função de sobrevivência marginal $S_2(y_2)$

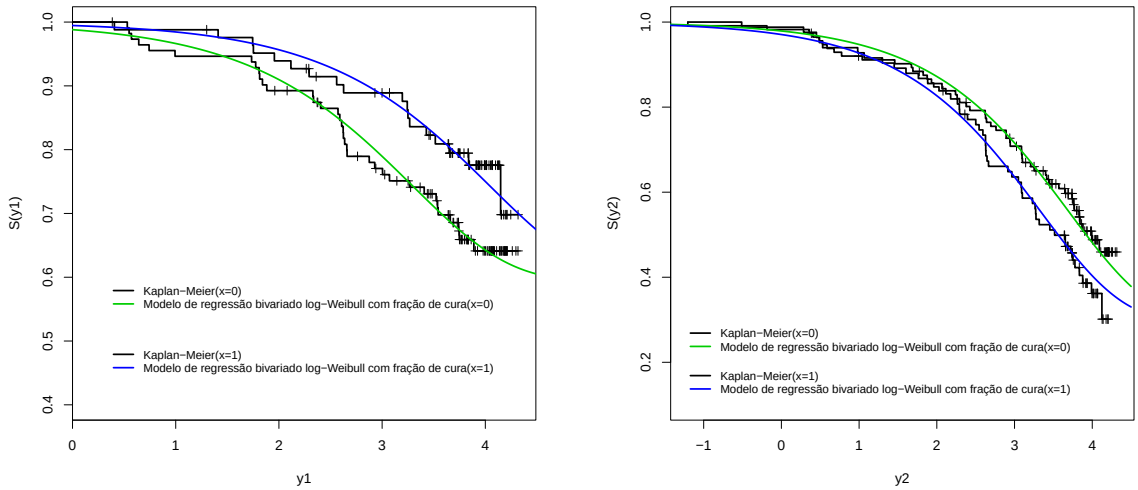

Figura 20 - Curvas de sobrevivência de Kaplan-Meier estratificadas por tipo de diabetes (0: diabetes juvenil, 1: diabetes adulto) e função de sobrevivência estimada para os dados de retinopatia. (a) função de sobrevivência marginal $S_1(y_1|\boldsymbol{x})$, e (b) função de sobrevivência marginal $S_2(y_2|\boldsymbol{x})$

3.7 Conclusões

Neste trabalho, foi proposto um modelo de regressão com fração de cura para dados bivariados por meio de cópulas, sendo este modelo uma modificação do modelo proposto por Wienke et al. (2006). A modificação considereada foi admitir um modelo de locação-escala para os tempos de sobrevivência. Outra contribuição deste trabalho foi propor um método de estimação usando o método de máxima verossimilhança sujeito às restrições lineares nos parâmetros, ao utilizar a função barreira adaptada para implementá-lo.

Uma análise de sensibilidade foi realizada devido a necessidade de verificar se as suposições do modelo foram atendidas e se a presença de observações extremas causaram distorções nos resultados da aplicação realizada. Não foram detectados afastamentos das suposições nem pontos de sensibilidade e o modelo proposto se mostrou adequado para descrever o tempo de sobrevivência bivariado com a presença de uma fração de indivíduos curados na

amostra avaliada.

Uma importante ressalva para este modelo é que problemas computacioanis podem aparecer caso não haja uma proporção de indivíduos curados em ambos os tempos de sobrevivência. Uma segunda ressalva discutida em diversos artigos da área está relacionada a dificuldades na implementação dos procedimentos de estimação em modelos semelhantes. O método estimação implementado utilizando a função barreira adaptada para o modelo proposto, entretanto, mostrou-se muito eficiente.

3.7.1 Propostas para trabalhos futuros

Como possíveis trabalhos futuros podem-se considerar os seguintes temas de pesquisa:

1. Realizar um estudo sobre a distribuição assintótica dos parâmetros com restrições lineares.

2. Desenvolver um teste para verificar a presença de indivíduos imunes.

3. Extender o modelo proposto neste capítulo ao estimar a proporção de indivíduos curados por meio de covariáveis.

4. Considerar outras cópulas arquimedianas como função de sobrevivência bivariada.

Referências

BARRIGA, G. D. C.; LOUZADA-NETO, F.; ORTEGA, E. M. M.; CANCHO, V. G. A bivariate regression model for matched paired survival data: local influence and residual analysis. **Statistical Methods and Applications**, New York, 2010.

BERKSON, J.; GAGE, R.P. Survival curve for cancer patients follwing treatment. **Journal of the American Statistical Association**, Alexandria, v. 47, p. 501-515, 1952.

CANCHO, V.G. **Métodos de monte carlo em análise de sobrevivência.** 1999. 207p. Tese (Doutorado em Estatística)-Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 1999.

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2nd ed. Pacific Grove: Thomson Learning, 2002. 660 p.

CASTRO, M.; CANCHO, V. G.; RODRIGUES, J. A bayesian long-term survival model parametrized in the cured fraction, **Biometrical Journal**, Weinheim, v. 51, n. 3, p. 443-455, 2009.

CHATTERJEE N., SHIH J. A bivariate cure-mixture approach for modeling familial association in diseases, **Biometrics**, Washington, v. 57, p. 779-786, 2001.

CHEN, M. H.; IBRAHIM, J; SINHA,D. A new bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Alexandria, v. 94, p. 909-919, 1999.

CLAYTON, D. G. A model for association in bivariate life-tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, London, v. 65, p. 141-151, 1978.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Edgard Blücher, 2006. 392 p.

COOK, R.D. Detection of influential observations in linear regression. **Technometrics**, Alexandria, v. 19, p. 15-18, 1977.

COOK, R. D.; WEISBERG, S. **Residuals and influence in regression**. New York: Chapman and Hill, 1982. 230 p.

COOK, R.D. Assement of local influence (with discussion). **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 48, n. 2, p. 133-169, 1986.

COOK, R. D.; PEÑA, D.; WEISBERG, S. The likelihood displacement: a unifying principle for influence. **Communications in Statistics: Part Theory and Methods**, New York, v. 17, n. 3, p. 623-640, 1988.

COX, D.R.; OAKES, D. **Analysis of survival data**. London: Chapman and Hall, 1984. 201 p.

ESCOBAR, L.A.; MEEKER, W.Q. Assessing influence in regression analysis with censored data. **Biometrics**, Washington, v. 48, n.2, p. 507-528, 1992.

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.

FAREWELL, V. T. The use mixture models for the analysis of survival data with log-term survivors. **Biometrics**, Washington, v. 38, p. 43-46, 1982.

FAREWELL, V. T. Mixture models in survival analysis: Are they worth the risk? **Canadian Journal Statistical**, Toronto, v. 14, p. 257-262, 1986.

FREUND, J. E. Bivariate Extension of the Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 56, p. 971-977, 1961.

GOLDMAN, A. Survivorship analysis when cure is a possibility: a Monte Carlo study. **Statistics in Medicine**, Chichester, v. 3, p. 153-163, 1984.

GOMES, E. M. C. **Análise de Sensibilidade e resíduos em modelos de regressão com respostas bivariadas por meio de cópulas.** 2007. 103p. Dissertação (Mestre em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

GOURIEROUX, C.; MONFORT, A. **Statistical and econometric models**. Cambridge: Cambridge University Press, 1995. v.2, 526 p.

GREENHOUSE, J.B.; WOLFE, R.A. A competing risks derivation of a mixture model for the analysis of survival data. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 13, p. 3133-3154, 1984.

GU, H; FUNG, W.K. Local influence for the restricted likelihood with applications. **Sankhya: The Indian Journal of Statistical**, Indian, v. 63, pt. 2, p. 250-259, 2001.

GUMBEL, E. J. Bivariate Exponential Distributions. **Journal of the American Statistical Association**, Alexandria, v. 55, p. 698-707, 1960.

HALPERN, J.; BROWN, B. Cure rate models: Power of the log-rank and generalized Wilcoxon tests. **Statistics in Medicine**, Chichester, v. 6, p. 483-489, 1987.

HASHIMOTO, E. M. **Modelo de Regressão para dados com censura intervalar e dados de sobrevivência agrupados** 2008. 121p. Dissertação (Mestre em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

HE, W.; LAWLESS, J. F. Bivariate location-scale models for regression analysis, with applications to lifetime data. **Journal of the Royal Statistical Society**, London, v. 67, n. 1, p. 63-78, 2005.

HOUGAARD, P. Fitting a multivariate failure time distribution **IEEE Transactions on Reliability**, New York, v. 38, p. 444-448, 1989.

HUSTER, W. J.; BROOKMEYER, R.; SELF, S.G. Modelling Paired Survival Data with Covariates **Biometrics**, Washington, v. 45, p. 145-156, 1989.

JOHNSON, R. A.; EVANS, J. W.; GREEN, D. W. Some bivariate distributions for modeling the strength properties of lumber. **United States Department of Agriculture**, Washington, FPL-RP-575, 1999.

KALBFLEISCH, J.D.; PRENTICE, R.L. **The statistical analysis of failure time data**. 2nd ed. New York: John Wiley, 2002. 439 p.

KWAN, C. W; FUNG, W. K. Assessing local influence for specific restricted likelihood: application to factor analysis. **Psychometrika**, New York, v. 63, n. 1, p. 35-46, 1998.

LANGE, K. **Numerical analysis for statisticians**. New York: Springer, 1999. 356 p.

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

LEE. E. T. **Statistical models and for survival data analysis**, 2nd ed., New York: Wiley, 1992. 482 p.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, Washington, v. 54, n. 2, p. 570-582, 1998.

LIANG, K.; STEVEN, G. S.; CHANG, Y. Modelling Marginal Hazards in Multivariate Failure Time Data. **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 55, p. 441-453, 1993.

MALLER, R.; ZHOU, X. **Survival Analysis with Long-Term Survivors**. New York: Wiley, 1996. 278 p.

MOESCHBERGER, M. L. Life tests under dependent competing causes of failure. **Technometrics**, Alexandria, v. 16, p. 39-47, 1974.

NELSEN, R. B. Properties of a one-parameter family of bivariate distributions with specified marginals. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 15, p. 3277-3285, 1986.

NELSON, W. **Applied life data analysis**. New York: Wiley, 1982. 634 p.

NÚÑEZ, J. S. R. **Modelagem Bayesiana para dados de sobrevivência bivariados através de cópulas.** 2005. 101p. Tese (Doutorado em Estatística)- Instituto de Matemática e Estatística, IME/USP, São Paulo, 2005.

ORTEGA, E. M. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. **Computational Statistics and Data Analysis**, New York, 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 Operation Reserch Transactions**, Catalunya, v. 30, n. 2, p. 171-192, 2006.

ORTEGA, E. M. M. ; PAULA, G. A. ; BOLFARINE, H. Deviance residuals in generalized log-gamma regression models with censored observations. **Journal of Statistical Computation and Simulation**, New York, v. 78, p. 747-764, 2008.

ORTEGA, E. M. M. ; CANCHO, V. G.; PAULA, G. A. Generalized log-gamma regression models with cure fraction. **Lifetime Data Analysis**, Boston, v. 15, p. 79-106, 2009a.

ORTEGA, E. M. M. ; RIZZATO, F. B.; DEMÉTRIO, C. G. B. The generalized log-gamma mixture model with covariates: local influence and residual analysis. **Statistical Methods and Application**, New York, v. 18, n. 3, p. 305-331, 2009b.

ORTEGA, E. M. M. ; CANCHO, V. G.; LANCHOS, V. H. A generalized log-gamma mixture models for cure rate: estimation and sensitivity analysis. **Sankhya: The Indian Journal of Statistical**, Indian, v. 71, p. 1-29, 2009c.

PAULA, G.; CYSNEIROS, F. J. A. Local influence under parameter constraints. **Communications in Statistics: Theory and Methods**, New York, v.88, p. 1-23, 2009.

R Development Core Team (2009). **R**: A language and environment for statistical computing. Disponível em:¡http://www.R-project.org¿. Acesso em: 17 maio 2011.

RIZZATO, F.B. **Modelos de Regressão log-gama generalizado com fração de cura.** 2007. 74p. Dissertação (Mestrado em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

SHIH, J. H.; LOUIS, T. A. Inferences on the association parameter in copula models for bivariate survival data. **Biometrics**, Washington, v. 51, p. 1384-1399, 1995.

TARUMOTO, M. H. **Um modelo Weibull bivariado para riscos competitivos.** 2001. 154p. Tese (Doutorado em Matemática Aplicada)- Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 2001.

TIBALDI, F. S. **Modeling of Correlated Data and Multivariate Survival Data.** 2004. 160p. Tese de Doutorado, Universidade de Hasselt, Hasselt, 2004.

WADA, C. Y.; HOTTA, L. K. Restricted alternatives tests in a bivariate exponential model with covariates. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 29, p. 193-210, 2000.

WIENKE A.; LICHTENSTEIN L.; YASHIN A.I. A bivariate frailty model with a cure fraction for modeling familial correlations in diseases, **Biometrics**, Washington, v.59, p.1178-1183, 2003.

WIENKE A.; LOCATELLI I.; YASHIN A.I. The modelling of a cure fraction in bivariate time-to-event data, **Austrian Journal of Statistics**, Austrian, v.35, p.67-76, 2006.

XIE, F.; WEI, B. Diagnostics analysis for log-Birnbaum-Saunders regression models. **Computational Statistics and Data Analysis**, Amsterdam, v. 51, p.4692-4706, 2007.

ZHU, H.; ZHANG, H. A diagnostic procedure based on local influence. **Biometrika**, Cambridge, v. 91, n. 3, p. 579-589, 2004.

4 MODELO DE REGRESSÃO LOG-LINEAR BIVARIADO COM FRAÇÃO DE CURA

Resumo

Neste trabalho foi considerada uma extensão do modelo introduzido por Chen et al. (2002), conhecido como modelo de tempo de promoção bivariado. A proposta de extensão refere-se a incluir covariáveis no logaritmo dos tempos, originando o modelo de regressão log-linear bivariado com fração de cura. O modelo proposto capta como as covariáveis podem influenciar no tempo de sobrevivência e na proporção de indivíduos curados. Para estimar os parâmetros do modelo foi utilizado o método da função barreira adaptada (LANGE, 1999). Uma análise de sensibilidade foi adaptada considerando as metodologias de Influência Global, Influência Local e Influência Local Total para verificar vários aspectos que envolvem a formulação e ajuste do modelo proposto. Finalmente, um conjunto de dados de retinopotia diabética foi analisado sob o modelo de regressão log-linear bivariado com fração de cura.

