UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Decomposição da variância para o modelo de regressão destrutivo Waring de longa duração

Jonathan Kevin Jordan Vasquez

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Jonathan Kevin Jordan Vasquez

Decomposição da variância para o modelo de regressão destrutivo Waring de longa duração

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *EXEMPLAR DE DEFESA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Josemar Rodrígues

USP – São Carlos
Setembro de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

J82d

Jordan Vasquez, Jonathan Kevin
Decomposição da variância para o modelo de regressão destrutivo Waring de longa duração / Jonathan Kevin Jordan Vasquez; orientador Josemar Rodrigues. -- São Carlos, .
111 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, .

1. Teoria de acidentes. 2. Aleatorização. 3. Mecanismo destrutivo. 4. Imunoterapia. 5. Distribuição Waring Generalizada. I. Rodrigues, Josemar , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Jonathan Kevin Jordan Vasquez

An useful variance decomposition for destructive Waring regression long-term model

Dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Statistics

Advisor: Prof. Dr. Josemar Rodrígues

USP – São Carlos
September 2019

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Jonathan Kevin Jordán Vásquez, realizada em 17/04/2020:

Prof. Dr. Josemar Rodrigues
USP

Profa. Dra. Elizabeth Mie Hashimoto
UTFPR

Prof. Dr. Vinicius Fernando Calsavara
CIPE

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Josemar Rodrigues Elizabeth Mie Hashimoto, Vinicius Fernando Calsavara e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Este trabalho é dedicado primeiramente a Deus,
à minha família e ao meu orientador por ter me dado
o apoio necessário para que eu chegasse até aqui.

AGRADECIMENTOS

Agradeço, em primeiro lugar a Deus por sua divina providência deu-me inspiração, sabedoria e força para todos os momentos de minha vida, e agora em minha dissertação de mestrado.

Aos meus pais Edilberto Jordan e Bacilia Vasquez, verdadeiramente os maiores mestres da minha vida, embora distantes fisicamente sempre apoiaram e confiaram em mim, possibilitando-me a concretização de mais um grande passo em direção à minha realização profissional, e aos meus irmãos Antony e Diego que sempre incentivaram-me a continuar.

Ao meu orientador, Prof. Dr. Josemar Rodrigues, pela compreensão, sábios conselhos e direcionamentos, pela disponibilidade para ajudar-me, pelas correções e sugestões que foram essenciais para realização deste trabalho e pela amizade. Ao Prof. Dr. Vicente Garibay Cancho que gentilmente proporcionou os códigos para nosso estudo. Aos membros da banca Profa. Dra. Elizabeth Mie Hashimoto e Prof. Dr. Vinícius Fernando Calsavara, pelas sugestões e correções, que foram úteis para melhorar a redação da dissertação. Aos professores e funcionários do programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs) de São Carlos que propiciaram-me condições para a realização deste trabalho.

À Katy Rocio C. Molina, que juntos começamos esta nova etapa do mestrado, e agradecer também pelo apoio, força, paciência e sentido de humor dado ao longo deste tempo.

À todos os meus amigos, que direta ou indiretamente colaboraram com o sucesso deste trabalho e que sempre estiveram torcendo por mim.

Finalmente o meu agradecimento para a Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES) - Código de Financiamento 001 pelo auxílio financeiro concedido, que permitiu-me dedicar inteiramente a este trabalho durante o mestrado.

“A vida é uma viagem
a três estações: ação, experiência e recordação.”
(Júlio Camargo)

RESUMO

VASQUEZ, J. J. K. **Decomposição da variância para o modelo de regressão destrutivo Waring de longa duração**. 2019. 111 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A finalidade deste trabalho é formular um modelo de regressão de longa duração em dois estágios, onde o mecanismo destrutivo dos fatores de riscos responsáveis pela sobrevivência do paciente está relacionado com três fontes de variabilidades: aleatória, externa e interna. O número de fatores de riscos é um efeito aleatório latente, que expressa o comportamento heterogêneo dos pacientes em relação ao risco básico da população, conhecido na literatura como fragilidade discreta. Esta fragilidade está diretamente conectada ao fenômeno de superdispersão e o mecanismo destrutivo. Várias distribuições discretas com caudas pesadas ("J-shaped") têm sido utilizadas para explicar o excesso de variabilidade, entretanto sem sucesso para separar a fragilidade interna, que corresponde ao mecanismo destrutivo, da fragilidade externa que corresponde a covariadas desconhecidas e não incluídas no modelo. A distribuição Binomial Negativa (BN) é a mais utilizada, porém não é flexível o suficiente para permitir uma destruição interna sem os ruídos externos. Neste contexto, a distribuição Waring é uma alternativa mais realista para o modelo de longa duração devido a existência de um mecanismo destrutivo individual e flexível. Consequentemente, a taxa de cura e o mecanismo destrutivo são personalizados e úteis no tratamento de câncer por imunoterapia, onde o paciente é o protagonista do tratamento. Um estudo de simulação Monte Carlo e aplicações com dados HIV e melanoma serão apresentados. A distribuição Waring é utilizada com sucesso na teoria de acidentes, onde os principais paradigmas serão adaptados na análise de sobrevivência de longa duração.

Palavras-chave: Teoria de acidentes, Aleatorização, Mecanismo destrutivo, Covariadas, Imunoterapia, Fragilidade interna, Fragilidade externa, Fragilidade, Distribuição Waring Generalizada.

ABSTRACT

VASQUEZ, J. J. K. **An useful variance decomposition for destructive Waring regression long-term model**. 2019. 111 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

The goal of this work is to formulate a two-stage regression long-term model, whose destructive mechanism of the competitive risk factors is flexible for measuring the impact on the survival function or cure rate of three variance components induced by: randomness effects, external effects or external frailties (unknown covariates) and destruction or internal frailty (destructive mechanism). The number of the risk factors which were not eliminated is unobservable random variable, called discrete frailty, and the choice of the frailty distribution must be appropriate to detect the sources of variability responsible for the variation between patients. The discrete frailty random variable of the first-stage of the model is based on the Waring distribution, which splits the variance into these three components, and was applied with success in the accident theory, epidemiology and biology. A simulation study and an application to a HIV and melanoma data, via likelihood approach, illustrate the utility of the Waring distribution to detect internal frailty, external frailty and model's uncertainty (randomness effect), which are not observable and responsible for the heterogeneity across patients. The cure rate is personalized and the patient is a protagonist for the treatment, and that could be useful to decide on preventive immunotherapy treatment for patients to fight cancer.

Keywords: Accident theory, Randomness, Covariates, Internal frailty, External frailty, Immunotherapy, Frailty, Destructive mechanism, Generalized Waring distribution.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BN	Binomial Negativa
DP	Desvio Padrão
EMV	Estimativa de Máxima Verosimilhança
EP	Erro Padrão
fgp	função geradora de probabilidade
MC	Monte Carlo
MRBNLD	Modelo de Regressão Binomial Negativa de Longa Duração
MRDWLD	Modelo de Regressão Destrutivo Waring de Longa Duração
PC	Probabilidade de Cobertura
WG	Waring Generalizada

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O modelo de longa duração em dois estágios (RODRIGUES *et al.*, 2009b) focaliza no primeiro estágio a cura dos pacientes em um cenário biológico, onde os fatores de riscos (células iniciadas ou bactérias) estão competindo para a ocorrência do evento de interesse dos pacientes. O primeiro estágio é o cérebro do modelo, onde o número de fatores de risco é controlado por um mecanismo destrutivo conhecido na área molecular ou carcinogênica como sistema imune. Neste contexto, mas sem considerar a existência de um mecanismo destrutivo ou superdispersão, Yakovlev, Tsodikov e Bass (1993) sugeriram a distribuição Poisson para o número latente de fatores de riscos, que devido a sua simplicidade ficou conhecido como modelo de promoção (CHEN; IBRAHIM; SINHA, 1999). Mais recentemente, dentro do cenário biológico, surgiu o interesse em investigar a superdispersão dos fatores de riscos supondo que a média da distribuição Poisson segue uma distribuição gama. Neste caso específico, a distribuição Poisson foi flexibilizada surgindo a distribuição Binomial Negativa (BN). Entretanto, o mecanismo de defesa da distribuição BN não é flexível o suficiente para permitir a possibilidade de eliminação dos fatores de riscos sem ruídos externos. Os ruídos externos são covariadas influentes não observadas e não incluídas no modelo, sendo conhecidas como efeito ambiental na teoria de acidentes apresentada no Apêndice A. A distribuição Waring Generalizada (WG) (IRWIN, 1975) é a mais apropriada para o modelo de longa duração em dois estágios, devido a existência de um mecanismo destrutivo flexível que depende exclusivamente dos pacientes. Em outras palavras, a distribuição WG é uma mistura da distribuição BN com a distribuição Beta que personaliza o mecanismo destrutivo e a taxa de cura, transformando os pacientes em protagonistas do tratamento. Este modelo misto com um mecanismo destrutivo inteligente (flexível) explica a heterogeneidade através de três fontes de variabilidades que impactam a taxa de cura e a sobrevivência dos pacientes: aleatorização, efeito externo ou fragilidade externa e efeito destrutivo ou fragilidade interna. Estas fontes de variabilidades são identificadas através da representação hierárquica em três níveis da distribuição WG (ver Apêndice A).

O número de fatores de riscos responsáveis pela sobrevivência dos pacientes pode ser

visto como um efeito aleatório latente (fragilidade discreta), que expressa o comportamento heterogêneo dos pacientes em relação ao risco básico. Este comportamento está diretamente conectado ao fenômeno de superdispersão dos fatores de riscos e o correspondente mecanismo de defesa. Na literatura mais recente sobre os modelos de longa duração em dois estágios, várias distribuições discretas com caudas pesadas ("J-shaped") têm sido utilizadas para explicar um possível excesso de variabilidade. Entretanto, estes modelos como a distribuição BN não conseguem separar o efeito individual (destrutivo) dos efeitos externos que interferem na superdispersão.