Palavras-chave: Fração de cura; Modelo de fragilidade; Verossimilhança sujeita a restrição nos parâmetros; Modelos de regressão; Dados bivariados e censurados; Análise de sensibilidade

Abstract

In this work was considered an extension to the model introduced by Chen et al. (2002), known as bivariate promotion time model. The proposed extension relates to the inclusion of covariates in the logarithm of time, creating the log-linear bivariate regression model with cured fraction. The proposed model captures the effects of covariates on the survival time and on the proportion of cured individuals. To estimate the parameters of the model, the adapted barrier function method (LANGE, 1999) was implemented. A sensitivity analysis was adapted considering the methodology of Global Influence, Local Influence and Total Local Influence to check various aspects of the formulation and adjustment of the model. Finally, a diabetic retinopathy data set was analyzed under the log-linear bivariate regression

model with cured fraction.

Keywords: Cured fraction; Bivariate data and censored; Likelihood subject to restriction on the parameters; Frailty model; Regression models; Sensitivity analysis

4.1 Introdução

Os modelos aqui abordados são propostos com objetivo de descrever situações experimentais ou observacionais nas quais o foco é o tempo até a ocorrência de determinado evento de interesse. As principais particularidades propostas para o modelo são referentes a possibilidade de analisar de forma simultânea o tempo até a ocorrência de dois distintos eventos de interesse, a relação entre eles, os efeitos causados por outros fatores ou covariáveis e contemplando ainda uma possível existência de alguns indivíduos no estudo que sejam imunes, curados ou não suscetíveis a vivenciar os eventos considerados.

Na literatura existem alguns modelos que contemplam as particularidades citadas, de forma isolada. A proposta é combinar os diferentes métodos e modelos existentes para tratar as especificidades de forma conjunta. A seguir são apresentados os modelos e métodos que formam a base das idéias a ser combinadas.

Desde 1952, Berkson e Gage trabalhavam com a idéia de uma modelagem para dados de sobrevivência com uma fração de curados no contexto univariado que ficou conhecida como modelo de mistura. Ao introduzir covariáveis nesse modelo de mistura, sua função de sobrevivência não apresenta estrutura de riscos proporcionais e esse fato dificulta os procedimentos computacionais de estimação dos parâmetros.

Como alternativa, Yakovlev et al. (1993) propõem uma nova classe de modelos de fração de cura denominado de tempo de promoção, cuja modelagem é introduzida no contexto biológico. Esse modelo também foi amplamente discutido por Yakovlev et al. (1994), Asselain et al. (1996), Yakovlev e Tsodikov (1996) e a formulação bayesiana foi dada por Chen et al. (1999).

Na abordagem bayesiana, Chen et al. (2002) propõem um modelo multivariado com fração de cura, que prova ser bastante útil para descrever dados multivariados com variáveis aleatórias de tempo de falha conjuntas associadas a uma fração de sobreviventes onde cada variável aleatória de tempo de falha marginal também está associada a uma fração de

indivíduos curados.

Com base no modelo introduzido por Chen et al. (2002) este trabalho considera que o tempo de sobrevivência pode ser influênciado por variáveis explicativas, obetendo-se dessa forma o modelo de regressão log-linear bivariado com fração de cura. A formulação do modelo impõe certas restrições ao espaço paramétrico demandando metodologia específica para estimação adequada. A proposta de Lange (1999) utiliza a função barreira adaptada, que é uma combinação do método barreira logaritmo com o algoritmo EM.

Uma vez que os métodos de estimação dos parâmetros do modelo são conduzidos por técnicas de maximização da função de verossimilhança, é conveniente considerar uma adaptação em alguns métodos de análise de sensibilidade baseados na função de verossimilhança. A análise de sensibilidade considerada neste trabalho para o modelo de regressão log-linear bivariado com fração de cura é composta pelas técnicas de Influência Global, Influência Local e Influência Local Total. São utilizadas algumas idéias de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009) que utilizam Influência Local na estrutura de verossimilhança sujeita a restrições nos parâmetros.

Este capítulo está estruturado de forma que: na seção 4.2 é apresentada uma revisão do modelo com fração de cura univariada seguindo abordagem de Yakovlev et al.(1993). Na seção 4.3 é descrito o modelo de tempo de promoção bivariado com fração de cura. Na seção 4.4 é desenvolvido o modelo de regressão log-linear bivariado com fração de cura. A seção 4.5 abrange uma descrição da metodologia de análise de sensibilidade para o modelo proposto. Na seção 4.6 é apresentada uma aplicação desse modelo. Para finalizar, a seção 4.7 relata as principais conclusões e um direcionamento de continuidade para este trabalho.

4.2 Fração de cura univariada seguindo abordagem de Yakovlev et al.(1993)

O modelo com fração de cura univariado introduzido por Berkson e Gage (1952), descrito na seção 3.2, foi extensivamente discutido na literatura por diversos autores, incluindo Farewell (1982, 1986), Goldman (1984), Greenhouse and Wolfe (1984), Halpern e Brown (1987), Maller e Zhou (1996), Cancho (1999) e Ortega et al. (2009a).

Uma observação importante é que ao introduzir covariáveis na estrutura de fração de cura essa modelagem não apresenta forma de riscos proporcionais para $S_{pop}(t)$. Alterna-

tivamente, Yakovlev et al. (1993) propõe uma nova classe de modelo com fração de cura, conhecida como modelo de tempo de promoção, cuja modelagem é introduzida no contexto biológico. Esse modelo também foi amplamente discutido por Yakovlev et al. (1994), Asselain et al. (1996), Yakovlev e Tsodikov (1996) e a formulação bayesiana foi dada por Chen et al. (1999).

O modelo introduzido por Yakovlev et al. (1993) supõe que para um indivíduo arbitrário da população existem N células carcinogênicas (causas competindo entre si) que podem produzir um câncer detectável, assumindo que N é uma variável aleatória que tem distribuição de Poisson com média θ. Para cada célula carcinogênica está associado um tempo aleatório até a ocorrência de um tumor detectável, sendo esse tempo denotado por $R_l, l = 1, 2, \ldots, N$, e conhecido como tempo de promoção para a l-ésima célula carcinogênica. Dado N, as variáveis aleatórias $R_l, l = 1, 2, \ldots, N$ são consideradas independentes e identicamente distribuídas com função de distribuição $F(.) = 1 - S(.)$. Dessa forma, o tempo observado até a ocorrência do câncer ou tempo de falha pode ser definido pela variável aleatória $T = min\{R_l, 0 \leq l \leq N\}$ em que $P(R_0 = \infty) = 1$.

A função de sobrevivência para T, e portanto, a função de sobrevivência populacional é dada por:

$$
\begin{aligned}
S_{pop}(t) &= P(\text{sem câncer no tempo } t) = P(T > t) \\
&= P\Big[\min(R_0, R_1, \ldots, R_l) > t\Big] = E\Big\{\mathbf{I}_{[\min(R_0,R_1,\ldots,R_l)>t]}\Big\} \\
&= E\left\{E\Big[\mathbf{I}_{[\min(R_0,R_1,\ldots,R_l)>t]}|N\Big]\right\} \\
&= E\{\varphi(N)\} = \sum_{l=0}^{\infty}\varphi(l)P(N=l) \\
&= \sum_{l=0}^{\infty}E\Big\{\mathbf{I}_{[\min(R_0,R_1,\ldots,R_l)>t]}\Big\}P(N=l) \\
&= \sum_{l=0}^{\infty}P\Big[\min(R_0, R_1, \ldots, R_l) > t\Big]P(N=l) \\
&= P(N=0) + \sum_{l=1}^{\infty}P(R_1 > t, R_2 > t, \ldots, R_l > t)P(N=l) \\
&= \exp(-\theta) + \sum_{l=1}^{\infty}P(R_1 > t)P(R_2 > t)\ldots P(R_l > t)P(N=l)
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \exp(-\theta) + \sum_{l=1}^{\infty} [S(t)]^l \frac{\exp(-\theta)(\theta)^l}{l!} \\
& = \exp(-\theta) \left\{ \sum_{l=0}^{\infty} \frac{[S(t)\theta]^l}{l!} \right\} \\
& = \exp(-\theta) \exp[S(t)\theta] \\
& = \exp[-\theta F(t)]. \qquad (44)
\end{aligned}
$$

Ao considerar a função de sobrevivência populacional (44), observa-se que

$$S_{pop}(\infty) = \lim_{t \longrightarrow \infty} S_{pop}(t) = \exp(-\theta) > 0,$$

indicando que (44) é uma função de sobrevivência imprópria, pois conforme o tempo aumenta a $S_{pop}(t)$ não converge para zero, e consequentemente, $S_{pop}(\infty) \equiv P(N = 0) = \exp(-\theta)$ que corresponde a fração de cura induzida pelo modelo (44). Observe ainda que, quando $\theta \longrightarrow \infty$, a fração de cura tende a zero, e por outro lado, quando $\theta \longrightarrow 0$, a fração de cura tende a um. Esse fato pode ser esperado pois, uma vez que θ cresce, a média de N aumenta, diminuindo a probabilidade de cura. Essa observação também é válida no caso de θ decrescente.

Como observado por Yakovlev e Tsodikov (1996), a função de sobrevivência populacional (44) mostra explicitamente a contribuição de duas características distintas de tumor de crescimento para o tempo de falha: o número inicial de células carcinogênicas e o risco de seu progresso. Assim, o modelo reune parâmetros com significado claramente biológico. Além da motivação biológica, o modelo (44) é apropriado para alguns tipos de dados de sobrevivência que não possuem uma explicação biológica, como citada anteriormente, mas possuem uma fração de sobreviventes e podem estar sendo generalizado por um número desconhecido N de riscos competitivos latentes. Sendo assim, o modelo pode ser usado para modelar vários tipos de dados de sobrevivência, incluindo tempo de reincidência, tempo de morte, tempo até a primeira infecção, entre outros.

Supondo que as variáveis $\{R_l, l = 1, 2, \ldots, N\}$ são absolutamente contínuas e denotando sua função de densidade de probabilidade por $f(t)$, a densidade correspondente a (44) é dada por:

$$f_{pop}(t) = -\frac{\partial[S_{pop}(t)]}{\partial t} = \theta f(t) \exp[-\theta F(t)], \qquad (45)$$

em que $f(t) = \frac{d}{dt} F(t)$ é uma função densidade de probabilidade própria.

Ao analisar a função densidade de probabilidade definida em (45), verifica-se que $f_{pop}(t)$ não é uma função de densidade de probabilidade própria, pois considerando as propriedades da função de densidade de probabilidade a integral de $f_{pop}(t)$, $\int_{-\infty}^{\infty} f_{pop}(t) = \theta \exp[-\theta]$, é diferente de 1, fato decorrente de $f_{pop}(t)$ estar associada a $S_{pop}(t)$.

A função de risco associada ao modelo (44) é escrita por;

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{\theta f(t) \exp[-\theta F(t)]}{\exp[-\theta F(t)]} = \theta f(t). \tag{46}$$

Note que $h_{pop}(t) \longrightarrow 0$ tão rápido quanto $t \longrightarrow \infty$.

O modelo de cura (44) revela uma forma atrativa da função de risco, pois (46) é composta pelo produto de dois componentes, θ e $f(t)$. Dessa forma, ao introduzir covariáveis em θ, ou seja, $\theta = \theta(\boldsymbol{x})$, a função de risco (46) caracteriza um modelo de riscos proporcionais, sendo essa a característica que diferencia o modelo de cura (44) do modelo de cura introduzido por Berkson e Gage (1952) dado em (27).

A função de sobrevivência populacional (44) pode ser escrita em termos do modelo de mistura (27), como:

$$S_{pop}(\infty) = \exp(-\theta) + [1 - \exp(-\theta)]S^*(t), \tag{47}$$

em que

$$S^*(t) = P(T > t | N \geq 1) = \frac{\exp[-\theta F(t)] - \exp(-\theta)}{1 - \exp(-\theta)}, \tag{48}$$

é uma função de sobrevivência própria da população não curada.

A equação (47) é um modelo de mistura, como definido por Berkson e Gage, com fração de cura igual a $1 - \phi = \exp(-\theta)$, e função de sobrevivência para a população não curada dada por $S^*(t)$. Isso mostra que todo modelo definido por (44) pode ser escrito como um modelo de mistura (27), com uma família específica de funções de sobrevivência $S^*(t)$ dada por (48). Esse resultado também implica que todo modelo de mistura corresponde a algum modelo da forma (44) para algum θ e $F(.)$.

A função de densidade própria para a população não curada é dada por:

$$f^*(t) = -\frac{d}{dt} S^*(t) = \frac{\exp[-\theta F(t)]}{1 - \exp(-\theta)} \theta f(t),$$

e a função de risco para a população não curada é representada por:

$$\begin{aligned} h^*(t) &= \frac{f^*(t)}{S^*(t)} = \frac{\exp[-\theta F(t)]}{\exp[-\theta F(t)] - \exp(-\theta)} h_{pop}(t) \\ &= \frac{1}{P(T < \infty | T > t)} h_{pop}(t). \end{aligned}$$

Como enunciado anteriormente, se covariáveis são introduzidas em θ, $S_{pop}(t)$ pode ser considerada um modelo com estrutura de riscos proporcionais de Cox, mas esse fato não é válido para $h^*(t)$. Pois, $\frac{1}{P(T<\infty|T>t)}$ nunca será independente de t para alguma $f(t)$ com suporte $(0, \infty)$.

Como apontado por Ibrahim et al. (2001), ao analisar as características do modelo (44), verifica-se que nesse modelo toda população é modelada com estrutura de riscos proporcionais. No modelo de cura (27), apenas o grupo de não curados pode ser modelado com estrutura de riscos proporcionais.

As covariáveis podem ser introduzidas no modelo (44) por meio da relação $\theta = \theta(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\gamma})$, em que $\boldsymbol{x}$ é o vetor de covariáveis e $\boldsymbol{\gamma}$ é o vetor de coeficientes de regressão, ambos de dimensão $p + 1$. Ibrahim et al. (2001) demonstram que inserir as covariáveis dessa forma corresponde a uma ligação canônica no modelo de regressão de Poisson. Assim, as regras de interpretação dos coeficientes de regressão podem ser utilizadas para a população curada e não curada.

Recentemente, Castro et al. (2009) propõem distribuição binomial negativa para o número de células carcinogênicas e relaciona as covariáveis com a fração de cura por meio de uma ligação logística, $\theta = \theta(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\gamma})/[1 + \exp(\boldsymbol{x}^T\boldsymbol{\gamma})]$, em que $\boldsymbol{x}$ é o vetor de covariáveis e $\boldsymbol{\gamma}$ é o vetor de coeficientes de regressão, ambos de dimensão $p + 1$. Na próxima seção, é descrito o modelo de tempo de promoção bivariado com fração de cura proposto por Chen et al. (2002) sob o enfoque bayesiano.

4.3 Modelo de tempo de promoção bivariado com fração de cura

Neste trabalho, já discutimos o interesse e a importância de modelar conjuntamente duas ou mais variáveis aleatórias que representam tempos de falha. Como uma alternativa ao modelo com fração de cura univariado discutido por Yakovlev et al. (1993), Chen et al. (2002) propõem um modelo com fração de cura multivariado. Com esse modelo é possível

descrever dados multivariados com variáveis aleatórias de tempo de falha associadas à uma fração conjunta de sobreviventes e cada variável aleatória do tempo de falha marginal também está associada à uma fração de indivíduos curado.

Chen et al.(2002) descrevem o modelo bivariado com fração de cura, considerando $T = (T_1, T_2)^T$ o tempo de falha bivariado, em que T_1 pode ser considerado o tempo até a ocorrência do primeiro evento de interesse e T_2 pode ser considerado o tempo até a ocorrência do segundo evento de interesse. É importante notar que os eventos denotados por T_1 e T_2 não são ordenados, de forma que a denominação de primeiro e segundo evento de interesse não está relacionada à ordem de ocorrência dos eventos mas somente para diferenciar as diferentes variáveis resposta consideradas.

Para um indivíduo arbitrário da população supõe-se que existe um par (N_1, N_2) de variáveis latentes representando o número de células carcinogênicas associadas à (T_1, T_2), respectivamente. É considerado que N_k tem distribuição Poisson com média $\theta_k m$, sendo m um componente de fragilidade para induzir a correlação entre as variáveis latentes (N_1, N_2) de forma que N_1 e N_2 são condicionalmente independentes dado m.

Assume-se distribuição estável positiva com parâmetro α para o componente de fragilidade m, em que $0 < \alpha < 1$. Esse componente m pode assumir várias distribuições, Chen et al. (2002) justificam utilizar a distribuição estável positiva por ser uma distribuição bastante flexível para dados de sobrevivência multivariados.

Para cada par (N_1, N_2) de quantidades de células carcinogênicas estão associados os tempos de promoção bivariados (R_{1l}, R_{2l}) até o l-ésimo fator de risco latente produzir um tumor detectável, sendo (R_{1l}, R_{2l}) conhecido como o tempo latente para (T_{1l}, T_{2l}). Os vetores aleatórios R_{1l} e R_{2l}, $l = 1, 2, \ldots, N_k$ são assumidos independentes e identicamente distribuídos, com função distribuição acumulada $F(t_k|\boldsymbol{\lambda}_k) = 1 - S(t_k|\boldsymbol{\lambda}_k)$, sendo $k = 1, 2$, $\boldsymbol{\lambda}_k$ o vetor de parâmetros desconhecidos da distribuição adotada para os tempos de falha e $F(t_k|\boldsymbol{\lambda}_k)$ é independente de N_k.

O tempo de sobrevivência observado é definido como a variável aleatória $T_k = min(R_{kl}, 0 \leq l \leq N_k)$, em que a $P(R_0 = \infty) = 1$ representa a cura e, N_k é independente da sequência $R_{k1}, R_{k2}, \ldots$, para $k = 1, 2$. De Chen et al. (2002), a função de sobrevivência

populacional condicionada ao componente de fragilidade m é escrita por:

$$\begin{aligned} S_{pop}(t_1,t_2|m) &= \prod_{k=1}^{2}[P(N_k=0)+P(R_{k1}>t_k,\ldots,R_{kN}>t_k,N_k\geq 1)] \\ &= \prod_{k=1}^{2}\left[\exp(-m\theta_k)+\left(\sum_{r=1}^{\infty}S(t_k|\boldsymbol{\lambda}_k)^r\frac{(m\theta_k)^r}{r!}\exp(-m\theta_k)\right)\right] \\ &= \prod_{k=1}^{2}[\exp(-m\theta_k+m\theta_k S(t_k|\boldsymbol{\lambda}_k))] \\ &= \exp[-m(\theta_1F(t_1|\boldsymbol{\lambda}_1)+\theta_2F(t_2|\boldsymbol{\lambda}_2))], \end{aligned} \tag{49}$$

em que $P(N_k=0)=P(T_k=\infty)=\exp(-\theta_k)$, para $k=1,2$. A variável de fragilidade m induz a correlação de T_1 e T_2 ao adicionar a mesma variação extra Poisson para N_1 e N_2 por meio das respectivas médias $\theta_1 m$ e $\theta_2 m$. É importante notar que N_k e R_{kl} apenas facilitam a construção do modelo e não necessitam de alguma interpretação biológica ou física para o modelo (49) ser válido.

Ao assumir distribuição estável positiva com parâmetro α para o componente de fragilidade m, a função de sobreviência polulacional não condicionada a m será encontrada utilizando a transformação de Laplace de m que é dada por $E(\exp(-sm))=\exp(-s^\alpha)$. Ao aplicar a transformação de Laplace na função (49), a função de sobrevivência populacional não condicionada a m é expressa por:

$$S_{pop}(t_1,t_2)=\exp\{-[\theta_1F(t_1|\boldsymbol{\lambda}_1)+\theta_2F(t_2|\boldsymbol{\lambda}_2)]^\alpha\}, \tag{50}$$

tendo estrutura de riscos proporcionais ao introduzir covariáveis no modelo por meio de (θ_1,θ_2).

A oconsiderar distribuição estável positiva com parâmetro α para o componente de fragilidade m, na função de sobrevivência (50) o parâmetro α, $0<\alpha<1$, mede a associação entre T_1 e T_2. Valores pequenos de α indicam alta associação, e quando $\alpha\longrightarrow 1$, implica baixa associação entre T_1 e T_2. Uma medida global de dependência, como o coeficente de correlação τ de Kendall e Pearson não é bem definido para o modelo (50) devido a função de sobrevivência imprópria.

Qunado T_1 e T_2 crescem conjuntamente, mostra-se que a função de sobrevivência populacional (50) é dada por,

$$S_{pop}(\infty,\infty)=\exp[-(\theta_1+\theta_2)^\alpha],$$

em que o termo $\exp[-(\theta_1+\theta_2)^{\alpha}]$ representa a fração de cura conjunta.

As funções de sobrevivência marginais de (50) são

$$S_1(t) = \lim_{t_2 \to -\infty} S_{pop}(t_1, t_2) = \exp\{-\theta_1^{\alpha}[F(t_1|\boldsymbol{\lambda}_1)]^{\alpha}\}$$

e

$$S_2(t) = \lim_{t_1 \to -\infty} S_{pop}(t_1, t_2) = \exp\{-\theta_2^{\alpha}[F(t_2|\boldsymbol{\lambda}_2)]^{\alpha}\},$$

com probabilidade de cura $\exp(-\theta_k^{\alpha})$ para T_k, $k = 1, 2$, sendo importante notar que cada função de sobrevivência marginal tem estrutura de riscos proporcionais ao introduzir covariáveis em θ_k.

As covariáveis são introduzidas no modelo (50) por meio da média de cada N_k, ao considerar $\theta_k = \theta_k(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\gamma}_k)$, ou $\theta_k = \theta_k(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\gamma}_k)/[1+\exp(\boldsymbol{x}^T\boldsymbol{\gamma}_k)]$, em que $\boldsymbol{x} = (x_0, x_1, x_2, \ldots, x_p)^T$ denota o vetor de covariáveis e $\boldsymbol{\gamma}_k = (\gamma_{0k}, \gamma_{1k}, \ldots, \gamma_{pk})^T$ é o vetor dos coeficientes de regressão, ambos de dimensão $(p+1)$.

4.4 Modelo de regressão log-linear bivariado com fração de cura

Em muitas situações práticas, é comum que o tempo de sobrevivência seja influenciado por uma ou mais covariáveis. Por exemplo, na área médica o tempo de sobrevivência de um paciente pode estar relacionado com tipo de tumor ou com tratamento ao qual é submetido. Na indústria, o tempo de sobrevivência de um determinado equipamento pode ser influenciado pelo tipo de matéria prima utilizado em sua fabricação ou pelo nível de voltagem ao qual é submetido.

Nas subsequentes definições do modelo muitas vezes será utilizado o exemplo de células carcinogênicas para contextualizar a utilização do mesmo, mas deve ficar registrado que o modelo não é restrito para tal tipo de exemplo sendo útil para as diversas aplicações de análise de sobrevivência.

Em situações que se deseja modelar o tempo até a ocorrência de um evento de interesse, duas importantes classes de modelos de regressão são consideradas: modelos de locação-escala e modelos de riscos proporcionais. Na literatura há um grande número de trabalhos publicados para ambos os tipos de modelo de regressão, veja por exemplo Klein e Moeschberger (1997), Lawless (2003), entre outros.

Neste trabalho, o modelo de locação-escala é extendido para o caso de dados de sobrevivência bivariados. Considere T_1 e T_2 duas variáveis aleatórias não negativas denotando o tempo de sobrevivência para cada um dos eventos de interesse, e $Y_k = \log(T_k)$ o logaritmo do tempo de falha do k-ésimo evento de interesse, para $k = 1, 2$. Ao assumir que há uma relação linear entre a variável Y_k e um vetor de variáveis explicativas $\boldsymbol{x} = (x_0, x_1, x_2, \ldots, x_p)^T$, o modelo de locação-escala é representado por:

$$\mathrm{Y_k} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{k}} + \sigma_{\mathrm{k}}\mathrm{Z_k}, \tag{51}$$

em que $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})^T$ denota o vetor de coeficientes de regressão associado ao vetor $\boldsymbol{x}$ de variáveis explicativas com $p + 1$ componentes, $\sigma_k > 0$ o parâmetro de escala e Z_k o erro aleatório com distribuição que não depende de $\boldsymbol{x}$.

Em algumas situações existem indivíduos que não vivenciam o evento de interesse mesmo após um período longo de acompanhamento, e por isso esses indivíduos são considerados imunes, curados ou não suscetíveis. Neste capítulo é proposto o modelo de regressão log-linear bivariado com fração de cura ao considerar que existe uma fração de sobreviventes associada ao logaritmo do tempo de falha de cada evento de interesse, sendo os logaritmos dos tempos de falha modelados seguindo a metodologia proposta por Chen et al. (2002) descrita na seção 4.3.

Como descrito anteriormente, para um indivíduo arbitrário da população supõe-se que existe um número de células carcinogênicas (causas competindo entre si) denotado por (N_1, N_2), e assume-se que N_k tem distribuição Poisson com média $\theta_k m$, em que a quantidade m representa um componente de fragilidade, com distribuição estável positiva, que induz a correlação entre as variáveis latentes N_1 e N_2.

Para cada célula carcinogênica associada a um evento de interesse existe o logaritmo do tempo de promoção até a ocorrência de um tumor detectável, sendo esse logaritmo do tempo representado por $(\log(R_{1l}), \log(R_{2l}))$. Os vetores aleatórios $\log(R_{kl}), l = 1, 2, \ldots, N_k$ são assumidos independentes e identicamente distribuídos com função distribuição acumulada $F(y_k|\boldsymbol{x}) = 1 - S(y_k|\boldsymbol{x})$, independente de N_k, pertencente a família de distribuições de locação e escala e com um vetor de covariáveis $\boldsymbol{x}$.

O logarítmo do tempo de sobrevivência observado é definido como a variável aleatória $Y_k = \log(T_k) = \min(\log(R_{kl}), 0 \leq l \leq N_k)$, em que a $P(\log(R_0) = \infty) = 1$ representa

a cura e, N_k é independente da sequência $\log(R_{k1}), \log(R_{k2}), \ldots$, para $k = 1, 2$. Com base no trabalho de Chen et al. (2002), a função de sobrevivência populacional de Y_1 e Y_2 condicionada ao componente de fragilidade m, é descrita por:

$$\begin{aligned} S_{pop}(y_1, y_2|m) &= \prod_{k=1}^{2}[P(N_k = 0) + P(\log(R_{k1}) > y_k, \ldots, \log(R_{kN}) > y_k, N_k \geq 1)] \\ &= \prod_{k=1}^{2}\left[\exp(-m\theta_k) + \left(\sum_{r=1}^{\infty} S(y_k|\boldsymbol{x})^r \frac{(m\theta_k)^r}{r!}\exp(-m\theta_k)\right)\right] \\ &= \prod_{k=1}^{2}[\exp(-m\theta_k + m\theta_k S(y_k|\boldsymbol{x}))] \\ &= \exp[-m(\theta_1 F(y_1|\boldsymbol{x}) + \theta_2 F(y_2|\boldsymbol{x}))], \end{aligned} \tag{52}$$

em que $P(N_k = 0) = P(Y_k = \infty) = \exp(-\theta_k)$, para $k = 1, 2$. Assumindo distribuição estável positiva com parâmetro α para o componente de fragilidade m, a função de sobrevivência populacional não condicionada a m será obtida ao considerar a transformação de Laplace de m dada por $E(\exp(-sm)) = \exp(-s^\alpha)$. Ao aplicar a transformção de Laplace na função (52), a função de sobrevivência populacional não condicionada a m para o modelo de regressão log-linear bivariado com fração de cura é expressa por:

$$S_{pop}(y_1, y_2|\boldsymbol{x}) = \exp\{-[\theta_1 F(y_1|\boldsymbol{x}) + \theta_2 F(y_2|\boldsymbol{x})]^\alpha\}. \tag{53}$$

Uma possível associação existente entre Y_1 e Y_2 é determinada pelo parâmetro escalar $\alpha \in (0, 1)$, que representa alta associação entre Y_1 e Y_2 quando α assume valores pequenos, e baixa associação quando $\alpha \longrightarrow 1$. Devido a função (53) ser uma função de sobrevivência imprópria, uma medida global de dependência como o coeficente de correlação τ de Kendall e Pearson não é bem definida.