O efeito destrutivo ou fragilidade interna é a principal causa da superdispersão dos fatores de riscos, e fundamental para entender a evolução e o tratamento da doença. Este conceito é análogo ao efeito individual chamado "proneness", que surgiu na teoria de acidentes com Greenwood e Yule (1920), e discutido com grande visibilidade em Bates e Neyman (1952), Irwin (1968). Estes autores, curiosamente, formularam a distribuição BN supondo que a taxa de ocorrência da Poisson é gerada de uma distribuição gama. Este parâmetro da distribuição Poisson foi utilizado para explicar a variabilidade interna responsável pelo acidente com o nome de "proneness". Se definirmos o acidente como um fator de risco em análise de sobrevivência de longa duração, este parâmetro é conhecido como "fragilidade" ou efeito aleatório latente responsável pela variação ou heterogeneidade entre os pacientes. Na teoria de acidentes, o excesso de variabilidade expresso pela distribuição BN é conhecido como sensibilidade. A sensibilidade na teoria de acidentes ou na área biológica é a mistura do efeito interno com o externo separada do efeito aleatório representado pela distribuição Poisson. A vantagem da distribuição WG sobre a distribuição BN é a separação da sensibilidade ou fragilidade em análise de sobrevivência em fragilidade interna (mecanismo destrutivo) e fragilidade externa (covariadas desconhecidas não incluídas no modelo).

Devido a relevância da teoria de acidentes (IRWIN, 1968) como motivação para este trabalho, um resumo é apresentado no Apêndice A com as principais fontes de variabilidades responsáveis pela dispersão dos acidentes. A primeira fonte de variabilidade é uma componente aleatória inerente ao modelo Poisson (média=variância), e as duas outras componentes são não aleatórias e não consideradas nos tradicionais modelos de regressão: o efeito externo (submissão) e o efeito interno (predisposição para evitar acidentes). Na teoria de acidentes a submissão é conhecida como "liability" e a predisposição como "proneness". Para caracterizar estas três fontes de heterogeneidade Irwin (1968) formulou um modelo hierárquico em três estágios, conhecido na teoria de acidentes como distribuição WG (Apêndice A). Recentemente este modelo foi utilizado com sucesso para entender o comportamento dos motoristas envolvidos em acidentes de tráfico (PENG; LORD; ZOU, 2014).

1.1 Objetivos do trabalho

Motivado pela teoria de acidentes (Apêndice A), os fatores de riscos do modelo de análise de sobrevivência de longa duração serão considerados como "acidentes", e a distribuição WG será utilizada para identificar as principais fontes de heterogeneidades do comportamento dos fatores de riscos responsáveis pela doença dos pacientes. Nesta dissertação o efeito "submissão" será considerado como uma fonte de variabilidade devido aos efeitos externos ou variáveis de regressão desconhecidas e não incluídas no modelo, que chamaremos de **"fragilidade externa"**. A "predisposição" como uma fonte de variabilidade interna ou subjetiva do paciente (mecanismo natural de defesa do paciente), que chamaremos de **"fragilidade interna"**, e a **"aleatorização"** como uma fonte de variabilidade aleatória gerada pela incerteza do modelo. Para analisar a superdispersão dos fatores de riscos em análise de sobrevivência será essencial, como na teoria de acidentes, um estudo das relações entre estas fontes de variabilidades em relação a complexidade da doença, isto é, quando o número médio de fatores de risco aumenta. Com esta adaptação da teoria de acidentes em análise de sobrevivência de longa duração, este trabalho está focalizado nos seguintes objetivos:

- Formular um modelo de longa duração em dois estágios com um mecanismo destrutivo (fragilidade interna) dos fatores risco competitivos, e avaliar o seu impacto na taxa de cura.
- Utilizar a distribuição WG para separar as fontes de variabilidades internas (fragilidade interna), externas (fragilidade externa) e aleatórias. A fragilidade interna e externa caracterizam um novo conceito de "fragilidade" úteis no planejamento preventivo e personalizado dos pacientes, que em analogia com a teoria de acidentes chamaremos de **"sensibilidade"**.
- Abordar a fragilidade discreta do primeiro estágio do modelo de longa duração como um paradigma diretamente conectado ao fenômeno de superdispersão (variabilidade acima da média), que pode ser do tipo aleatório, externo ou interno.
- Definir uma taxa de cura personalizada, isto é, que dependa do mecanismo destrutivo dos pacientes e comparar com os modelos alternativos.
- Formular o modelo de regressão destrutivo Waring (caso particular da WG) de longa duração e avaliar as suas vantagens e desvantagens através de simulações e dados reais.

1.2 Organização dos Capítulos

Este trabalho está organizado como segue: No Capítulo 2 apresentamos o modelo unificado de longa duração em dois estágios e os resultados mais importantes. No Capítulo 3 apresentamos o modelo de regressão Binomial Negativo de longa duração e uma aplicação com dados de sinusite em pacientes HIV positivos e negativos. No Capítulo 4 apresentamos o modelo de regressão destrutivo Waring de longa duração. No Capítulo 5 um estudo de simulação e duas aplicações do modelo de regressão destrutivo Waring de longa duração são apresentados: análise dos dados de sinusite em pacientes HIV positivos e negativos e a análise de dados de melanoma. No Capítulo 6 apresentamos as conclusões finais e as pesquisas futuras. Um resumo da teoria de acidentes, a distribuição Waring Generalizada , provas dos resultados obtidos e os programas das simulações e aplicações em RSudio são apresentados nos Apêndices A , B e C, respectivamente.

CAPÍTULO

2

MODELO UNIFICADO DE LONGA DURAÇÃO

A teoria unificada dos modelos de longa duração (RODRIGUES; CANCHO; CASTRO, 2008; RODRIGUES *et al.*, 2009a) está baseada em um processo biológico com dois estágios: iniciação e promoção dos fatores de riscos do evento de interesse. Identificar as fontes de variabilidades da superdispersão dos fatores de riscos, que influenciam a sobrevivência dos pacientes, é o principal desafio na formulação de um modelo com taxa de cura, onde o paciente é o protagonista do tratamento.

2.1 Teoria unificada de longa duração

A teoria unificada de longa duração é caracterizada pelos seguintes estágios:

- **Estágio de Iniciação:** Seja M uma variável aleatória representando o número de fatores de risco ou as causas da ocorrência de um evento de interesse com a seguinte distribuição de probabilidades

$$p_m = P[M = m],\ m = 0,1,2,\ldots \tag{2.1}$$

- **Estágio de Promoção:** Dado $M = m$, seja $Z_j, j = 1,2,\ldots,m$ variáveis aleatórias contínuas independentes (não negativas) com distribuição $F(z) = 1 - S(z)$ e independentes de M, representando o tempo de ocorrência ou promoção do evento de interesse devido ao j-ésimo fator de risco. O tempo de ocorrência do evento de interesse é definido como

$$T = \begin{cases} Y, & M \geq 1 \\ \infty, & M = 0, \end{cases}$$

onde $Y = \min\{Z_1, Z_2, \ldots, Z_M\}$ e $P[Z_0 = \infty] = 1$.

O estágio de iniciação é o cérebro do modelo de cura em dois estágios, onde os fatores de risco estão competindo e comunicando-se para determinar a ocorrência do evento de interesse. A variável M representa a heterogeneidade dos pacientes ao longo da população, conhecida como fragilidade discreta e a sua superdispersão será fundamental para entender este comportamento heterogêneo entre os pacientes. O segundo estágio conecta a variável latente M com os dados para entender a origem de sua variabilidade, e obter soluções inferenciais mais precisas e menos viciadas. O evento de interesse pode ser a morte do paciente, ocorrência de uma sinusite ou reincidência de um câncer. Os fatores de riscos podem ser células defeituosas que causariam o tumor ou bactérias que causam a sinusite. O número de fatores de riscos M é desconhecido com distribuição de probabilidades p_m previamente definida. A escolha da distribuição de M é fundamental para identificar as possíveis fontes de variabilidades da superdispersão, tal que o paciente seja o protagonista do tratamento. O estágio de iniciação permite a possibilidade de cura com probabilidade p_0. A variável resposta T definida no estágio de promoção é observada ou censurada. O impacto da superdispersão na sobrevivência do paciente fica evidente através da função de sobrevivência de longa duração em dois estágios dada por Rodrigues *et al.* (2009a):

$$S_p(t) = P[T > t] = E_M[S^M(t)] = A_M[S(t)], \tag{2.2}$$

onde $A_M[.]$ é a função geradora de probabilidade (fgp) da variável M. A expressão em (2.2) resume os dois estágios em um mecanismo biológico consistente, que naturalmente explica a variabilidade dos dados através da superdispersão de M e da função de sobrevivência dos pacientes em risco, $S(t)$. A expressão em (2.2) pode ser vista como o clássico modelo de fragilidade com estrutura de risco proporcional de Cox (COX, 1972) dada por

$$S_p(t) = E_M[e^{-MH(t)}], \tag{2.3}$$

onde $H(t) = -log[S(t)]$ é a função básica de risco acumulado comum aos pacientes da população. Para uma justificativa da equação (2.2) sugerimos ver Apêndice B. Dada a função de sobrevivência de longa duração, $S_p(t)$, temos que

$$\lim_{t\to\infty} S_p(t) = P[M = 0] = p_0,$$

onde p_0 é a proporção de indivíduos "curados" ou "imunes", que podem estar presentes na população. O resultado é obtido trivialmente a partir de (2.2). Um indivíduo é considerado imune se não está sujeito ao evento de interesse, e a sua correspondente probabilidade p_0 é a taxa de cura. Para que o paciente seja o protagonista do tratamento é importante que a taxa de cura seja personalizada, sendo um dos principais objetivos da teoria unificada com superdispersão apresentada neste trabalho.

A função de sobrevivência de longa duração pode ser reescrita como um modelo de mistura padrão entre imunes e não imunes (BERKSON; GAGE, 1952) como

$$S_p(t) = p_0 + (1 - p_0) S_p^*(t), \tag{2.4}$$

onde

$$S_p^*(t) = \sum_{m=1}^{\infty} p_m^* \{S(t)\}^m,\ p_m^* = \frac{p_m}{1-p_0}.$$

A prova da equação (2.4) está disponível no Apêndice B.

2.2 Propriedades

A seguir apresentamos as principais propriedades do modelo unificado, que serão importantes para a solução dos problemas inferenciais e a sua respectiva implementação computacional. Para mais detalhes sugerimos ver Rodrigues, Cancho e Castro (2008) e Rodrigues *et al.* (2009a).

- **Função de densidade:**

$$f_p(t) = f(t)\left(\frac{dA(s)}{ds}|_{s=S(t)}\right).$$

- **Função de risco:**

$$h_p(t) = \frac{f_p(t)}{S_p(t)} = f(t)\frac{\frac{dA(s)}{ds}|_{s=S(t)}}{S_p(t)}.$$

- **Função de densidade própria:**

$$f^*(t) = -\frac{dS_p^*(t)}{dy} = f(t)\frac{\frac{dA(s)}{ds}|_{s=S(t)}}{1-p_0}.$$

- **Função de risco própria:**

$$h^*(t) = \frac{f^*(t)}{S_p^*(t)} = \frac{S_p(t)h_p(t)}{(1-p_0)S_p^*(t)} = \frac{S_p(t)}{S_p(t)-p_0}h_p(t),$$

ou

$$h_p(t) = h^*(t)\frac{(1-p_0)S_p^*(t)}{S_p(t)}.$$

A função de densidade própria representa uma função de densidade com as propriedades usuais conhecidas em probabilidades. A função densidade $f_p(t)$ é impropria no sentido que a integral correspondente é menor que um.

Motivado pelo fenômeno da superdispersão apresentado no Apêndice A (teoria de acidentes), a escolha da distribuição de probabilidades do primeiro estágio é crucial para avaliar o processo destrutivo como um efeito individual análogo ao efeito "proneness" e o efeito externo como "liability". Estes efeitos juntamente com o efeito aleatório são essenciais para avaliar as fontes de dispersão responsáveis pela superdispersão dos fatores de riscos e o impacto na

sobrevivência dos pacientes. Neste sentido, as distribuições BN e WG serão importantes para uma interpretação biológica mais realista, que serão apresentadas com detalhes nos próximos capítulos.

A fragilidade discreta M em (2.3) representa a heterogeneidade dos pacientes na população (diferentes riscos) em relação ao risco básico acumulado $H(u)$. Infelizmente, os modelos de fragilidades existentes na literatura não discutem ou focalizam a origem da superdispersão para um determinado paciente. Entender a origem da superdispersão é fundamental para entender a evolução da doença e formular tratamentos preventivos eficientes. Por exemplo, na medicina moderna ou de precisão já existe a preocupação em entender subjetivamente a origem da superdispersão, e formular tratamentos adicionais personalizados que colaborem para o sucesso do tratamento principal.

É evidente a partir da fgp da variável M, que a superdispersão pode causar um grande impacto na função de sobrevivência de um paciente, e identificar as fontes que geram esta superdispersão é o principal objetivo nos próximos capítulos. A seguir apresentamos os seguintes modelos probabilísticos com as suas respectivas fontes de variabilidades, que serão apresentados e utilizados na formulação de novos modelos de regressão de longa duração:

- Distribuição Poisson: Efeito aleatório ou equidispersão.
- Distribuição Binomial Negativa (BN: superdispersão): Efeito externo (covariadas desconhecidas não incluídas no modelo) combinado com o efeito destrutivo.
- Distribuição Waring Generalizada (WG: superdispersão): Efeitos aleatório, externo e destrutivo.

A distribuição de WG é uma flexibilização da distribuição BN (Apêndice A) ao introduzir um terceiro nível no modelo hierárquico, que separa o efeito externo do efeito individual (destrutivo). Com esta separação é possível avaliar, para cada paciente, a influência dos efeitos aleatórios, externos e o mecanismo destrutivo na superdispersão dos fatores de riscos, e elaborar um tratamento mais eficiente e personalizado. Na literatura existem vários conceitos de fragilidades, mas nenhum deles discutem o impacto externo e individual na fragilidade, sobrevivência e taxa de cura do paciente. Para uma discussão mais profunda, sobre a origem da fragilidade do ponto de vista médico, demográfico e gerontológico, sugerimos ver Wienke (2010).

CAPÍTULO

3

DECOMPOSIÇÃO DA VARIÂNCIA PARA O MODELO DE REGRESSÃO BINOMIAL NEGATIVO DE LONGA DURAÇÃO

Como foi amplamente discutido na teoria de acidentes apresentada no Apêndice A, a distribuição de Poisson com parâmetro λ pode ser flexibilizada através de uma distribuição de mistura gama com parâmetros a e v, tal que $M \sim BN(a,p)$, com $p = \frac{1}{1+v}$ e fgp dada por

$$A_M(s) = \left[\frac{p}{1-(1-p)s}\right]^a. \tag{3.1}$$

A média e variância são dadas respectivamente, por

$$\begin{aligned} E[M] &= \frac{a}{\frac{1}{v}} = av = \mu \\ V[M] &= \mu(1+v), \end{aligned}$$

onde v é o parâmetro de superdispersão e $a = \frac{\mu}{v}$ o índice de variação da superdipersão em relação a média (μ: fixada) ("indice parameter", (XEKALAKI, 2014)). Quando o parâmetro v decresce e o parâmetro a cresce, a distribuição BN tende a distribuição de Poisson. Nesse sentido, vamos chamar a de índice de precisão. Portanto, $M \sim BN(a,p)$, onde $a = \frac{\mu}{v}$ é o índice de precisão, $p = \frac{1}{1+v}$ a probabilidade de destruir o fator de risco e v o parâmetro de superdispersão.

3.1 Decomposição da variância

O modelo de regressão BN de longa duração em dois estágios é caracterizado pela função de sobrevivência (RODRIGUES *et al.*, 2009a)

$$S_p(t) = A_M[S(t)] = \left[\frac{p}{1-(1-p)S(t)}\right]^a, \tag{3.2}$$

e pela seguinte função de ligação que varia ao longo da população de pacientes:

$$\mu = av = e^{\beta_0 + x^T \beta},$$

onde $x^T = (x_1, x_2 \ldots, x_q)$ é o vetor de covariadas de dimensão q, β o vetor de coeficientes de regressão e β_0 o parâmetro de referência para os pacientes da população.

A Tabela 1 apresenta as fontes de variabilidades da dispersão dos fatores de riscos do i-ésimo paciente: aleatorização e sensibilidade. Como na teoria de acidentes, a sensibilidade é definida como a mistura do efeito externo (λ) e o efeito individual (p):

Tabela 1 – Decomposição da variância total do modelo BN.

Fonte de Variabilidade	**Variância**	**Taxa de Variação**
Aleatório	μ	$\frac{a}{a+\mu} = p$
Sensibilidade	$\frac{1}{a}\mu^2$	$\frac{\mu}{a+\mu} = 1-p$
Total	$\mu + \frac{1}{a}\mu^2$	1

A última coluna expressa o impacto da sensibilidade em relação ao efeito aleatório na dispersão do fatores de riscos, e consequentemente na sobrevivência e cura dos pacientes. Para a fixo, um número médio muito grande de fatores de riscos implica em um forte impacto da sensibilidade na cura e sobrevivência dos pacientes. Para μ pequeno predomina o efeito aleatório similar ao modelo de promoção. Infelizmente, o modelo de regressão de longa duração BN em dois estágios não captura separadamente o efeito externo (λ) e o efeito individual (p) na dispersão dos fatores de riscos.

3.2 Aplicação: Dados de sinusite em pacientes HIV positivos e negativos

Como uma aplicação ilustrativa da necessidade de conhecer as fontes de variabilidades da superdispersão dos fatores de riscos, vamos analisar nesta seção os dados de sinusite em Colosimo e GIOLO (2006), onde as bactérias responsáveis pela ocorrência da sinusite correspondem as células danificadas no processo carcinogênico. Estes dados consistem de 103 pacientes que foram acompanhados no período de março de 1993 a fevereiro de 1995, com o objetivo de verificar a hipótese que a infecção pelo HIV aumenta o risco de ocorrência da sinusite (em dias). A percentagem de observações censuradas foi de 80,6% e as estimativas de Kaplan-Meier da sobrevivência dos pacientes HIV positivos e negativos com risco de sinusite são apresentadas na Figura 1.

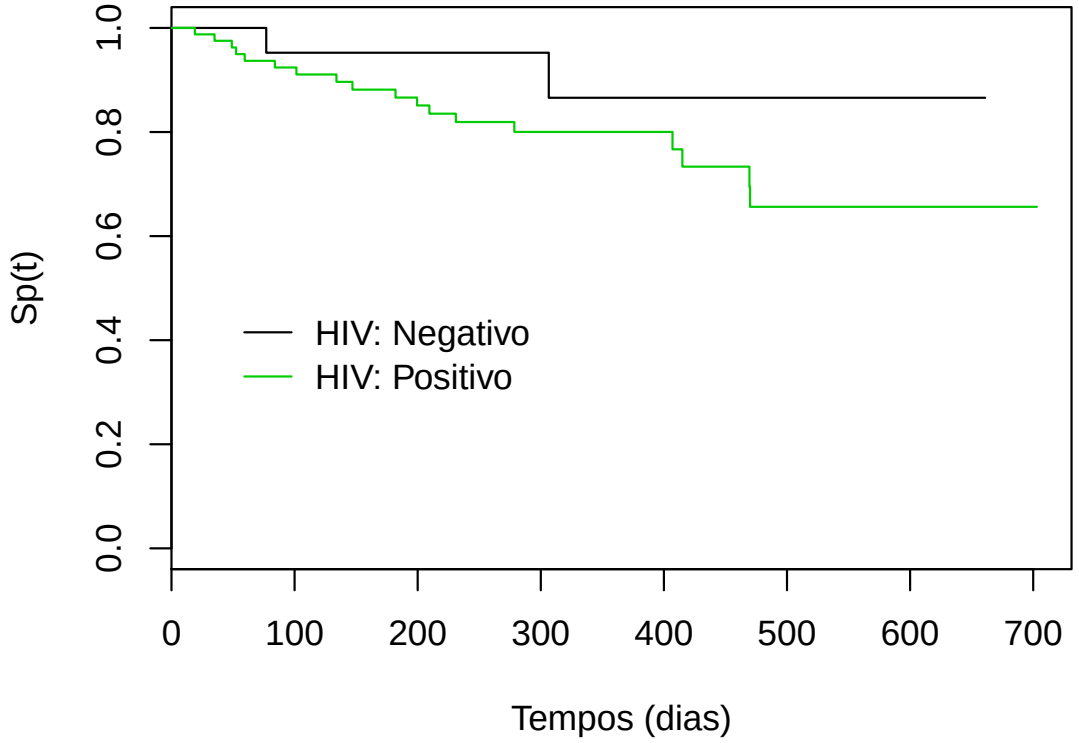

Figura 1 – Curva de sobrevivência estimada por meio do estimador de Kaplan-Meier para os dados de sinusite em pacientes HIV positivos e negativos

Fonte: Elaborada pelo autor.