A fração de cura conjunta do modelo (53), é expressa por

$$S_{pop}(\infty, \infty) = \exp[-(\theta_1 + \theta_2)^\alpha]$$

e as funções de sobrevivência marginais são

$$S_1(y_1|\boldsymbol{x}) = \lim_{y_2 \to -\infty} S_{pop}(y_1, y_2) = \exp\{-\theta_1^\alpha [F(y_1|\boldsymbol{x})]^\alpha\}$$

e

$$S_2(y_2|\boldsymbol{x}) = \lim_{y_1 \to -\infty} S_{pop}(y_1, y_2) = \exp\{-\theta_2^\alpha [F(y_2|\boldsymbol{x})]^\alpha\},$$

com probabilidade de cura $\exp(-\theta_k^{\alpha})$ para Y_k, $k = 1, 2$. A grande vantagem deste capítulo é propor um modelo que consiga captar o efeito das covariáveis no logaritmo do tempo de sobrevivência observado e na proporção de indivíduos curados por meio da relação $\theta_k = \theta_k(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\gamma}_k)$, em que $\boldsymbol{x} = (x_0, x_1, x_2, \ldots, x_p)^T$ denota o vetor de covariáveis com dimensão $(p+1)$ e $\boldsymbol{\gamma}_k = (\gamma_{0k}, \gamma_{1k}, \ldots, \gamma_{pk})^T$ denota o vetor $(p+1)$-dimensional dos coeficientes de regressão associados a $\boldsymbol{x}$ referentes a fração de cura.

4.4.1 Inferência para o modelo de regressão log-linear bivariado com fração de cura

Ao considerar n indivíduos e para cada indivíduo i, $i = 1, \ldots, n$, tem-se associado uma amostra observada contendo variáveis $(y_{1k}, \delta_{1k}, \boldsymbol{x}_1), \ldots, (y_{nk}, \delta_{nk}, \boldsymbol{x}_n)$, sendo y_{ik} o logaritmo do tempo do k-ésimo evento de interesse do i-ésimo indivíduo, δ_{ik} a respectiva variável indicadora de censura no k-ésimo evento de interesse do i-ésimo indivíduo e $\boldsymbol{x}_i$ o vetor de covariáveis associado ao i-ésimo indivíduo por meio do logaritmo do tempo do k-ésimo evento de interesse e pela fração de sobreviventes, em que $k = 1, 2$ e $i = 1, 2, \ldots, n$.

O logaritmo da função de verossimilhança para o modelo de regressão log-linear bivariado com fração de cura é obtido ao considerar a função de verossimilhança descrita em Lawless (2003), representado por:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n} \left\{ \delta_{i1}\delta_{i2} f_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) + \delta_{i1}(1-\delta_{i2}) \left[\frac{-\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}} \right] + \right. \\ &\left. (1-\delta_{i1})\delta_{i2} \left[\frac{-\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}} \right] + (1-\delta_{i1})(1-\delta_{i2}) S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) \right\} \end{aligned} \quad (54)$$

em que $S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)$ é a função de sobrevivência populacional definida na equação (53), a função densidade conjunta de (y_{i1}, y_{i2}) é dada por $f_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) = \frac{\partial^2 S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}\partial y_{i2}}$, $\boldsymbol{\varphi} = (\alpha, \boldsymbol{\beta}_k^T, \sigma_k^T, \boldsymbol{\gamma}_k^T)^T$ é o vetor de parâmetros desconhecidos, sendo que $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$ e $\boldsymbol{\gamma}_k = (\gamma_{0k}, \gamma_{1k}, \ldots, \gamma_{pk})^T$, para $k = 1, 2$.

Para maximizar o logaritmo da função de verossimilhança definido em (54) sujeito às seguintes restrições nos parâmetros, $\sigma_k > 0$ e $0 < \alpha < 1$, é necessário considerar o método da função barreira adaptada (LANGE, 1999) que é uma combinação do método barreira logaritmo com o algoritmo EM. Para escrever o logaritmo da função de verossimi-

lhança utilizando o método da função barreira adaptada considere o vetor de parâmetros $\boldsymbol{\varphi}$ sob 4 restrições de inequações lineares $\boldsymbol{u}_j^T\boldsymbol{\varphi} - c_j \geq 0$, em que $\boldsymbol{u}_j, j = 1, 2, \ldots, 4$ são vetores de dimensão 11×1 e c_j são escalares assumindo valores 0 ou 1 dependendo da restrição de interesse. O vetor $\boldsymbol{u}_j$ é escrito por $(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, assumindo valor 1 na posição que o parâmetro de interesse encontra-se.

O logaritmo da função de verossimilhança sujeita as restrições lineares é representado por

$$l_R(\boldsymbol{\varphi}, \vartheta) = l(\boldsymbol{\varphi}) + \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T\boldsymbol{\varphi} - c_j) \tag{55}$$

em que o parâmetro de ajuste é uma constante positiva, $\vartheta > 0$, $\boldsymbol{u}_j^T\boldsymbol{\varphi} - c_j$ é o conjunto de restrições de inequações lineares para $j = 1, 2, \ldots, q$, $\boldsymbol{\varphi} = (\alpha, \boldsymbol{\beta}_k^T, \sigma_k^T, \boldsymbol{\gamma}_k^T)^T$, $\boldsymbol{\beta}_k^T = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk})$, e $\boldsymbol{\gamma}_k = (\gamma_{0k}, \gamma_{1k}, \ldots, \gamma_{pk})^T$.

O estimador de máxima verossimilhança sujeito às restrições nos parâmentros, $\hat{\boldsymbol{\varphi}}$, pode ser obtido numericamente ao maximizar o logarítmo da função de verossimilhança definida em (55). Neste trabalho o *software* R (R DEVELOPMENT CORE TEAM, 2009) foi utilizado para obter $\hat{\boldsymbol{\varphi}}$ por meio da função *constrOptim*. Maiores detalhes sobre o método da função Barreira adaptada ver, por exemplo, Lange (1999).

Procedimentos inferênciais para os parâmetros com restrição não serão considerados e o tema pode ser abordado em pesquisas futuras. Sob certas condições de regularidade, propriedades assintóticas são consideradas para realizar inferências dos parâmetros de regressão, $\boldsymbol{\varphi}^* = (\boldsymbol{\beta}_k^T, \boldsymbol{\gamma}_k^T)^T$. Estimativas da matriz de covariância para os estimadores de máxima verossimilhança $\hat{\boldsymbol{\varphi}^*}$ podem ser obtidas utilizando a matriz hessiana. Nesse caso, os estimadores de máxima verossimilhança para os parâmetros de regressão têm distribuição assintótica normal com média $\boldsymbol{\varphi}^*$ e matriz de covariância dada por $\boldsymbol{I}^{-1}(\boldsymbol{\varphi}^*)$ com $\boldsymbol{I}(\boldsymbol{\varphi}^*) = E[\ddot{\boldsymbol{L}}_R(\boldsymbol{\varphi}^*)]$, em que $\ddot{\boldsymbol{L}}_R(\boldsymbol{\varphi}^*) = -\left\{\frac{\partial^2 l_R(\boldsymbol{\varphi}, \vartheta)}{\partial \boldsymbol{\varphi}^* \partial \boldsymbol{\varphi}^{*T}}\right\}$.

Para dados de sobrevivência, sabe-se que não é possível calcular $\boldsymbol{I}(\boldsymbol{\varphi}^*)$ devido à presença de observações censuradas, então pode-se utilizar, alternativamente, a matriz $[\ddot{\boldsymbol{L}}(\boldsymbol{\varphi}^*)]$ avaliada em $\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}^*}$, denominada matriz de informação observada, que é um estimador consistente de $\boldsymbol{I}(\boldsymbol{\varphi}^*)$. A diagonal principal da matriz $\ddot{\boldsymbol{L}}^{-1}(\boldsymbol{\varphi}^*)$ é utilizada como uma estimativa para a variância dos estimadores.

Um intervalo de confiança de $(1-\alpha)100\%$ para um parâmetro $\boldsymbol{\varphi}^*_{lk}$, em que $l = 0, 1, 2, \ldots, p$ e $k = 1, 2$, é expresso por:

$$\hat{\boldsymbol{\varphi}^*_{lk}} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\boldsymbol{\varphi}^*_{lk}})}.$$

Para realizar testes de hipóteses sobre os parâmetros utiliza-se, por exemplo, a estatística definida por:

$$z = \frac{\hat{\boldsymbol{\varphi}^*_{lk}} - \boldsymbol{\varphi}^*_{l0}}{\sqrt{\widehat{Var}(\hat{\boldsymbol{\varphi}^*_{lk}})}},$$

em que $Z \sim N(0,1)$ e $\boldsymbol{\varphi}^*_{l0}$ é o valor do parâmetro considerado na hipótese nula do teste.

4.5 Análise de sensibilidade

A análise de sensibilidade tem como objetivo estudar o comportamento do modelo proposto após esse ser ajustado a um conjunto de dados, ou seja, verificar se as suposições do modelo são válidas e identificar características inesperadas nos dados que possam influênciar as conclusões obtidas.

A metodologia proposta por Cook (1977), conhecida como Influência Global, ou deleção de casos, considera a influência do i-ésimo indivíduo sobre as estimativas dos parâmetros. Entretanto, quando a deleção de casos é utilizada todas as informações de um único indivíduo são deletadas. Logo fica difícil verificar se aquele indivíduo tem alguma influência sobre algum aspecto específico do modelo.

Uma solução encontrada por Cook (1986) está na metodologia de Influência Local, que tem o objetivo de avaliar a influência conjunta das observações sob pequenas mudanças (perturbações) no modelo. Existem inúmeras formas de perturbar o modelo proposto, cuja escolha deve levar em consideração quais os aspectos da análise se deseja monitorar, lembrando que os esquemas de perturbação devem ser interpretáveis. Lesaffre e Verbeke (1998) desenvolvem a medida de Influência Local Total.

Na literatura de modelos com fração de cura na abordagem de Yakovlev et al. (1993) e Asselain et al. (1996), existem vários trabalhos com diferentes abordagens utilizando as medidas de influência citadas anteriormente. Por exemplo, Mizoi (2004) estuda modelos de sobrevivência com fração de cura e aplica uma análise de Influência Local nos modelos estudados; Cancho et al. (2009) realizam uma análise de resíduo e de Influência Local para o modelo

de regressão log-Weibull-exponenciado com fração de cura; Ortega et al. (2009a) desenvolvem uma análise de sensibilidade e resíduo para o modelo de regressão log-gama generalizado com fração de cura. Contudo, quando considera-se o método de máxima verossimilhança com restrição nos parâmetros, destaca-se os trabalhos de Kwan e Fung (1998), Gu e Fung (2001) e Paula e Cysneiros (2009). Ao considerar esses trabalhos, nas próximas seções encontram-se as medidas de análise de Influência Global e Local propostas por Cook (1977, 1986), sob o enfoque da metodologia de verossimilhança sujeita as restrições nos parâmetros.

4.5.1 Influência Global sob verossimilhança restrita

A metodologia de deleção de casos é considerada para avaliar o efeito da i-ésima observação nas estimativas, e o quanto a deleção de um caso pode alterar os resultados do modelo proposto.

O modelo caso deleção para a variável aleatória contínua $Y_k = \log(T_k)$, com $k = 1, 2$, considerando o modelo (51) é representado por:

$$\boldsymbol{y}_{(i)k} = \boldsymbol{x}_{(i)k}^T \boldsymbol{\beta}_k - \sigma_k \boldsymbol{z}_{(i)k}, \qquad i = 1, \ldots, n, \tag{56}$$

em que o subscrito $(i)k$ representa que a i-ésima observação do k-ésimo evento foi retirada da amostra. O logaritmo da função de verossimilhança com restrição nos parâmetros é denotado por $l_{R(i)}(\boldsymbol{\varphi}, \vartheta)$. Seja $\hat{\boldsymbol{\varphi}}_{(i)}$ o estimador de máxima verossimilhança obtido a partir de $l_{R(i)}(\boldsymbol{\varphi}, \vartheta)$. A influência da i-ésima observação no estimador de máxima verossimilhança sujeito às restrições é avaliada por meio da diferença $\hat{\boldsymbol{\varphi}}_{(i)} - \hat{\boldsymbol{\varphi}}$. Se essa diferença $\hat{\boldsymbol{\varphi}}_{(i)} - \hat{\boldsymbol{\varphi}}$ é relativamente grande, as observações podem ser consideradas influentes. Nessa situação, deve-se analisar com prudência as observações consideradas influentes.

Uma vez que $\hat{\boldsymbol{\varphi}}$ é um vetor, avaliar a diferença $\hat{\boldsymbol{\varphi}}_{(i)} - \hat{\boldsymbol{\varphi}}$ não é trivial e algumas medidas precisam ser definidas. Uma primeira medida de Influência Global é definida como a norma padronizada de $\hat{\boldsymbol{\varphi}}_{(i)} - \hat{\boldsymbol{\varphi}}$, conhecida como a Distância de Cook Generalizada e é dada por:

$$GD_i(\boldsymbol{\varphi}) = (\hat{\boldsymbol{\varphi}}_{(i)} - \hat{\boldsymbol{\varphi}})^T \boldsymbol{M} (\hat{\boldsymbol{\varphi}}_{(i)} - \hat{\boldsymbol{\varphi}}),$$

em que, podem ser consideradas várias escolhas de $\boldsymbol{M}$, segundo Cook e Weisberg (1982).

Entretanto, as escolhas mais utilizadas entre os pesquisadores é considerar $\boldsymbol{M} = -\ddot{L}(\hat{\boldsymbol{\varphi}})$ ou $\boldsymbol{M} = [-\ddot{L}(\hat{\boldsymbol{\varphi}})]^{-1}$.

Outra medida utilizada para verificar a existência de pontos influentes no modelo, é conhecida como Afastamento da Verossimilhança, expressa por:

$$LD_i(\boldsymbol{\varphi}) = 2[l_R(\hat{\boldsymbol{\varphi}}, \vartheta) - l_{R(i)}(\hat{\boldsymbol{\varphi}}, \vartheta)].$$

4.5.2 Influência Local sob verossimilhança restrita

Nesta seção uma metodologia de Influêncial local sob verossimilhança restrita para o modelo de regressão log-linear bivariado com fração de cura é considerada seguindo a mesma sequência da análise de Influêncial local sob verossimilhança restrita desenvolvida na seção 2.4.3 para o modelo de regressão Kumaraswamy Weibull bivariado.