As estimativas de Kaplan-Meier na Figura 1 sugerem a possibilidade de cura dos pacientes com HIV (positivo) e sem HIV (negativo) após um longo período de observação. Os modelos tradicionais de análise de sobrevivência formulados em Colosimo e GIOLO (2006) não levam em conta a possibilidade de não ocorrência do evento de interesse. Os modelos que foram utilizados consideram a alta taxa de censura como uma justificativa da ausência de um bom ajuste após um tempo prolongado. Motivado pelo estimador de Kaplan-Meier, vamos propor para estes dados um modelo BN de longa duração com dois estágios, e avaliar o impacto da superdispersão no número de bactérias responsáveis pela sinusite.

O modelo de regressão BN de longa duração para estes dados consiste nos seguintes estágios:

- **Estágio de Iniciação:** Seja $M_i,\ i = 1,...,103$, variáveis aleatórias que representam o número de bactérias que competem entre si para a ocorrência da sinusite em 103 pacientes HIV positivos e negativos. Além disso, vamos supor que estas variáveis discretas são independentes com distribuição BN com parâmetros a e v_i. A taxa de cura para o i-ésimo

paciente é dada por

$$p_{0,i} = P[M_i = 0] = \left(\frac{1}{1+v_i}\right)^a, i = 1, ..., 103. \tag{3.3}$$

A taxa de cura depende totalmente do mecanismo destrutivo do paciente definido como $p_i = \frac{1}{1+v_i}$, onde a representa o índice de precisão.

- **Estágio de Promoção:** Suponhamos que as variáveis não observáveis, $Z_1, ..., Z_{M_i}$ ($M_i \geq 1$) (independentes dado M_i), que representam o tempo que cada bactéria do i-ésimo paciente leva para promover a sinusite, seguem uma distribuição Weibull com função de sobrevivência $S(t) = e^{-e^{\gamma_t}\alpha}$, com $\alpha > 0$ e $-\infty < \gamma < \infty$.

Segue de Rodrigues *et al.* (2009a) o seguinte modelo de longa duração com superdispersão:

$$S_p(t) = \left[\frac{p}{1-(1-p)S(t)}\right]^a \text{ e a correspondente fdp imprópria}$$
$$f_p(t) = af(t)\left[\frac{p}{1-(1-p)S(t)}\right]^a \left[\frac{1-p}{1-(1-p)S(t)}\right].$$

O modelo de regressão BN de longa duração é finalizado com a seguinte covariada:

$$x_i = \begin{cases} 1, \text{HIV positivo}, \\ 0, \text{HIV negativo}. \end{cases}$$

A covariável x_i indica se o i-ésimo paciente é HIV positivo ou negativo. A ligação entre o parâmetro p_i e a covariada x_i do i-ésimo paciente é definida como:

$$\mu_i = av_i = e^{\beta_0+\beta_1 x_i} \Rightarrow v_i = \frac{e^{\beta_0+\beta_1 x_i}}{a},$$
$$p_i = (1+v_i)^{-1} \Rightarrow p_i = \left[1 + \frac{e^{\beta_0+\beta_1 x_i}}{a}\right]^{-1}.$$

Analisando os dados de $n = 103$ pacientes com ocorrência da sinusite em Colosimo e GIOLO (2006), a correspondente função de verosimilhança para $\vartheta = (\alpha, \gamma, a, \beta_0, \beta_1)$ é dada por:

$$\begin{aligned} L(\vartheta|D) &= \prod_{i=1}^{103} \left\{ [f_p(t_i)]^{\delta_i} [S_p(t_i)]^{1-\delta_i} \right\} \\ &= \prod_{i=1}^{103} \left\{ af(t_i) \left[\frac{p_i}{1-(1-p_i)S(t_i)}\right]^a \left[\frac{1-p_i}{1-(1-p_i)S(t_i)}\right] \right\}^{\delta_i} \\ &\quad \left\{ \left[\frac{p_i}{1-(1-p_i)S(t_i)}\right]^a \right\}^{1-\delta_i}, \end{aligned}$$

onde $D = \{(t_i, \delta_i), i = 1, 2, \ldots, n\}$ e δ_i a variável que indica a ocorrência ou não da censura dada por

$$\delta_i = \begin{cases} 0: \ t_i \text{ é censurado,} \\ 1: \ t_i \text{ é não censurado.} \end{cases}$$

A Estimativa de Máxima Verosimilhança (EMV) $\hat{\vartheta}$, o erro padrão (EP) e o p-valor foram obtidos numericamente pelo método L-BFGS-B (R Core Team, 2018), utilizando os códigos disponíveis no Apêndice C. Os resultados obtidos foram os seguintes:

Tabela 2 – Estimativa de máxima verosimilhança (EMV), Erro padrão (EP).

	EMV	EP	p-valor
$\hat{\alpha}$	0,933	0,164	0,000
$\hat{\gamma}$	-4,708	0,904	0,000
$\hat{a}$	5,000	7,669	0,514
$\hat{\beta}_0$	-2,188	0,718	0,002
$\hat{\beta}_1$	0,980	0,755	0,194

Na Tabela 2, observamos que a estimativa do parâmetro a apresenta um alto Erro Padrão (EP) em relação as outras estimativas dos parâmetros envolvidos no modelo. Uma justificativa seria a forte influência do parâmetro de forma (a) e do coeficiente de regressão β_0, na ligação entre o coeficiente de regressão β_1 e o parâmetro de escala v, isto é,

$$e^{\beta_1 x_i} = \frac{a}{e^{\beta_0}} v_i, \ i = 1, 2, \ldots, n,$$

para $\beta_0 \neq \log(a)$.

A influência dos parâmetros a e β_0 e a sensibilidade de a como parâmetro de forma, justificam um problema de identificabilidade destes parâmetros produzindo estimativas viciadas e imprecisas. Um estudo de simulação será apresentado no Capítulo 5 para avaliar o impacto da falta de identificabilidade do modelo de regressão BN de longa duração, e propor no Capítulo 4 o modelo destrutivo de regressão Waring de longa duração como uma alternativa para minimizar o problema de identificabilidade. O modelo destrutivo de regressão Waring de longa duração adiciona mais um nível na representação estocástica da BN tornando o modelo mais identificável, que também no Capítulo 5 será justificado através de simulações.

Como o tempo de promoção das bactérias para ocorrência da sinusite segue uma distribuição de Weibull com parâmetros α e γ , a função de sobrevivência de longa duração para o paciente HIV positivo ou negativo é dada por:

$$\hat{S}_p(t) = \begin{cases} \left[\dfrac{\left[1+\frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{\hat{a}}\right]^{-1}}{1-\left(1-\left[1+\frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{\hat{a}}\right]^{-1}\right)e^{-e^{\hat{\gamma}_t}\hat{\alpha}}} \right]^{a} & : \text{ se } x=1, \\ \left[\dfrac{\left[1+\frac{e^{\hat{\beta}_0}}{\hat{a}}\right]^{-1}}{1-\left(1-\left[1+\frac{e^{\hat{\beta}_0}}{\hat{a}}\right]^{-1}\right)e^{-e^{\hat{\gamma}_t}\hat{\alpha}}} \right]^{a} & : \text{ se } x=0. \end{cases} \tag{3.4}$$

A Figura 2 apresenta uma comparação gráfica da função de sobrevivência de longa duração em (3.4) com o estimador de Kaplan-Meier para pacientes HIV positivos e negativos.

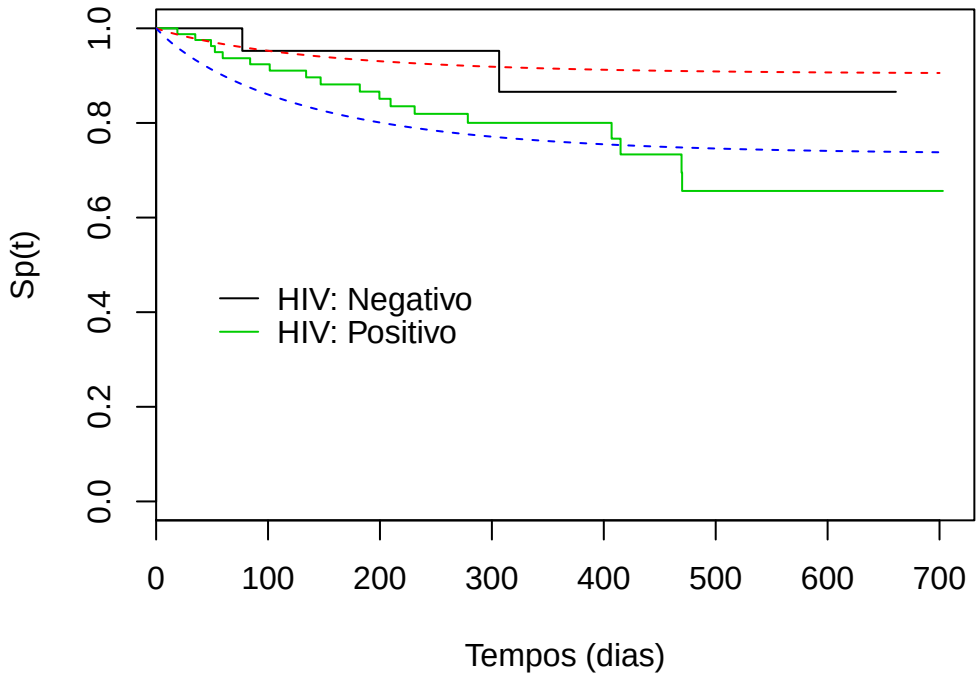

Figura 2 – Estimador de Kaplan-Meier e as estimativas da função de sobrevivência de longa duração: Distribuição BN

Fonte: Elaborada pelo autor.