Ao considerar um vetor de perturbações $\boldsymbol{w}$ de dimensão $b \times 1$, o logaritmo da função de verossimilhança perturbada sujeita as restrições lineares é definido por

$$\mathrm{l_R}(\boldsymbol{\varphi}, \vartheta|\boldsymbol{w}) = \mathrm{l}(\boldsymbol{\varphi}|\boldsymbol{w}) + \vartheta \sum_{\mathrm{j}=1}^{\mathrm{q}} (\boldsymbol{u}_{\mathrm{j}}^{\mathrm{T}} \boldsymbol{\varphi} - \mathrm{c_j}) \tag{57}$$

em que a constante $\vartheta > 0$ é o multiplicador do termo barreira. Da equação (57) verifica-se que as restrições não são alteradas pelo esquema de perturbação, que garante solução no subespaço paramétrico. Se $\boldsymbol{w}_0$ denotar o vetor de não perturbação, então $l_R(\boldsymbol{\varphi}, \vartheta|\boldsymbol{w}) = l_R(\boldsymbol{\varphi}, \vartheta)$.

Para cada um dos esquemas de perturbação é necessário obter a matriz $\boldsymbol{\Delta}$, com componentes $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_\alpha, \boldsymbol{\Delta}_{\boldsymbol{\beta}_k}, \boldsymbol{\Delta}_{\sigma_k}, \boldsymbol{\Delta}_{\boldsymbol{\gamma}_k})^T$, definida por:

$$\Delta = \left(\Delta_{vi}\right)_{\left[(4p+7)\times n\right]} = \left(\frac{\partial^2 l_R(\boldsymbol{\varphi}, \vartheta|\boldsymbol{w})}{\partial \varphi_v \partial w_i}\right)_{\left[(4p+7)\times n\right]},$$

em que $v = 1, 2, \ldots, 4p + 7$ e $i = 1, 2, \ldots, n$, utilizando-se o modelo definido em (53) e o logaritmo da função de verossimilhança sujeita as restrições lineares (57). Neste trabalho as matrizes referentes a cada esquema de perturbação são obtidas numericamente.

A seguir são descritos alguns esquemas de perturbação interpretáveis para o modelo de regressão log-linear bivariado com fração de cura. A idéia de perturbar o modelo seguindo o esquema de perturbação de casos pode ser interpretada como uma flexibilização da deleção de casos, de forma que a influência conjunta das observações pode ser investigada, e os

casos de perturbação das respostas ou covariáveis podem ser interpretados como uma forma de detecção de pontos atípicos, com erro ou simplesmente mal modelados.

4.5.2.1 Perturbação de casos

Ao considerar o vetor de perturbação $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^T$, o logaritmo da função de verossimilhança perturbada sujeita as restrições lineares é expresso por:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n} w_i \left\{ \delta_{i1}\delta_{i2} f_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) + \delta_{i1}(1-\delta_{i2}) \left[\frac{-\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}} \right] \right\} + \\ &\quad \sum_{i=1}^{n} w_i \left\{ (1-\delta_{i1})\delta_{i2} \left[\frac{-\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}} \right] + (1-\delta_{i1})(1-\delta_{i2}) S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) \right\} + \\ &\quad \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T \boldsymbol{\varphi} - c_j), \end{aligned}$$

em que o vetor correspondente à não perturbação é o vetor $\boldsymbol{w}_0 = (1, \ldots, 1)^T$, n-dimensional.

4.5.2.2 Perturbação da variável resposta

Para esse esquema de perturbação considera-se que cada variável resposta y_{i1} e y_{i2} é perturbada como $y_{i1}^* = y_{i1} + S_{y_1} w_i$ e $y_{i2}^* = y_{i2} + S_{y_2} w_i$, em que S_{y_k} são fatores de escala que podem ser a estimativa do desvio padrão da variável Y_k, $k = 1, 2$ e $w_i \in \mathbf{R}$. O logaritmo da verossimilhança perturbada sujeita as restrições lineares é dado por:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n} \left\{ \delta_{i1}\delta_{i2} f_{pop}(y_{i1}^*, y_{i2}^*|\boldsymbol{x}_i) + \delta_{i1}(1-\delta_{i2}) \left[\frac{-\partial S_{pop}(y_{i1}^*, y_{i2}^*|\boldsymbol{x}_i)}{\partial y_{i1}^*} \right] \right\} + \\ &\quad \sum_{i=1}^{n} \left\{ (1-\delta_{i1})\delta_{i2} \left[\frac{-\partial S_{pop}(y_{i1}^*, y_{i2}^*|\boldsymbol{x}_i)}{\partial y_{i2}^*} \right] + (1-\delta_{i1})(1-\delta_{i2}) S_{pop}(y_{i1}^*, y_{i2}^*|\boldsymbol{x}_i) \right\} + \\ &\quad \vartheta \sum_{j=1}^{q} (\boldsymbol{u}_j^T \boldsymbol{\varphi} - c_j), \end{aligned}$$

em que $y_{ik}^* = y_{ik} + S_{y_k} w_i$ e o vetor de não perturbação $\boldsymbol{w}_0 = (0, \ldots, 0)^T$. Nesse caso, podem ser consideradas três possibilidades de perturbação, ou seja, perturbar apenas a variável resposta y_{i1}, ou apenas y_{i2}, ou y_{i1} e y_{i2} conjuntamente.

4.5.2.3 Perturbação de uma covariável no logaritmo do tempo

Nesse caso, o objetivo é avaliar a sensibilidade do modelo a pequenas perturbações em uma determinada variável explicativa contínua, X_l. Considere uma perturbação aditiva para a variável explicativa, $x_{ilw} = x_{il} + S_{xl}w_i$, em que S_{xl} é um fator de escala que pode ser a estimativa do desvio padrão de X_l e $w_i \in \mathbf{R}$. O logaritmo da verossimilhança perturbada sujeita as restrições lineares é dado por:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n}\left\{\delta_{i1}\delta_{i2}f_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i^*) + \delta_{i1}(1-\delta_{i2})\left[\frac{-\partial S_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i^*)}{\partial y_{i1}}\right]\right\} + \\ &\sum_{i=1}^{n}\left\{(1-\delta_{i1})\delta_{i2}\left[\frac{-\partial S_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i^*)}{\partial y_{i2}}\right] + (1-\delta_{i1})(1-\delta_{i2})S_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i^*)\right\} + \\ &\vartheta\sum_{j=1}^{q}(\boldsymbol{u}_j^T\boldsymbol{\varphi} - c_j), \end{aligned}$$

em que $\boldsymbol{x}_i^{*T}\boldsymbol{\beta}_k = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \ldots + \beta_{lk}(x_{il} + S_{xl}w_i) + \ldots + \beta_{pk}x_{ip}$ e o vetor de não perturbação $\boldsymbol{w}_0 = (0,\ldots,0)^T$.

4.5.2.4 Perturbação de uma covariável na fração de cura

Para esse caso, o interesse é detectar a sensibilidade do modelo quando uma covariável contínua que esteja influenciando a fração de cura é submetida a uma perturbação aditiva $x_{ilw} = x_{il} + S_{xl}w_i$, em que S_{xl} é um fator de escala que pode ser a estimativa do desvio padrão de X_l e $w_i \in \mathbf{R}$. O logaritmo da verossimilhança perturbada sujeita as restrições lineares é dado por:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n}\left\{\delta_{i1}\delta_{i2}f^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i) + \delta_{i1}(1-\delta_{i2})\left[\frac{-\partial S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}}\right]\right\} + \\ &\sum_{i=1}^{n}\left\{(1-\delta_{i1})\delta_{i2}\left[\frac{-\partial S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}}\right] + (1-\delta_{i1})(1-\delta_{i2})S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i)\right\} + \\ &\vartheta\sum_{j=1}^{q}(\boldsymbol{u}_j^T\boldsymbol{\varphi} - c_j), \end{aligned}$$

em que $S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i) = \exp\{-[\theta_1^*F(y_{i1}|\boldsymbol{x}_i) + \theta_2^*F(y_{i2}|\boldsymbol{x}_i)]^\alpha\}$, $f^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i) = \frac{\partial^2 S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}\partial y_{i2}}$ e $\theta_k^* = \exp\{\gamma_{0k} + \gamma_{1k}x_{i1} + \ldots + \gamma_{lk}(x_{il} + S_{xl}w_i) + \ldots + \gamma_{pk}x_{ip}\}$. Vale observar que a covariável

pode influenciar a fração de cura existente no logaritmo do tempo de falha da variável resposta y_{i1}, ou da variável resposta y_{i2}, ou conjuntamente.

4.5.2.5 Perturbação de uma covariável na fração de cura e no logaritmo do tempo

Agora, deseja-se verificar a sensibilidade do modelo quando uma covariável contínua, X_l, influência tanto na fração de cura como no logaritmo do tempo de falha. A perturbação aditiva é dada por $x_{ilw} = x_{il} + S_{xl}w_i$, em que S_{xl} é um fator de escala que pode ser a estimativa do desvio padrão de X_l e $w_i \in \mathbf{R}$. O logaritmo da verossimilhança perturbada sujeita as restrições lineares é dado por:

$$\begin{aligned} l(\boldsymbol{\varphi}) &= \sum_{i=1}^{n}\left\{\delta_{i1}\delta_{i2}f^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}^*_i) + \delta_{i1}(1-\delta_{i2})\left[\frac{-\partial S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}^*_i)}{\partial y_{i1}}\right]\right\} + \\ &\sum_{i=1}^{n}\left\{(1-\delta_{i1})\delta_{i2}\left[\frac{-\partial S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}^*_i)}{\partial y_{i2}}\right] + (1-\delta_{i1})(1-\delta_{i2})S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}^*_i)\right\} + \\ &\vartheta\sum_{j=1}^{q}(\boldsymbol{u}_j^T\boldsymbol{\varphi} - c_j), \end{aligned}$$

em que $S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i) = \exp\{-[\theta^*_1F(y_{i1}|\boldsymbol{x}_i)+\theta^*_2F(y_{i2}|\boldsymbol{x}_i)]^\alpha\}$, $f^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i) = \frac{\partial^2 S^*_{pop}(y_{i1},y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}\partial y_{i2}}$, $\theta^*_k = \exp\{\gamma_{0k} + \gamma_{1k}x_{i1} + \ldots + \gamma_{lk}(x_{il} + S_{xl}w_i) + \ldots + \gamma_{pk}x_{ip}\}$ e $\boldsymbol{x}^{*T}_i\boldsymbol{\beta}_k = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \ldots + \beta_{lk}(x_{il} + S_{xl}w_i) + \ldots + \beta_{pk}x_{ip}$. Nesse caso, também pode-se considerar que a covariável pode influenciar a fração de cura existente no logaritmo do tempo de falha da variável resposta y_{i1}, ou da variável resposta y_{i2}, ou conjuntamente.

4.6 Aplicação

O conjunto de dados considerado nesta aplicação foi analisado por Huster et al. (1989), Liang et al. (1993), Wada e Hotta (2000) e Tarumoto (2001). Os dados foram coletados a partir de 1971 e são referntes a tempos até a perda de acuidade visual decorrente de retinopatia diabética. Pacientes com retinopatia diabética em ambos os olhos e com acuidade visual menor ou igual a 20/100 para ambos os olhos, fizeram parte do estudo. Um olho foi selecionado aleatoriamente para receber o tratamento de fotocoagulação a laser e o outro foi observado sem tratamento. Os pacientes foram observados por dois períodos de 4 meses completos e foi considerada falha a ocorrência de nível de acuidade visual menor que 5/200.

No total, 1742 pacientes foram acompanhados durante 7 anos, e no final 197 pacientes fizeram parte do subconjunto em estudo definido por critério de estudo de retinopatia diabética. Para cada paciente i, $i = 1, 2, \ldots, 197$ as variáveis associadas são:

- t_{i1}: tempo de queda da acuidade visual até o nível definido para o olho com tratamento (evento 1);
- t_{i2}: tempo de queda da acuidade visual até o nível definido para o olho sem tratamento (evento 2);
- δ_{i1}: indicador de censura do evento 1;
- δ_{i2}: indicador de censura do evento 2;
- x_{i1}: tipo de diabetes (0: diabetes juvenil, 1: diabetes adulto).

4.6.1 Análise Descritiva

Ao realizar uma análise exploratória dos dados considerando as variáveis respostas referentes aos tempos até a ocorrência do evento de interesse no olho com tratamento e no olho sem tratamento. Pelo fato das respostas serem medidas no mesmo indivíduo espera-se que essas sejam correlacionadas. O coeficiente de correlação de τ de Kendall foi calculado para as variáveis respostas obtendo $\tau = 0,39$. Esse resultado indica a existência de uma pequena associação positiva entre os tempos dos eventos em estudo. As estimativas de sobrevivência de Kaplan-Meier e o ajuste marginal do modelo com distribuição Weibull para ambos os eventos 1 e 2 são apresentados na Figura 21, verificando que é adequado supor distribuição Weibull para os tempos de sobrevivência, e, consequentemente, distribuição do valor extremo para o logaritmo dos tempos. Observa-se também a existência de uma significativa fração de indivíduos curados para ambos os tempos, pois o limite da estimativa de sobrevivência de t_1 tende a $0,67$, $\lim_{t_1 \to \infty} \hat{S}(t_1) = 0,67$, e de t_2 tende a $0,39$, $\lim_{t_1 \to \infty} \hat{S}(t_2) = 0,39$, indicando que os dados de retinopatia diabética devem ser analisados por modelos que consigam captar uma fração de indivíduos curados.