Na Figura 2, as curvas em verde e preta representam as estimativas de Kaplan-Meier e as linhas de cor vermelha e azul são as estimativas das curvas de sobrevivência baseadas no modelo de regressão BN de longa duração, para os pacientes HIV negativos e positivos, respectivamente. A Figura 2 sugere que o modelo de regressão BN-Weibull de longa duração

não ajusta adequadamente aos dados de HIV com taxas de cura dadas por:

$$\begin{aligned} \hat{p_1} &= 0,89 \text{ (HIV negativo)}, \\ \hat{p_2} &= 0,75 \text{ (HIV positivo)}. \end{aligned}$$

A Tabela 3 e Tabela 4 apresentam a decomposição da variância total para os pacientes HIV negativos e positivos, respectivamente. As Figura 3 e Figura 4 ilustram graficamente (boxplot) os respectivos comportamentos das fontes de variabilidades.

Tabela 3 – Decomposição da variância da distribuição BN para os pacientes HIV negativos $(x = 0)$.

Fonte de Variabilidade	**Variância**	**Taxa de Variação**
Aleatório	$0,110$	$0,980$
Sensibilidade	$0,002$	$0,020$
Total	$0,112$	1

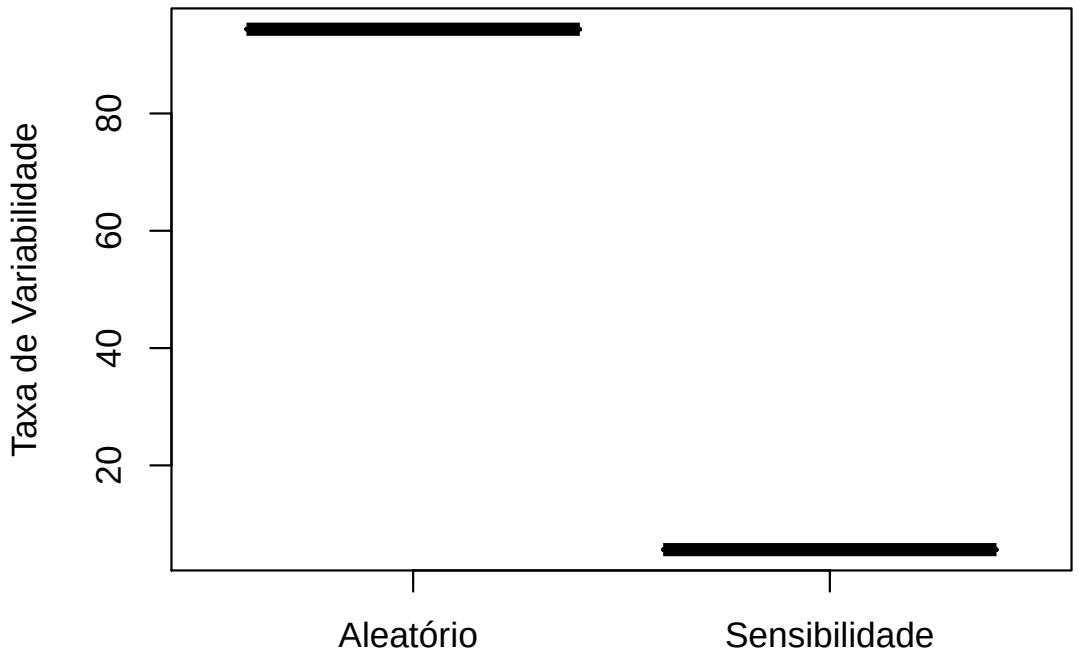

Figura 3 – Fontes de variabilidades para os pacientes HIV negativos: Distribuição BN

Fonte: Elaborada pelo autor.

Tabela 4 – Decomposição da variância da distribuição BN para os pacientes HIV positivos $(x = 1)$.

Fonte de Variabilidade	**Variância**	**Taxa de Variação**
Aleatório	$0,300$	$0,940$
Sensibilidade	$0,018$	$0,060$
Total	$0,318$	1

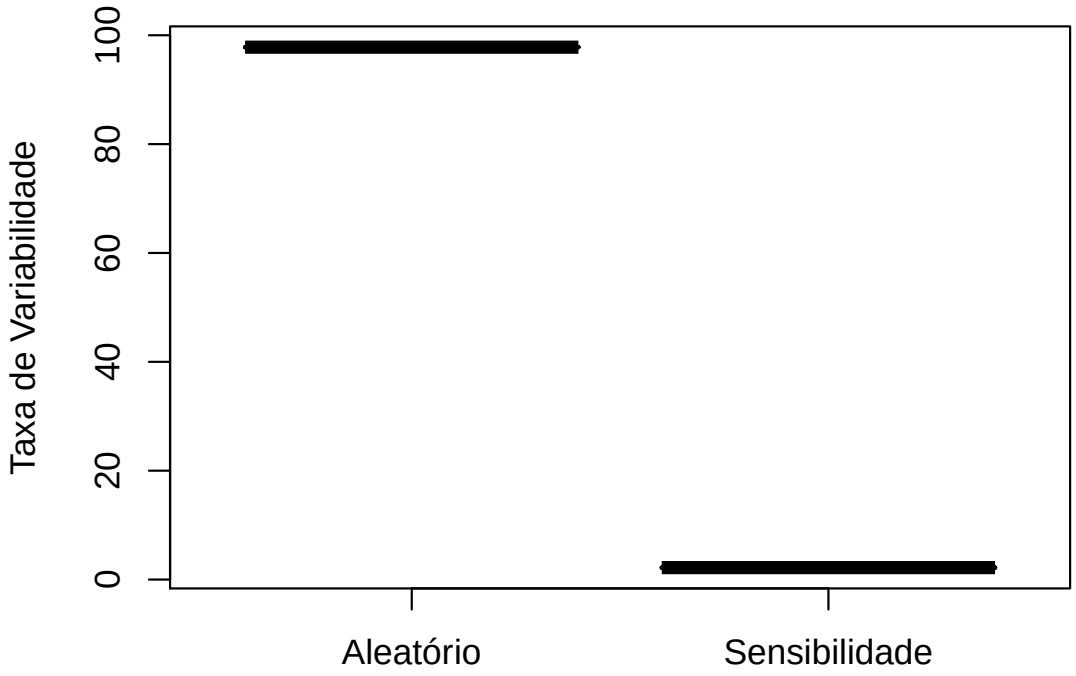

Figura 4 – Fontes de variabilidades para os pacientes HIV positivos: Distribuição BN

Fonte: Elaborada pelo autor.

As Figura 3 e Figura 4 mostram que a fonte aleatória domina fortemente a sensibilidade para os pacientes HIV positivos e negativos. O efeito HIV positivo não causa impacto significativo no comportamento dispersivo das bactérias sugerindo a equidispersão (Poisson), cujo comportamento não corresponde a realidade da doença devido a um possível problema de identificabilidade comentado anteriormente. Neste caso específico é necessário um modelo de regressão de longa duração mais flexível para verificar se existe ou não superdispersão das bactérias na presença do HIV. O modelo de regressão alternativo baseado na distribuição WG será proposto no próximo Capítulo.

CAPÍTULO

4

MODELO DE REGRESSÃO DESTRUTIVO DE LONGA DURAÇÃO

Inicialmente apresentaremos o modelo de regressão destrutivo de longa duração, onde o número de fatores de riscos do primeiro estágio segue a distribuição WG introduzida no Apêndice A, com ênfase no mecanismo destrutivo e a decomposição da variância total do primeiro estágio. As propriedades do segundo estágio podem ser obtidas diretamente do Apêndice A. Um caso particular do modelo de regressão destrutivo WG de longa duração, que será chamado de Modelo de Regressão Destrutivo Waring de Longa Duração (MRDWLD), será visto juntamente com o conceito de fragilidade interna e externa. Um estudo de simulação Monte Carlo para investigar o problema de identificação e aplicações para dados reais, deste particular modelo de regressão, serão apresentados no Capítulo 5.

4.1 Modelo regressão destrutivo Waring Generalizado de longa duração

Motivado pela teoria de acidentes apresentada no Apêndice A, vamos supor que o número de fatores de riscos do primeiro estágio do modelo de longa duração segue a distribuição WG, isto é, $M \sim WG(a,\rho,k), a>0, k>0, \rho>2$.

O parâmetro $p=\frac{1}{1+\nu}$ do terceiro estágio do modelo hierárquico da WG (ver Apêndice A) é a probabilidade do paciente eliminar o fator de risco, que será o efeito destrutivo do paciente. O efeito destrutivo corresponde ao efeito predisposição (“proneness”) na teoria de acidentes. Os parâmetros (ρ,k,p) caracterizam o mecanismo destrutivo do paciente, sem ruídos externos (covariadas desconhecidas e não incluídas no modelo de regressão destrutivo) conhecidos na teoria de acidentes (ver Apêndice A) como submissão (“liability”).

Nas próximas seções, as covariadas ou variáveis auxiliares para cada paciente serão

introduzidas no modelo através do parâmetro a, utilizando uma função de ligação associada ao parâmetro $\mu = E[M]$. Na área carcinogênica a variável M é o número de células danificadas não destruídas, que serão divididas sem controle gerando um tumor. Do ponto vista biológico e preventivo, a distribuição WG é o modelo de superdispersão mais apropriado para estimar a taxa de cura pelas seguintes razões:

- Se os parâmetros ρ e k crescem, a distribuição WG aproxima-se da distribuição BN com parâmetros μ e $p = \frac{1}{1+\theta}$, onde $\theta = \frac{k}{\rho-1}$, com média e variância dadas por (IRWIN, 1968)

$$E[M] = a\theta \tag{4.1}$$
$$Var[M] = \mu[1+\theta], \text{ respectivamente,} \tag{4.2}$$

  onde θ é o parâmetro de superdispersão. Neste caso especial, a distribuição é conhecida como distribuição "BNeg I" (RODRÍGUEZ-AVI *et al.*, 2009).

- Como $\mu = E[M] = \frac{ak}{\rho-1}$ (ver Apêndice A), temos que

$$a = \frac{\mu}{\theta},$$

  onde a é o índice de superdispersão dos fatores de riscos em relação a média, que chamaremos de índice de precisão da distribuição WG.