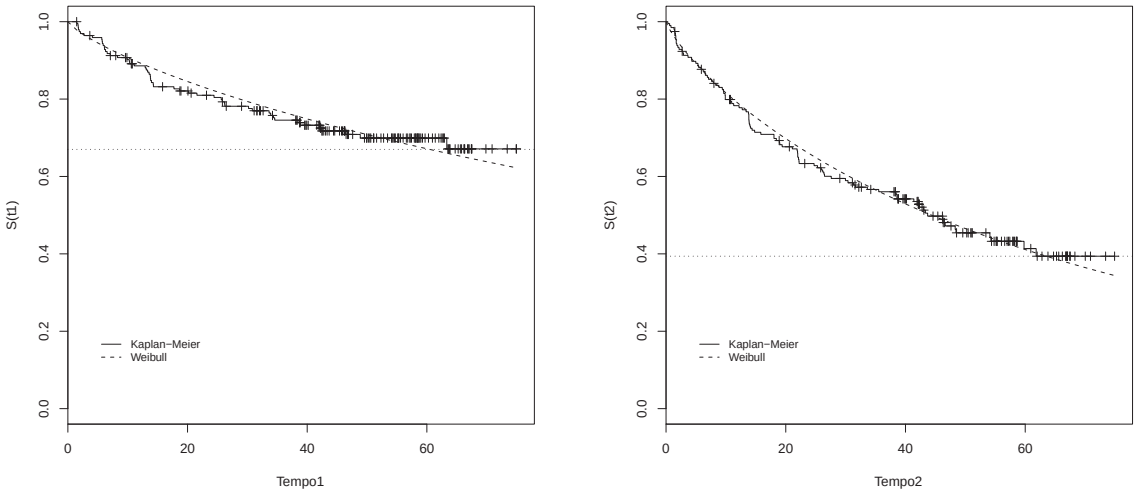

Figura 21 - Curvas de sobrevivência estimadas marginalmente por Kaplan-Meier para os dados de retinopatia

4.6.2 Ajuste do modelo de regressão log-linear bivariado com fração de cura

Ao constatar por meio da análise descritiva que as suposições do modelo de regressão log-linear bivariado com fração de cura estão satisfeitas. Nesta seção, dar-se-a início à análise dos dados de retinopatia diabética considerando o modelo proposto. Primeiramente, o modelo de regressão marginal de locação-escala, apresentado na seção 4.4, é dado por:

$$y_{ik} = \beta_{0k} + \beta_{1k} x_{i1} + \sigma_k z_{ik}, \tag{58}$$

em que $y_{ik} = \log(t_{ik})$ representa o logaritmo do tempo de falha, $i = 1, \ldots 197, k = 1, 2$, β_{pk} são os parâmetros de locação, σ_k são os parâmetros de escala, e z_{ik} são variáveis aleatórias independentes com função distribuição acumulada Valor extremo definida por:

$$F_k = F(z_{ik}) = 1 - \exp[-\exp(z_{ik})]. \tag{59}$$

Supondo que exista uma proporção de indivíduos curados em cada vetor $\boldsymbol{y}_k$, então essa proporção será modelada por: $\theta_k = \exp(\gamma_{0k} x_0 + \gamma_{1k} x_1)$. Desta forma, a função de

sobrevivência populacional é dada por:

$$S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) = \exp\left\{ - \left[\theta_1 \left[1 - \exp\left(- \exp\left(\frac{y_{i1} - \beta_{01} - \beta_{11}x_{i1}}{\sigma_1 z_{i1}}\right)\right)\right] + \right.\right.$$
$$\left.\left. \theta_2 \left[1 - \exp\left(- \exp\left(\frac{y_{i2} - \beta_{02} - \beta_{12}x_{i1}}{\sigma_2 z_{i1}}\right)\right)\right]\right]^{\alpha}\right\}, \qquad (60)$$

com probabilidade de cura conjunta $\exp[-(\theta_1 + \theta_2)^{\alpha}]$. As funções de sobrevivência marginais de (60) são

$$S_k(y_{ik}|\boldsymbol{x}_i) = \exp\left\{ - [\exp(\gamma_{0k}x0 + \gamma_{1k}x1)]^{\alpha} \left[1 - \exp\left(- \exp\left(\frac{y_{ik} - \beta_{01} - \beta_{1k}x_{i1}}{\sigma_k z_{i1}}\right)\right)\right]^{\alpha}\right\},$$

com probabilidade de cura $\exp\{-[\exp(\gamma_{0k}x_0 + \gamma_{1k}x_1)]^{\alpha}\}$ associada ao logarimo do tempo de falha do k-ésimo evento de interesse y_{ik}, para $k = 1, 2$.

Para estimar os parâmetros de interesse do modelo proposto, foi utilizado o método de máxima verossimilhança sujeita as restrições nos parâmetros definido na seção 4.4.1. O logaritmo da função de verossimilhança expresso em (55) para os dados de retinopatia diabética é composto por:

$$-\frac{\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i1}} = \frac{(\theta_1 F_1 + \theta_2 F_2)^{\alpha-1}\alpha\theta_1 e^{z_{i1}} e^{-e^{z_{i1}}} S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\sigma_1},$$

$$-\frac{\partial S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\partial y_{i2}} = \frac{(\theta_1 F_1 + \theta_2 F_2)^{\alpha-1}\alpha\theta_2 e^{z_{i2}} e^{-e^{z_{i2}}} S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\sigma_2},$$

$$f_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i) = \frac{(\theta_1 F_1 + \theta_2 F_2)^{\alpha}\alpha\theta_1 e^{z_{i1}} e^{-e^{z_{i1}}} \theta_2 e^{z_{i1}} e^{-e^{z_{i1}}} S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)}{\sigma_2(\theta_1 F_1 + \theta_2 F_2)^2\sigma_1} \times$$
$$[\alpha + 1 + (\theta_1 F_1 + \theta_2 F_2)^{\alpha}\alpha],$$

em que $S_{pop}(y_{i1}, y_{i2}|\boldsymbol{x}_i)$ é definida em (60) e F_k são as funções distribuição acumulada do valor extremo definidas em (59).

As estimativas dos parâmetros do modelo de regressão log-linear bivariado com fração de cura, os respectivos erros padrões e significâncias encontram-se na Tabela 5. Os resultados indicam que apenas o intercepto é significativo tanto para o logaritmo dos tempos de vida, como para a fração de cura da variável resposta 1. A estimativa do parâmetro de associação indica uma pequena associação entre os logaritmos dos tempos de vida, o que confirma o resultado obtido na análise descritiva realizada na seção 4.6.1.

A proporção de cura conjunta média estimada para ambos os logaritmos dos tempos de vida é de $16,85\%$, a proporção de cura média estimada para o logaritmo do tempo de vida da variável resposta 1 é de $64,25\%$, e a proporção de cura média estimada para o logaritmo do tempo de vida da variável resposta 2 é de $22,53\%$.

Tabela 5 - Estimativa de máxima verossimilhança para o modelo de regressão bivariado com fração de cura na estrutura de riscos competitivos

Parâmetro	Estimativa	Erro-Padrão	valor-p
β_{01}	3,4137	0,3108	0,0000
β_{11}	0,5508	0,7209	0,4449
β_{02}	4,1323	0,7150	0,0000
β_{12}	0,8830	1,8947	0,6412
σ_1	0,7717	0,1093	-
σ_2	0,8892	0,1112	-
α	0,8043	0,0589	-
γ_{01}	-0,9334	0,2704	0,0006
γ_{11}	-0,1968	0,6676	0,7681
γ_{02}	0,0705	0,5542	0,8988
γ_{12}	1,2307	1,8700	0,5104

Dos resultados obtidos pelo modelo proposto e das probabilidades calculadas empiricamente pelas curvas de Kaplan-Meier apresentadas na Figura 21, conclui-se que existe uma proporção de invidíduos curados da perda da acuidade visual para o olho com tratamento, para o olho sem tratamento, e conjuntamente. Pode-se observar que a covariável tipo de diabetes não se mostrou significativa em relação ao tempo até a perda da acuidade visual e nem na proporção de indivíduos curados, tanto para os olhos tratados como para os não tratados. O tratamento aplicado a um dos olhos, entretanto, mostrou-se eficiente devido a uma maior proporção de curados observada para os olhos tratados. Devido a semelhança entre os interceptos de regressão para ambos os tempos pode-se concluir que a efetividade do tratamento ocorre apenas na capacidade de curar e não de retardar a perda de acuidade visual.

4.6.3 Análise de Influência Global

Com o objetivo de verificar a existência de possíveis observações influenciando o ajuste do modelo de regressão log-linear bivariado com fração de cura para os dados de retinopatia diabética, foram calculadas as medidas de Influência Global, Afastamento da Verossimilhança ($LD_i(\boldsymbol{\varphi})$) e Distância de Cook Generalizada ($GD_i(\boldsymbol{\varphi})$), como definidas na seção 4.5.1. Dessa análise nota-se que as observações #5 e #102, se destacam das demais, como mostra

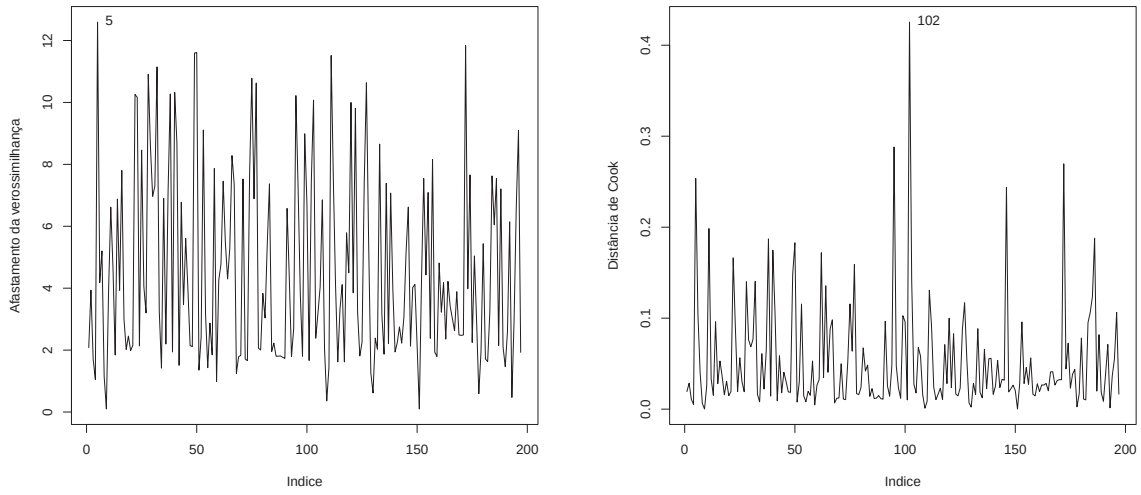

Figura 22 - Gráficos de medidas de Influência Global para modelo de regressão log-linear bivariado com fração de cura para os dados de retinopatia diabética. (a) Afastamento da Verossimilhança (b) Distância de Cook

a Figura 22. A observação #5 refere-se ao indivíduo com tempo de censura médio para a variável resposta 1 e o menor tempo de falha para a variável resposta 2. A observação #102 refere-se ao indivíduo com um grande tempo de falha para o olho com tratamento e com tempo de falha médio para o olho sem tratamento. Essas diferentes características de cada indivíduo que o torna detectável na análise de Influência Global.

4.6.4 Análise de Influência Local

Para dar continuidade à análise de sensibilidade aos dados de retinopatia diabética considerando o modelo proposto, foi realizada uma análise de Influência Local incluindo os esquemas de perturbação de casos, perturbação nos logaritmos dos tempos até a perda da acuidade visual para o olho com tratamento, perturbação nos logaritmos dos tempos até a perda da acuidade visual para o olho sem tratamento, e a perturbação conjunta para os logarimtos dos tempos até a perda da acuidade visual para os olhos com tratamento e sem tratamento. Os esquemas de perturbação envolvendo covariáveis não foram considerados, pois a única covariável presente no estudo é categórica.

Para os esquema de perturbação citados, foram claculados os vetores $\boldsymbol{d}_{max}$, $\boldsymbol{d}_{max}(y_1)$, $\boldsymbol{d}_{max}(y_2)$ e $\boldsymbol{d}_{max}(y_1y_2)$ correspondente à direção da maior curvatura, e os autovalores das curvaturas máximas que são dados por: $C_{\boldsymbol{d}_{max}}(\boldsymbol{\varphi}) = 1,27$, $C_{\boldsymbol{d}_{max}}(y_1)$ =5,84, $C_{\boldsymbol{d}_{max}}(y_2) = 6,98$ e $C_{\boldsymbol{d}_{max}}(y_1y_2) = 7,08$, respectivamente. Os gráficos para as medidas de

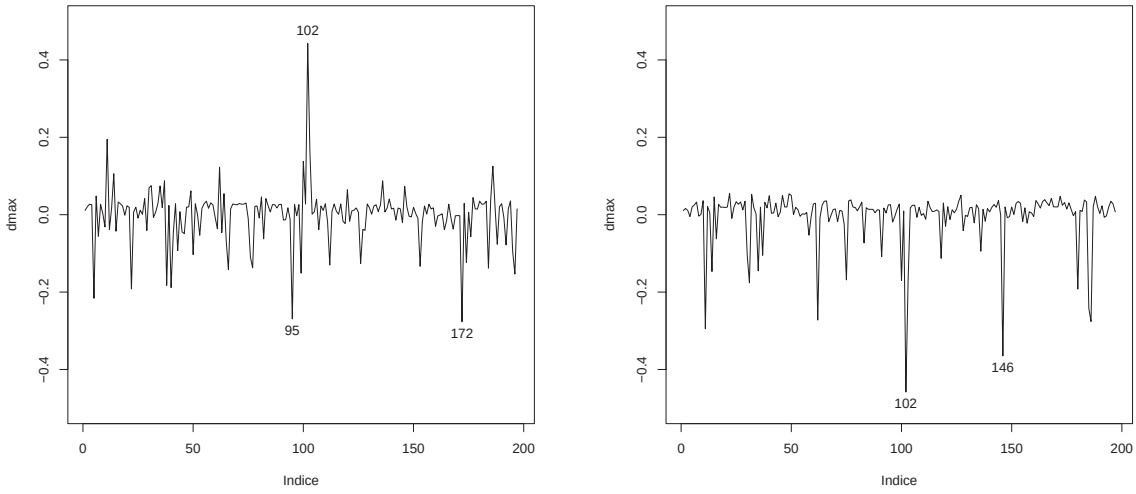

Figura 23 - Gráfico de medidas de influência do modelo de regressão log-linear bivariado com fração de cura considerando o esquema de perturbação de casos e da variável resposta 1 para os dados de retinopatia diabética. (a) Influência Local $\boldsymbol{d}_{max}$ casos (b) Influência Local $\boldsymbol{d}_{max}(y_1)$

Influência Local $\boldsymbol{d}_{max}$ $\boldsymbol{d}_{max}(y_1)$, $\boldsymbol{d}_{max}(y_2)$ e $\boldsymbol{d}_{max}(y_1y_2)$, e para as medidas de Influência Local Total, C_i, $C_{y_{i1}}$, $C_{y_{i2}}$ e $C_{y_{i1}y_{i2}}$ contra o índice das observações, encontram-se nas Figuras 23, 24, 25 e 26. Por meio dessas figuras, verifica-se que as observações #95 e #102 requerem uma atenção especial por se destacarem das demais. Isso ocorre porque a observação #95 refere-se ao indivíduo com pequeno tempo de falha até a perda da acuidade visual para o olho com tratamento e um grande tempo de censura até a perda da acuidade visual para o olho sem tratamento. E a observação #102 representa o indivíduo com um grande tempo de falha para o olho com tratamento e com tempo de falha médio para o olho sem tratamento.

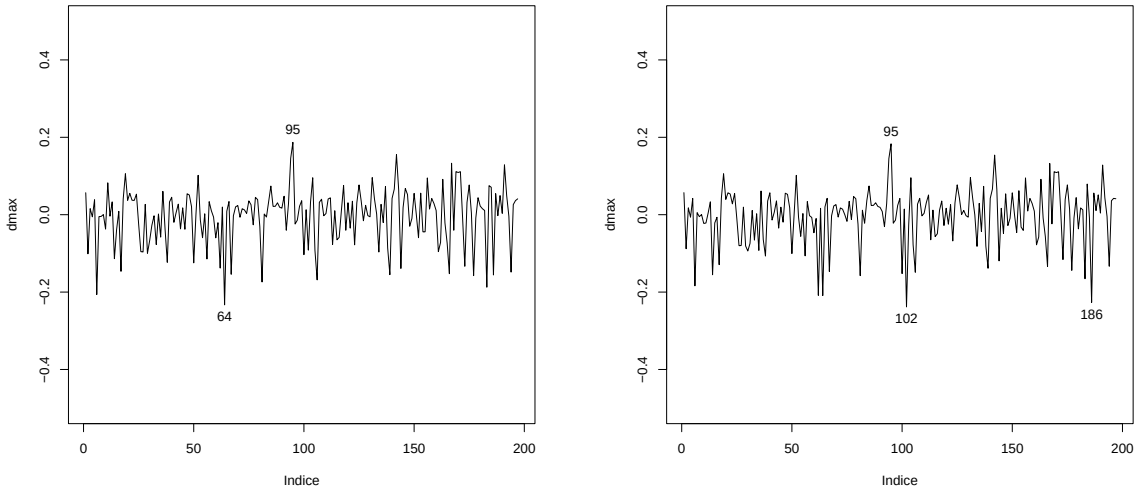

Figura 24 - Gráfico de medidas de influência do modelo de regressão log-linear bivariado com fração de cura considerando o esquema de perturbação da variável resposta 2 e a perturbação conjunta de ambas as variáveis respostas para os dados de retinopatia diabética. (a) Influência Local $\boldsymbol{d}_{max}(y_2)$ (b) Influência Local $\boldsymbol{d}_{max}(y_1y_2)$

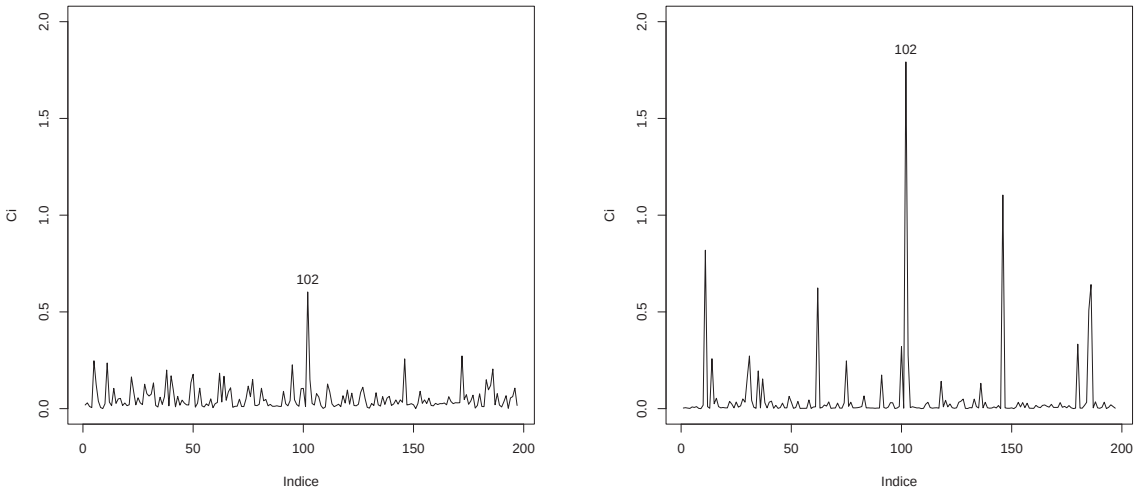

Figura 25 - Gráfico de medidas de influência do modelo de regressão log-linear bivariado com fração de cura considerando o esquema de perturbação de casos e da variável resposta 1 para os dados de retinopatia diabética. (a) Influência Local Total C_i casos (b) Influência Local Total $C_{y_{i1}}$

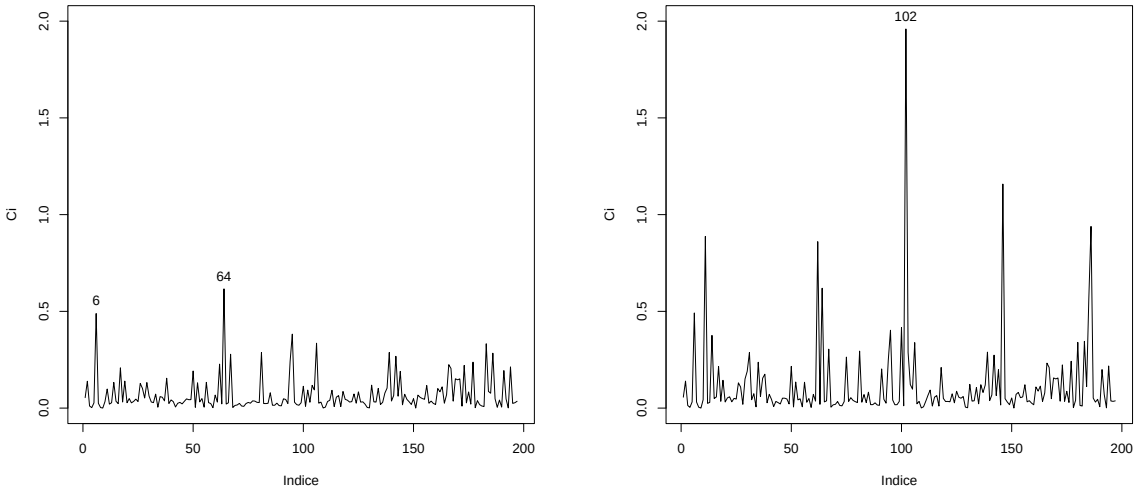

Figura 26 - Gráfico de medidas de influência do modelo de regressão log-linear bivariado com fração de cura considerando o esquema de perturbação da variável resposta 2 e a perturbação conjunta de ambas as variáveis respostas para os dados de retinopatia diabética. (a) Influência Local Total $C_{y_{i2}}$ (b) Influência Local Total $C_{y_{i1}y_{i2}}$

4.6.5 Impacto das observações influentes

Ao considerar os resultados obtidos na análise de sensibilidade para os dados de retinopatia considerando o modelo de regressão log-linear bivariado com fração de cura, verifica-se que as observações #95 e #102 foram as que mais se destacaram das demais, sendo a observação #95 referente ao indivíduo com pequeno tempo de falha até a perda da acuidade visual para o olho com tratamento e um grande tempo de censura até a perda da acuidade visual para o olho sem tratamento, e a observação #102 referente ao indivíduo com um grande tempo de falha para o olho com tratamento e com tempo de falha médio para o olho sem tratamento. Após realizar algumas verificações, constata-se que essas observações não apresentam indícios de erro na coleta ou transcrição dos dados, dessa forma devem ser mantidas no conjunto de observações.

O impacto dessas observações sob o modelo deve ser analisado para avaliar a sensibilidade do modelo e suas estimativas. Para realizar tal análise, novas estimativas para os parâmetros do modelo foram obtidas a partir de subamostras referentes à retirada dessas observações, individualmente e em grupo.

As subamotras obtidas a partir da exclusão individual e conjunta dos valores possivelmente influentes, as estimativas para os parâmetros, os respectivos p-valores e as mudanças relativas de cada parâmetro encontram-se na Tabela 6. A mudança relativa foi definida como $\boldsymbol{RC}_{\boldsymbol{\varphi}_j} = [(\hat{\boldsymbol{\varphi}}_j - \hat{\boldsymbol{\varphi}}_{j(I)})/\hat{\boldsymbol{\varphi}}_j]$, sendo (I) o índice referente as observações excluidas da amostra.

Ao analisar os resultados da Tabela 6, verifica-se que com a inclusão ou exclusão das observações em estudo não há mudanças nas significancias dos parâmetros do modelo, fato que torna o modelo de regressão log-linear bivariado com fração de cura robusto nesta aplicação.

Tabela 6 - Mudança relativa [RC], estimativas dos parâmetros, e correspondentes (p-valor)

Sub-amostra	$I-\{completo\}$	$I-\{95\}$	$I-\{102\}$	$I-\{95,102\}$
	[-]	[-0,1897]	[1,2701]	[1,0361]
β_{01}	3,4137	3,4202	3,3703	3,3783
	(0,0000)	(0,0000)	(0,0000)	(0,0000)
	[-]	[-31,5434]	[77,0646]	[60,6910]
β_{11}	0,5508	0,7245	0,1263	0,2165
	(0,4449)	(0,4231)	(0,7801)	(0,6654)
	[-]	[-0,3417]	[-0,4869]	[-0,9246]
β_{02}	4,1323	4,1464	4,1524	4,1705
	(0,0000)	(0,0000)	(0,0000)	(0,0000)
	[-]	[-55,1053]	[21.3154]	[-15,8765]
β_{12}	0,8830	1,3696	0,6948	1,0232
	(0,6412)	(0,6946)	(0,6578)	(0,6609)
	[-]	[0,8140]	[4,1360]	[4,7177]
σ_1	0,7717	0,7654	0,7397	0,7353
	(-)	(-)	(-)	(-)
	[-]	[0,7397]	[-0,4799]	[0,2961]
σ_2	0,8892	0,8826	0,8935	0,8866
	(-)	(-)	(-)	(-)
	[-]	[1,7239]	[-0,2146]	[1,4827]
α	0,8043	0,7904	0,8060	0,7924
	(-)	(-)	(-)	(-)
	[-]	[-1,6452]	[-1,9893]	[-3,5556]
γ_{01}	-0,9334	-0,9499	-0,9519	-0,9665
	(0,0006)	(0,0006)	(0,0002)	(0,0002)
	[-]	[40,5894]	[-179,5069]	[-191,6330]
γ_{11}	-0,1968	-0,1169	-0,5501	-0,5740
	(0,7681)	(0,8939)	(0,2114)	(0,2367)
	[-]	[-4,7466]	[-18,1376]	[-28,4416]
γ_{02}	0,0705	0,0738	0,0832	0,0905
	(0,8988)	(0,8970)	(0,8837)	(0,8780)
	[-]	[-43,9936]	[15,3464]	[-14,3147]
γ_{12}	1,2307	1,7722	1,0419	1,4069
	(0,5104)	(0,6280)	(0,4884)	(0,5519)

4.6.6 Qualidade de Ajuste

Com o objetivo de verificar a qualidade do ajuste do modelo de regressão log-linear bivariado com fração de cura aos dados de retinopatia diabética, forma construidos gráficos representando as funções de sobrevivência estimadas pelo método de Kaplan-Meier e as funções de sobrevivência marginais estimadas para o modelo de regressão log-linear bivariado com fração de cura. A Figura 27 ilustra as funções citadas sem considerar covariáveis e a Figura 28 ilustra a incorporação da covariável tipo de diabetes nas funções citadas.

Ao analisar as Figuras 27 e 28, verifica-se que o modelo conseguiu captar muito bem a proporção de indivíduos curados presente no evento 1, contudo para o evento 2 há um indicativo de que o modelo sub estimou a proporção de indivíduos imunes presente nesse evento. Mas de forma geral, verifica-se um bom ajuste do modelo, pois observa-se que a curva do modelo de regressão log-linear bivariado com fração de cura acompanha o gráfico da função de sobrevivência estimada por Kaplan-Meier.