- O mecanismo destrutivo do paciente fica totalmente personalizado pela distribuição Beta do terceiro nível da representação estocástica da distribuição WG (ver Apêndice A): Dado o par (ρ,k), a probabilidade individual ou interna de destruição do fator de risco (p) satisfaz a condição

$$E[p \mid (\rho,k)] = \frac{\rho}{\rho+k}. \tag{4.3}$$

  A interpretação dos parâmetros (p,ρ,k) envolvidos no mecanismo destrutivo do paciente e a dependência entre ρ e k podem ser justificadas através das parametrizações formuladas em Ferrari e Cribari-Neto (2004) dadas por:

$$\rho = \mu^{\star}\phi \tag{4.4}$$
$$k = (1-\mu^{\star})\phi, \tag{4.5}$$

  onde $\mu^{\star} = E[p] = \frac{\rho}{k+\rho}$ e $\phi = k+\rho$. Para ϕ fixo temos uma correlação negativa entre k e ρ, onde ϕ é o parâmetro de precisão satisfazendo $V[p] = \frac{\mu^{*}(1-\mu^{*})}{(1+\phi)}$. O parâmetro ρ é o índice de precisão em relação a $\mu^{\star}$, e o parâmetro k é o índice de precisão em relação a $1-\mu^{\star}$.

- As fontes de variabilidades responsáveis pela superdispersão da variável M são identificadas como: aleatorização, efeito externo e destruição (ver Figura 5 ou Apêndice A). Estas fontes são úteis no planejamento preventivo, tratamento e diagnóstico da doença de uma forma personalizada, que atualmente é conhecida como imunoterapia.

- A taxa de cura fica totalmente personalizada pelo efeito destrutivo, sendo mais realista que os modelos mais recentes de longa duração (ver Apêndice A).

- Os modelos Waring (IRWIN *et al.*, 1963) e Yule (YULE, 1924) são versões mais simples do modelo WG ($k = 1$ e $k = a = 1$, respectivamente), e atrativos devido a existência de uma imensa literatura e tradição histórica em biologia, medicina e tráfego. Estes modelos particulares não apresentam problemas de identificabilidade como ocorre no modelo WG. Detalhes sobre o problema de identificabilidade da distribuição WG, para $k > 1$, sugerimos (RODRÍGUEZ-AVI *et al.*, 2009).

Uma observação interessante é a similaridade nas interpretações do parâmetro a do mecanismo destrutivo da BN com o parâmetro ρ do mecanismo destrutivo da distribuição WG:

$$a = \frac{\mu}{v}, \tag{4.6}$$

$$\rho = \frac{\mu^\star}{v^\star}, v^\star = \frac{1}{\phi}, \tag{4.7}$$

onde v é o parâmetro de superdispersão da BN , $v^\star$ o parâmetro de superdispersão da distribuição Beta e ϕ o correspondente parâmetro de precisão.

A heterogeneidade populacional dos pacientes, em relação a função risco básica acumulada, depende do mecanismo destrutivo dos pacientes como pode ser visto a seguir:

$$H_p(t) = E_M[MH(t)] = E[M]H(t) = a\frac{k}{\rho - 1}H(t). \tag{4.8}$$

A Figura 5 apresenta a representação estocástica da fragilidade discreta M, com as suas respectivas fontes de variabilidades, σ_1^2, σ_2^2 e σ_3^2, responsáveis pela superdispersão. Estas fontes de variabilidades caracterizam a decomposição da variância da fragilidade discreta M, que é o principal foco deste trabalho na formulação de modelos de longa duração. As fontes de variabilidades, no contexto da teoria de acidentes, foram discutidas nos Apêndice A e B e serão adaptadas para o modelo de regressão de longa duração na próxima seção.

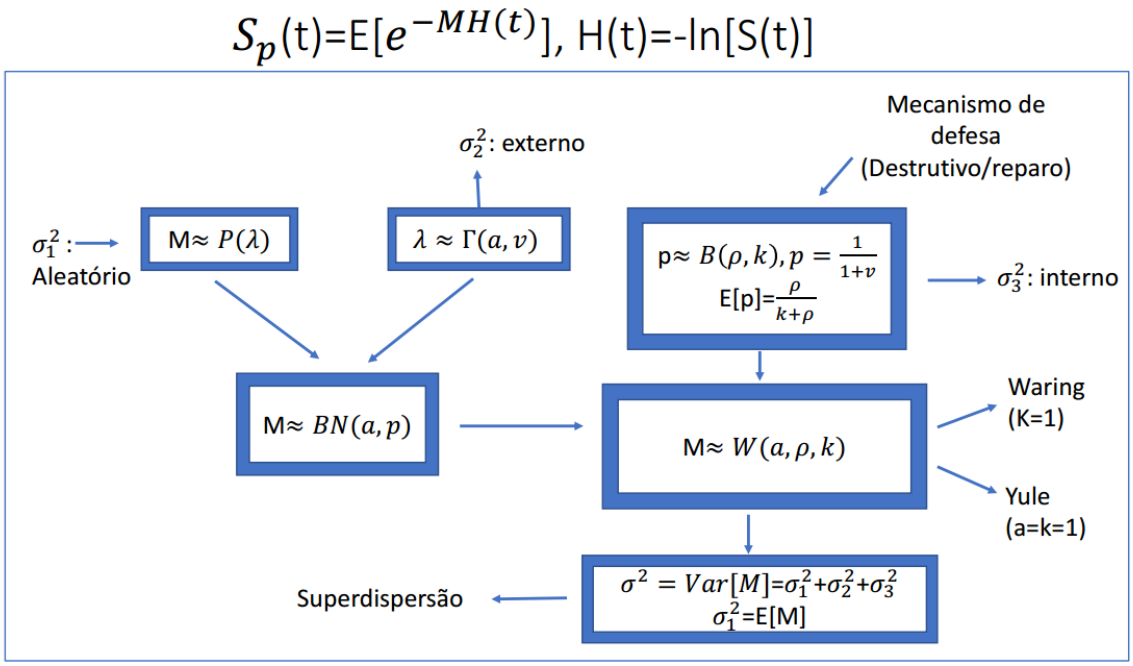

Figura 5 – Representação estocástica da fragilidade discreta M.

Fonte: Elaborada pelo autor.

Para evitar o problema de identificação da distribuição WG (RODRÍGUEZ-AVI *et al.*, 2007), na próxima seção vamos supor que o número de fatores de riscos M segue uma versão mais simples da distribuição WG, conhecida na teoria de acidentes como distribuição Waring (IRWIN *et al.*, 1963) $(k = 1)$.

4.2 Decomposição da variância para o modelo de regressão destrutivo Waring de longa duração

As propriedades da distribuição Waring, utilizadas nesta seção, foram obtidas diretamente do Apêndice A com $k = 1$. Para $k = 1$, a distribuição WG ou simplesmente Waring (IRWIN *et al.*, 1963) é dada por

$$p_m = P[M = m] = \rho \frac{(a)_m}{(a+\rho)_{m+1}}, m = 0,1,2,\ldots, \tag{4.9}$$

onde $a > 0, \rho > 2$.

A fgp é expressa como:

$$\begin{aligned} A_M(s) &= \frac{\rho}{a+\rho} {}_2F_1(a,1,a+\rho+1;s), \\ &= \frac{\rho}{a+\rho}\left\{1+\frac{as}{(\rho+a+1)}+\frac{a(a+1)s^2}{(\rho+a+1)(\rho+a+2)}+\cdots+\frac{(a)_r s^r}{(\rho+a+1)_r}+\ldots\right\}, \end{aligned} \tag{4.10}$$

onde ${}_2F_1(\cdot)$ é a função Gaussiana hipergeométrica (IRWIN, 1975). Para este modelo específico temos que a média é dada por:

$$E[M] = \mu = \frac{a}{\rho - 1} \Longrightarrow a = \mu(\rho - 1). \tag{4.11}$$

A variância total de M é dividida em três fontes de variabilidades: aleatório, efeito externo e efeito destrutivo (ver Apêndice A):

$$Var[M] = \mu + \frac{2}{\rho - 2}\mu + \frac{\rho\mu^2}{\rho - 2}. \tag{4.12}$$

Como os efeitos externos e destrutivos estão diretamente associados a superdispersão dos fatores de riscos dos pacientes, isto é, interpretam a variabilidade da fragilidade discreta M e a heterogeneidade em relação ao risco básico da população dos pacientes, vamos daqui para frente chamar de "fragilidade externa" o efeito externo e "fragilidade interna" o efeito destrutivo. Lembramos que o efeito externo refere-se as covariadas desconhecidas que são influentes na sobrevivência do paciente e não foram incluídas no modelo de regressão destrutivo.

A Tabela 5 resume as fontes de variabilidades da distribuição Waring com os novos conceitos de fragilidades:

Tabela 5 – Decomposição da variância total da distribuição Waring: $\mu = \frac{a}{\rho-1}, k = 1$.

Fonte de Variabilidade	**Variância**	**Taxa de variação**
Aleatório: Poisson	$\sigma_1^2 = \mu$	$\frac{\rho-2}{\rho}\frac{1}{1+\mu}$
Fragilidade externa: Gama	$\sigma_2^2 = \frac{2}{(\rho-2)}\mu$	$\frac{2}{\rho}\frac{1}{1+\mu}$
Fragilidade interna: Beta	$\sigma_3^2 = \frac{\rho}{(\rho-2)}\mu^2$	$\frac{\mu}{1+\mu}$
Total	$\sigma^2 = \frac{\rho}{\rho-2}(\mu+\mu^2)$	1

Para μ fixo e $\rho \longrightarrow \infty$, segue que $\sigma_1^2 = \frac{1}{1+\mu}, \sigma_2^2 = 0, \sigma_3^2 = \frac{\mu}{1+\mu}$. A sensibilidade é totalmente individual e consistente com a convergência da distribuição Waring para a distribuição $\mathbf{BN}(1, \frac{\mu}{1+\mu})$ (RODRÍGUEZ-AVI *et al.*, 2009).

A função de sobrevivência de longa duração é a composição dos dois estágios do modelo de longa duração, através da fgp da distribuição Waring dada por

$$S_p(t) = A_M[S(t)] = \frac{\rho}{a+\rho}{}_2F_1(a, 1; a+1+\rho; S(t)), \quad t > 0, \tag{4.13}$$

onde ${}_2F_1(a, 1; a+1+\rho; s)$ é a função hipergeométrica gaussiana (IRWIN, 1975), e $a = \mu(\rho - 1), \rho > 2$. A taxa de cura é dada por

$$p_0 = P[M = 0] = \frac{\rho}{a+\rho}, \rho > 2. \tag{4.14}$$

A taxa de cura expressa a probabilidade individual de eliminar todos os fatores de riscos, isto é, a porcentagem da precisão total $a+\rho$ que corresponde ao mecanismo destrutivo. A taxa de cura é personalizada devido a dependência do parâmetro ρ, confirmando o protagonismo do paciente durante tratamento.