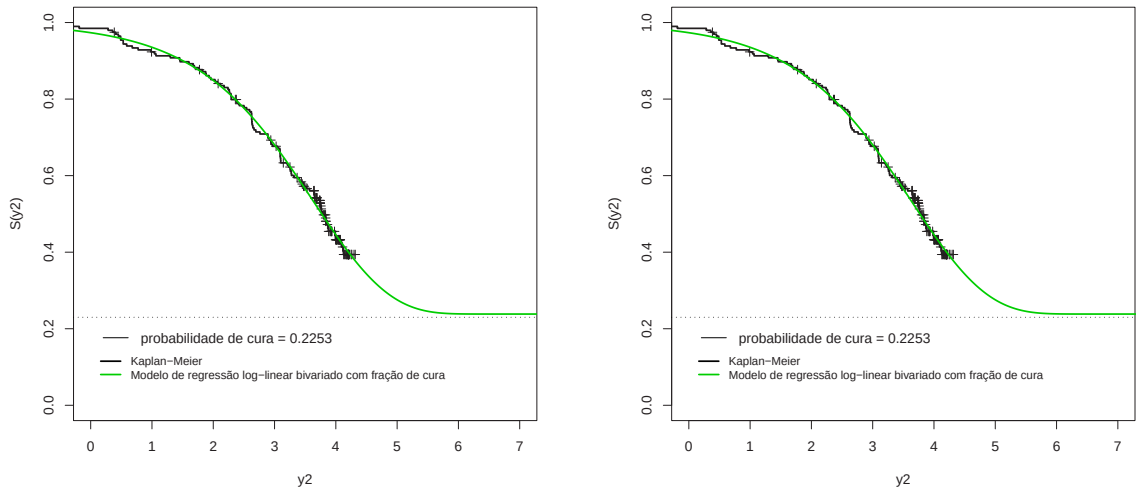

Figura 27 - Curvas de sobrevivência de Kaplan-Meier e função de sobrevivência estimada para os dados de retinopatia. (a) função de sobrevivência marginal $S_1(y_1)$, e (b) função de sobrevivência marginal $S_2(y_2)$

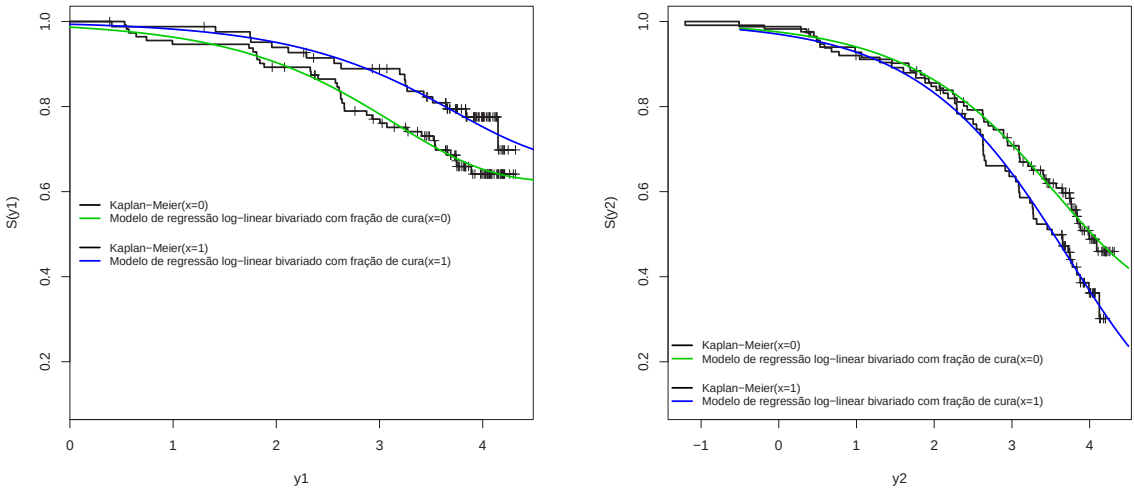

Figura 28 - Curvas de sobrevivência de Kaplan-Meier estratificadas por tipo de diabetes (0: diabetes juvenil, 1: diabetes adulto) e função de sobrevivência estimada para os dados de retinopatia. (a) função de sobrevivência marginal $S_1(y_1|\boldsymbol{x})$, e (b) função de sobrevivência marginal $S_2(y_2|\boldsymbol{x})$

4.7 Conclusões

Neste capítulo foi proposto um modelo de regressão log-linear bivariado com fração de cura baseado no trabalho de Chen et al. (2002), para dados bivariados com a presença de uma significativa proporção de indivíduos curados e com a presença de covariáveis. Os parâmetros do modelo foram estimados utilizando uma metodologia de máxima verossimilhança sujeita a restrições lineares nos parâmetros, e uma análise de sensibilidade foi conduzida para avaliar a robustez dos resultados obtidos.

A metodologia de estimação foi implementada no *software* R (R DEVELOPMENT CORE TEAM, 2009), utilizando diversas funções já existentes no mesmo, assim como novas funções com os cálculos das derivadas parciais da função de verossimilhança foram incluidas na programação utilizada. No processo de otimização, os valores iniciais para os parâmetros de regressão foram obtidos de estimativas de modelos marginais supondo distribuição do Valor extremo para cada variável resposta. Para o

parâmetro α foi utilizado como referência a interpretação da associação indicada pelo valor calculado para o coeficiente τ - Kendall. O modelo mostrou-se computacionalmente flexível, desde que as restrições dos parâmetros sejam corretamente consideradas. É importante ressaltar que esse modelo só deve ser considerado quando de fato existe um indicativo da presença de indivíduos curados.

Ao considerar as técnicas de análise de sensibilidade, Influência Global, Influência Local e Influência Local Total algumas observações foram identificadas como possíveis pontos influentes. Após algumas verificações no conjunto de dados descartou-se a possibilidade de erros de trasncrição e administração dos dados. Um análise para verificar a sensibilidade do modelo na presença ou ausência dessas observações também foi realizada constatando que o modelo proposto neste capítulo é robusto a pequenas mudanças nos dados.

Para os dados de retinopatia diabética analisado, o modelo conseguiu captar uma pequena associação presente entre as variáveis respostas, a proporção de indivíduos curados presente no evento 1 e sub estimou a presença de indivíduos curados presente no evento 2 comparado com os resultados obtidos na análise exploratória dos dados, além de estimar a presença de indivíduos curados conjuntamente. Fato esse que só um modelo de fração de cura consegue diagnosticar.

Com base nos resultados obtidos e discutidos sobre o modelo de regressão log-linear bivariado com fração de cura, conclui-se que este modelo pode ser usado como uma nova ferramenta para analisar dados de sobrevivência com resposta bivariada e presença de uma proporção de indivíduos curados, e ainda consta com uma metodologia de análise de sensibilidade desenvolvida para verificar a adequação das suposições do modelo e validar seus resultados.

4.7.1 Proposta para trabalhos futuros

Como possíveis trabalhos futuros podem-se considerar os seguintes temas de pesquisa:

1. Construir uma nova formulação da função de sobrevivência populacional bivariada ao considerar outras distribuições para o número de células carcinogênicas.

2. Supor outras distribuições para induzir a correlação entre as variáveis latentes (número de células carcinogênicas).

Referências

ASSELAIN, B.; FOURQUET, A.; HOANG, T.; TSODIKOV, D.; YAKOVLEV, A. Y. A parametric regression model of tumor recurrence: An application to the analysis of clinical data on breast cancer. **Statistics and Probability Letters**, North-Holland, v. 29, p. 271-278, 1996.

BERKSON, J.; GAGE, R.P. Survival curve for cancer patients follwing treatment. **Journal of the American Statistical Association**, Alexandria, v. 47, p. 501-515, 1952.

CANCHO, V.G. **Métodos de monte carlo em análise de sobrevivência.** 1999. 207p. Tese (Doutorado em Estatística)-Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 1999.

CANCHO, V.G.; ORTEGA, E. M. M.; BOLFARINE, H. The log-exponentiated-Weibull regression models with cure rate: local influence and residual analysis. **Journal of Data Science**, New York, v. 7, p. 433-458, 2009.

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2nd ed. Pacific Grove: Thomson Learning, 2002. 660 p.

CASTRO, M.; CANCHO, V. G.; RODRIGUES, J. A bayesian long-term survival model parametrized in the cured fraction. **Biometrical Journal**, Weinheim, v. 51, n. 3, p. 443-455, 2009.

CHEN, M. H.; IBRAHIM, J; SINHA,D. A new bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Alexandria, v. 94, p. 909-919, 1999.

CHEN, M. H.; IBRAHIM, J; SINHA,D. Bayesian inference for multivariate survival data with a cure fraction. **Journal of Multivariate Analysis**, New York, v. 80, p. 101-126, 2002.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Edgard Blücher, 2006. 392 p.

COOK, R.D. Detection of influential observations in linear regression. **Technometrics**, Alexandria, v. 19, p. 15-18, 1977.

COOK, R. D.; WEISBERG, S. **Residuals and influence in regression**. New York: Chapman and Hill, 1982. 230 p.

COOK, R.D. Assement of local influence (with discussion). **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 48, n. 2, p. 133-169, 1986.

COOK, R. D.; PEÑA, D.; WEISBERG, S. The likelihood displacement: a unifying principle for influence. **Communications in Statistics: Part Theory and Methods**, New York, v. 17, n. 3, p. 623-640, 1988.

COX, D.R.; OAKES, D. **Analysis of survival data**. London: Chapman and Hall, 1984. 201 p.

ESCOBAR, L.A.; MEEKER, W.Q. Assessing influence in regression analysis with censored data. **Biometrics**, Washington, v. 48, n.2, p. 507-528, 1992.

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.

FAREWELL, V. T. The use mixture models for the analysis of survival data with log-term survivors. **Biometrics**, Washington, v. 38, p. 43-46, 1982.

FAREWELL, V. T. Mixture models in survival analysis: Are they worth the risk? **Canadian Journal Statistical**, Toronto, v. 14, p. 257-262, 1986.

GOLDMAN, A. Survivorship analysis when cure is a possibility: a Monte Carlo study. **Statistics in Medicine**, Chichester, v. 3, p. 153-163, 1984.

GOMES, E. M. C. **Análise de Sensibilidade e resíduos em modelos de regressão com respostas bivariadas por meio de cópulas.** 2007. 103p. Dissertação (Mestre em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

GOURIEROUX, C.; MONFORT, A. **Statistical and econometric models**. Cambridge: Cambridge University Press, 1995. v.2, 526 p.

GREENHOUSE, J.B.; WOLFE, R.A. A competing risks derivation of a mixture model for the analysis of survival data. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 13, p. 3133-3154, 1984.

GU, H; FUNG, W.K. Local influence for the restricted likelihood with applications. **Sankhya: The Indian Journal of Statistical**, Indian, v. 63, pt. 2, p. 250-259, 2001.

HALPERN, J.; BROWN, B. Cure rate models: Power of the log-rank and generalized Wilcoxon tests. **Statistics in Medicine**, Chichester, v. 6, p. 483-489, 1987.

HASHIMOTO, E. M. **Modelo de Regressão para dados com censura intervalar e dados de sobrevivência agrupados** 2008. 121p. Dissertação (Mestre em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

HE, W.; LAWLESS, J. F. Bivariate location-scale models for regression analysis, with applications to lifetime data. **Journal of the Royal Statistical Society**, London, v. 67, n. 1, p. 63-78, 2005.

HUSTER, W. J.; BROOKMEYER, R.; SELF, S.G. Modelling Paired Survival Data with Covariates **Biometrics**, Washington, v. 45, p. 145-156, 1989.

IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. **Bayesian survival analysis**. New York: Springer-Verlag, 2001, 479p.

KALBFLEISCH, J.D.; PRENTICE, R.L. **The statistical analysis of failure time data**. 2nd ed. New York: John Wiley, 2002. 439p.

KLEIN, J.P; MOESCHBERGER, M.L. **Survival analysis techniques for censored and truncated data**. 1nd ed. New York: Springer-Verlag, 1997. 357p.

KWAN, C. W; FUNG, W. K. Assessing local influence for specific restricted likelihood: application to factor analysis. **Psychometrika**, New York, v. 63, n. 1, p. 35-46, 1998.

LANGE, K. **Numerical analysis for statisticians**. New York: Springer, 1999. 356 p.

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

LEE. E. T. **Statistical models and for survival data analysis**, 2nd ed., New York: Wiley, 1992. 482 p.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, Washington, v. 54, n. 2, p. 570-582, 1998.

LIANG, K.; STEVEN, G. S.; CHANG, Y. Modelling Marginal Hazards in Multivariate Failure Time Data. **Journal of the Royal Statistical Society: Series B, Statistical Methodology**, Oxford, v. 55, p. 441-453, 1993.

MALLER, R.; ZHOU, X. **Survival Analysis with Long-Term Survivors**. New York: Wiley, 1996. 278 p.

MIZOI, M. F. **Influência local em modelos de sobrevivência com fração de cura.** 2004. 95p. Tese (Doutorado em Estatística)-Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2004.

MOESCHBERGER, M. L. Life tests under dependent competing causes of failure. **Technometrics**, Alexandria, v. 16, p. 39-47, 1974.

ORTEGA, E. M. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. **Computational Statistics and Data Analysis**, New York, 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 Operation Reserch Transactions**, Catalunya, v. 30, n. 2, p. 171-192, 2006.

ORTEGA, E. M. M. ; PAULA, G. A. ; BOLFARINE, H. Deviance residuals in generalized log-gamma regression models with censored observations. **Journal of Statistical Computation and Simulation**, New York, v. 78, p. 747-764, 2008.

ORTEGA, E. M. M. ; CANCHO, V. G.; PAULA, G. A. Generalized log-gamma regression models with cure fraction. **Lifetime Data Analysis**, Boston, v. 15, p. 79-106, 2009a.

ORTEGA, E. M. M. ; RIZZATO, F. B.; DEMÉTRIO, C. G. B. The generalized log-gamma mixture model with covariates: local influence and residual analysis. **Statistical Methods and Application**, New York, v. 18, n. 3, p. 305-331, 2009b.

ORTEGA, E. M. M. ; CANCHO, V. G.; LANCHOS, V. H. A generalized log-gamma mixture models for cure rate: estimation and sensitivity analysis. **Sankhya: The Indian Journal of Statistical**, Indian, v. 71, p. 1-29, 2009c.

PAULA, G.; CYSNEIROS, F. J. A. Local influence under parameter constraints. **Communications in Statistics: Theory and Methods**, New York, v.88, p. 1-23, 2009.

R Development Core Team (2009). **R**: A language and environment for statistical computing. Disponível em:¡http://www.R-project.org¿. Acesso em: 17 maio 2011.

RIZZATO, F.B. **Modelos de Regressão log-gama generalizado com fração de cura.** 2007. 74p. Dissertação (Mestrado em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

TARUMOTO, M. H. **Um modelo Weibull bivariado para riscos competitivos.** 2001. 154p. Tese (Doutorado em Matemática Aplicada)- Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 2001.

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**, Alexandria, v. 98, p. 1063-1078, 2003.

WADA, C. Y.; HOTTA, L. K. Restricted alternatives tests in a bivariate exponential model with covariates. **Communications in Statistics - Theory and Methods**, Philadelphia, v. 29, p. 193-210, 2000.

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.; ASSELAIN, B.; BARDOU, V., FOURQUET, A. HOANG, T. ROCHEFEDIERE; TSODIKOV,A.D. A Stochastic models of tumor latency and their biosta-tistical applications. **Biometrie et analyse de Donnes Spatio-Temporelles**, Paris, v. 12, p. 66-82, 1993.

YAKOVLEV, A. Letter to the Editor. **Statistics in Medicine**, Chichester, v. 13, p. 983-986, 1994.

YAKOVLEV, A.; TSODIKOV,A.D. **Stochastic models of tumor latency and their biostatistical applications**, New Jersey: World Scientific, 1996. 288 p.

ZHU, H.; ZHANG, H. A diagnostic procedure based on local influence. **Biometrika**, Cambridge, v. 91, n. 3, p. 579-589, 2004.