As informações externas observáveis ou covariadas, $x^T = (1, x_1, x_2, \ldots, x_q)$ estão associadas a média μ através da função de ligação:

$$\log(\mu) = x^T \beta, \tag{4.15}$$

onde $\beta^T = (\beta_0, \beta_1, \ldots, \beta_q)$ é o coeficiente de regressão. Neste caso temos um modelo regressão destrutivo Waring de longa duração (MRDWLD), onde $a = (\rho - 1)e^{x^T\beta} = e^{\log(\rho-1)+x^T\beta}$. O parâmetro β_0 foi incluído como uma constante de referência para o vetor de covariadas x^T para cada paciente. A inclusão deste parâmetro gera um problema de identificação, que será analisado via simulação Monte Carlo (MC) no Capítulo 5.

A função de densidade para resolver os problemas inferenciais, via método de máxima-verossimilhança, é dada por

$$\begin{aligned} f_p(t) &= \frac{\rho}{a+\rho} f(t) \frac{d_2 F_1(a, 1; a+\rho+1; s)}{ds} \Big|_{s=S(t)} \\ f_p(t) &= f(t) \frac{a\rho}{(a+\rho)(a+\rho+1)} {}_2F_1(a+1, 2, a+\rho+2, S(t)), \quad t > 0. \end{aligned} \tag{4.16}$$

Para $a = 1$ temos o modelo de longa duração com distribuição de Yule (1924), onde o modelo de regressão correspondente fica definido pela função de ligação

$$\rho = 1 + e^{x^T\beta}. \tag{4.17}$$

A distribuição Waring é apropriada para análise de sobrevivência de longa duração, devido a flexibilidade em identificar o impacto das fontes de variabilidades na sobrevivência e taxa de cura dos pacientes. Os modelos de longa duração em dois estágios mais recentes não apresentam esta flexibilidade, que é importante no tratamento dos pacientes. Uma vantagem da distribuição Waring em Análise de Sobrevivência é a representação estocástica:

- Compondo a $M \mid \lambda \sim \textbf{Poisson}(\lambda)$ com $\lambda \mid a, p \sim \textbf{Gama}(a, \frac{1-p}{p})$ segue que

$$E[M \mid a, p] = E_\lambda[\lambda \mid a, p] = a\frac{1-p}{p}. \tag{4.18}$$

  A previsão do número de fatores de riscos não eliminados é controlado pelo processo BN definido por (a, p).

- Como $p \mid (\rho, 1) \sim \textbf{Beta}(\rho, 1)$, o previsor final de M é dado por

$$E[M] = \mu = e^{x^T\beta} = \frac{a}{\rho - 1}. \tag{4.19}$$

  O previsor de M depende do mecanismo destrutivo do paciente (fragilidade interna), dado por $E[p] = \frac{\rho}{1+\rho}$.

As Figura 6 e Figura 7 mostram a influência dos parâmetros a e ρ da distribuição Waring (1o. estágio do modelo de cura) no comportamento da função de sobrevivência e densidade do tempo de ocorrência do evento de interesse. O tempo de ocorrência depende do 1o. estágio e do tempo de promoção Weibull (2o. estágio do modelo de cura). Para a fixo e ρ crescendo, os gráficos mostram que os pacientes tem uma maior chance de sobrevivência com altas taxas de cura. Esta observação é consequência do sucesso do mecanismo destrutivo (fragilidade interna baixa) em eliminar os fatores de riscos. Para ρ fixo e a crescendo (fragilidade externa alta) o impacto é oposto, isto é, a chance de sobrevivência é menor com taxas de cura também menores. Embora a variável M seja latente, o impacto dos parâmetros da distribuição Waring são visíveis no tempo de vida dos pacientes. O parâmetro ρ expressa a fragilidade interna e a a fragilidade externa.

Quando a e ρ crescem para o infinito (ver Apêndice A), a distribuição Waring aproxima-se da distribuição BN$(1, \frac{1}{1+\mu})$, onde μ é o parâmetro de superdispersão, conhecida como "Negbin II" (RODRÍGUEZ-AVI *et al.*, 2009). O modelo de regressão destrutivo Waring inclui o modelo destrutivo BN ou Geométrico com probabilidade destrutiva, $p = \frac{1}{1+\mu}$, como caso limite. Como $\mu = \frac{a}{\rho-1}$, para a e ρ crescendo, a taxa de cura $p_0 = \frac{\rho}{a+\rho}$ converge para $p = \frac{1}{1+\mu}$. Neste caso limite irá prevalecer a fragilidade interna (sensibilidade: p) sobre a aleatorização somente quando a média cresce, que por outro lado está fortemente correlacionada com as covariadas através da função de ligação.

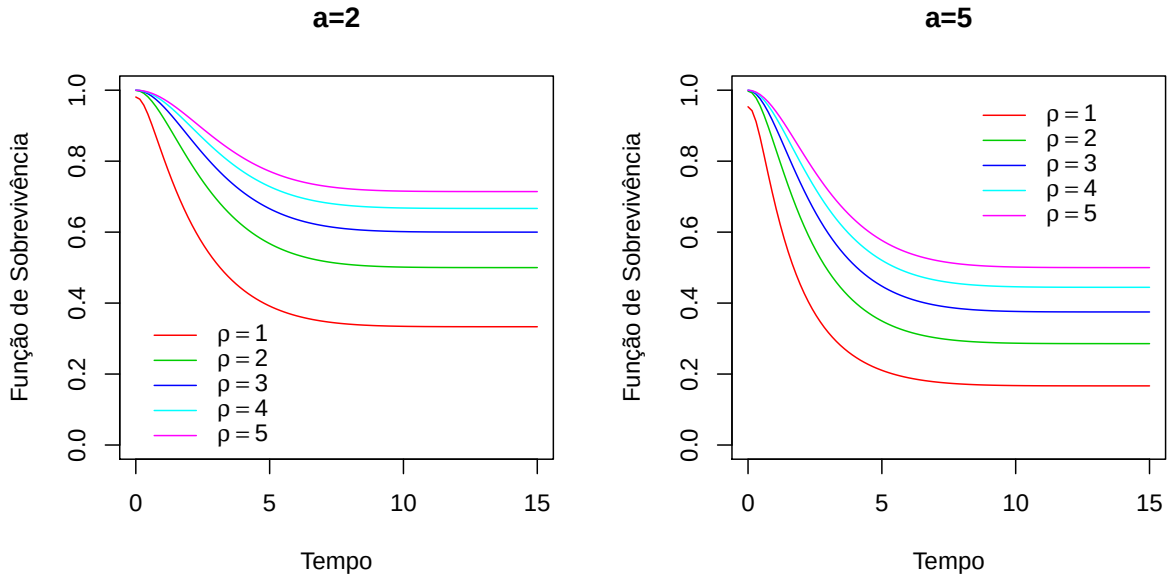

Figura 6 – Função de sobrevivência do modelo regressão destrutivo Waring de longa duração e tempo de promoção Weibull ($\alpha = 2, \gamma = -3$).

Fonte: Elaborada pelo autor.

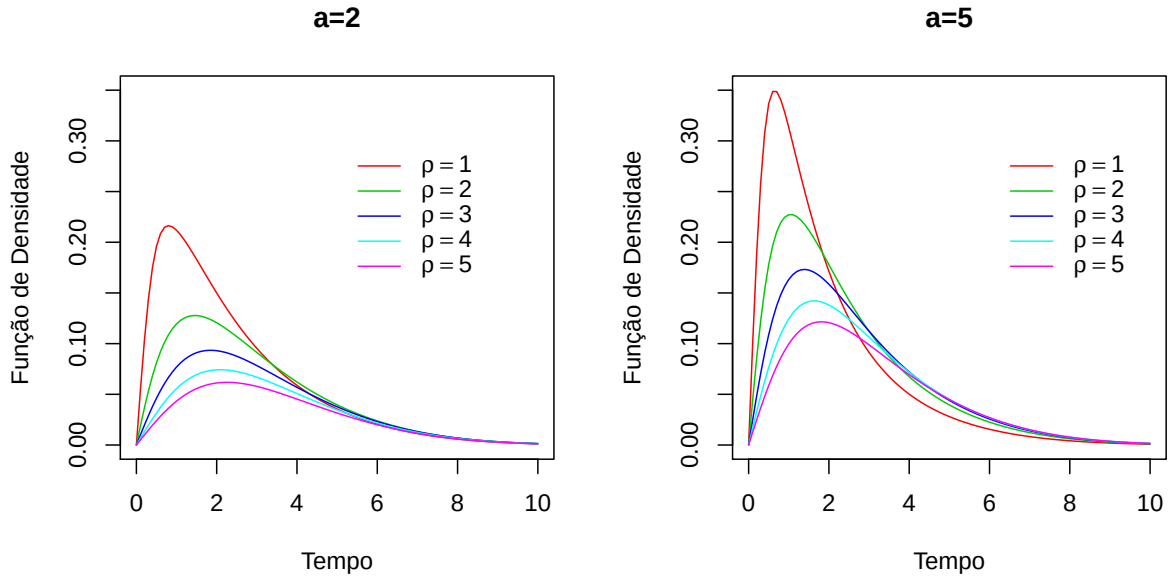

Figura 7 – Função de densidade do modelo destrutivo Waring de longa duração e tempo de promoção Weibull ($\alpha = 2, \gamma = -3$).

Fonte: Elaborada pelo autor.

CAPÍTULO

5

MODELO DE REGRESSÃO DESTRUTIVO WARING DE LONGA DURAÇÃO: SIMULAÇÃO E APLICAÇÕES

Inicialmente será apresentado um estudo de simulação Monte Carlo para estudar o problema de identificabilidade do Modelo de Regressão Binomial Negativa de Longa Duração (MRBNLD) apresentado no Capítulo 3. Em seguida será apresentado um estudo de simulação do MRDWLD para discutir as suas vantagens em relação ao modelo de regressão destrutivo BN de longa duração apresentado anteriormente. Na simulação do MRDWLD será dado uma atenção especial ao EMV dos parâmetros ρ e β_0 em relação aos outros estimadores dos parâmetros do modelo proposto. Como ilustração do MRDWLD dois conjuntos de dados serão analisados nesta seção: Dados sobre pacientes com HIV em Colosimo e GIOLO (2006) e os dados de melanoma que estão disponíveis no R Core Team (2018). Os códigos em R foram gentilmente cedidos pelo Prof. Dr. Vicente G. Cancho do ICMC-USP (ver Apêndice C).

5.1 Estudo de simulação

Para discutir o problema de identificabilidade do MRBNLD, formulado como uma mistura da distribuição de Poisson dado λ e $\lambda \sim \textbf{Gama}(a, v)$ (ver Apêndice A), um estudo de simulação Monte Carlo foi realizado com seguinte função de ligação:

$$\mu = E[M] = av = e^{\beta_0+\beta_1 x_1+\beta_2 x_2}, \tag{5.1}$$

onde o parâmetro de escala v não depende das covariadas com taxa, $\frac{Var[M]}{E[M]}$, constante. O modelo BN com esta suposição é conhecido como Negbin I (COLIN; TRIVEDI, 1986).

Para analisar o comportamento dos EMV dos parâmetros envolvidos no MRBNLD, 1000 amostras de tamanhos $n = 200, 300, 500$ foram simuladas para $v = 4$ (parâmetro de escala da

distribuição gama) com os seguintes coeficientes de regressão: $\beta_0 = 2, \beta_1 = -1, \beta_2 = -1$. As covariadas e a censura foram geradas das seguintes distribuições: $X_1 \sim \text{Binomial}(1,0,5)$, $X_2 \sim U(0,1)$ e a censura de uma uniforme no intervalo [2,5], respectivamente. O tempo de promoção foi gerado da distribuição exponencial com parâmetro $\lambda = 1$. A Tabela 6 e a Figura 8 apresentam os resultados da simulação utilizando os códigos do *software* R-Studio que estão disponíveis no Apêndice C.

Tabela 6 – Resumo das estimativas para os dados gerados com $\beta_0 = 2, \beta_1 = -1, \beta_2 = -1, \lambda = 1$ e $v = 4$.

n		λ	v	β_0	β_1	β_2
200	Média	0,811	2,801	-0,282	-1,041	-1,010
	DP	0,310	3,076	0,602	0,431	0,689
	VIÉS	-0,188	-1,198	-2,282	-0,042	-0,010
	EQM	0,362	3,300	2,360	0,432	0,689
	PC	1,000	0,890	0,428	0,965	0,946
300	Média	0,807	2.516	-0,315	-1,031	-1,029
	DP	0,273	3,107	0,538	0,353	0,532
	VIÉS	-0,193	-1,484	-2,315	-0,032	-0,029
	EQM	0,334	3,442	2,377	0,355	0,532
	PC	0,998	0,781	0,239	0,948	0,945
500	Média	0,827	1,890	-0,414	-1,020	-0,996
	DP	0,235	2,441	0,445	0,258	0,409
	VIÉS	-0,173	-2,110	-2,414	-0,020	0,004
	EQM	0,292	3,226	2,455	0,258	0,409
	PC	0,987	0,641	0,088	0,956	0,954

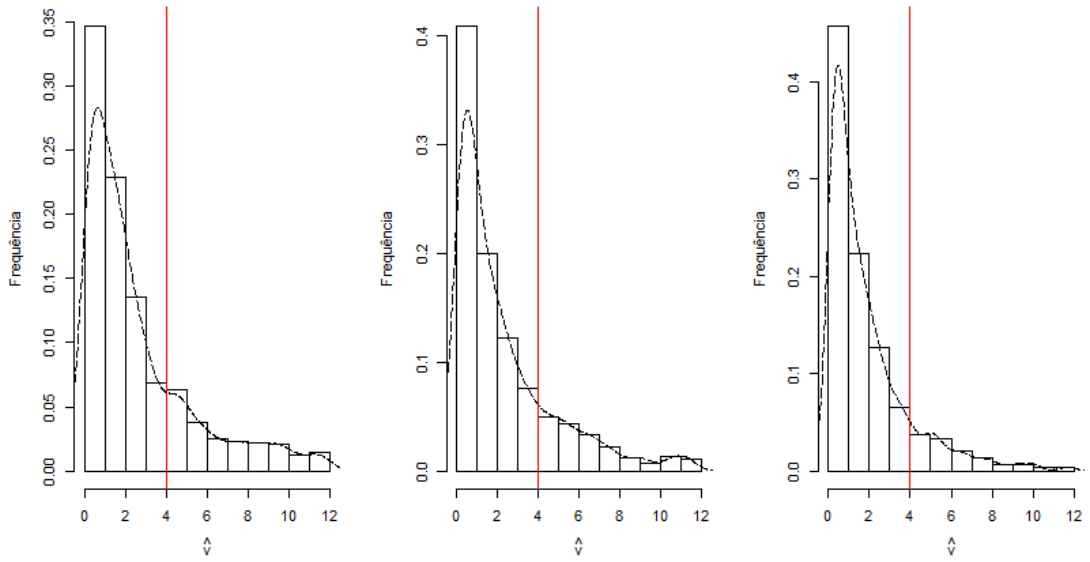

Figura 8 – Histogramas das estimativas de v para os dados gerados (da esquerda para a direita) com $n = 200, 300$ e 500.

Fonte: Elaborada pelo autor.

Os resultados da Tabela 6 mostram um alto Desvio Padrão (DP) e viés dos EMV dos parâmetros v e β_0 em relação aos outros parâmetros para $n = 200, 300$ e 500. Para os outros parâmetros os respectivos EMV apresentam um baixo viés e desvio padrão com probabilidades de coberturas convergente para grandes amostras. Os resultados da simulação confirmam o problema de identificabilidade dos parâmetros v e β_0. Fixamos o parâmetro de escala v e não o parâmetro de forma a por ser menos sensível, e obter uma comparação mais justa como o MRDWLD. A Figura 8 mostra que a distribuição do parâmetro de escala v é mais assimétrica à esquerda, quando n cresce de 200 para 500, com uma forte inconsistência do EMV do parâmetro v. A Tabela 7 e a Figura 9 apresentam os resultados da simulação do MRDWLD no mesmo cenário da simulação anterior (códigos disponíveis no Apêndice C), mas agora com $\rho = 4$.

Tabela 7 – Resumo das estimativas para os dados gerados com $\beta_0 = 2, \beta_1 = -1, \beta_2 = -1, \lambda = 1$ e $\rho = 4$.

n		λ	ρ	β_0	β_1	β_2
200	Média	1,006	4,416	2,052	-0,995	-1,020
	DP	0,166	2,083	0,302	0,268	0,434
	VIÉS	0,006	0,416	0,051	0,005	-0,020
	EQM	0,166	2,124	0,307	0,267	0,434
	PC	0,980	0,948	0,990	0,949	0,957
300	Média	1,011	4,626	2,012	-1,010	-0,988
	DP	0,140	2,129	0,265	0,210	0,388
	VIÉS	0,011	0,626	0,011	-0,010	0,012
	EQM	0,141	2,219	0,265	0,210	0,388
	PC	0,975	0,964	0,989	0,950	0,944
500	Média	1,000	4,69	2,01	-1,000	-1,000
	DP	0,111	1,996	0,215	0,164	0,295
	VIÉS	0,004	0,691	0,007	-0,002	-0,002
	EQM	0,111	2,112	0,215	0,164	0,294
	PC	0,971	0,973	0,993	0,951	0,953

Os resultados da Tabela 7 confirmam uma maior dispersão (DP) e viés do EMV do parâmetro ρ em relação aos outros parâmetros para $n = 200, 300$ e 500, mas com uma Probabilidade de Cobertura (PC) não muito distante de 95% para n=300 e n=500. Apesar desta falha, o MRDWLD é superior ao modelo de regressão BN de longa duração, devido ao terceiro nível da representação estocástica. Para os outros parâmetros, inclusive β_0, os respectivos EMV apresentam boas propriedades assintóticas. Esta inconveniência do modelo de regressão destrutivo Waring de longa duração em relação ao parâmetro ρ poderá ser evitado com a generalização da distribuição Beta (RODRÍGUEZ-AVI *et al.*, 2007), que será estudado posteriormente. Esta falha também ocorre de forma mais acentuada na distribuição WG para dados de contagem, que segundo Rodríguez-Avi *et al.* (2009) é devido a alta correlação entre os parâmetros k e ρ.

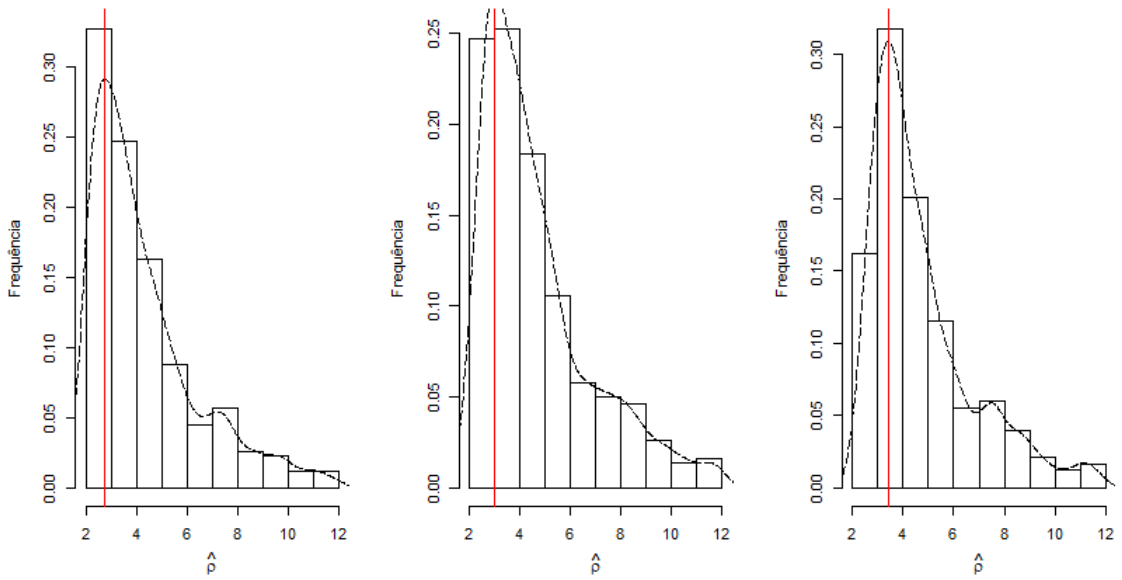

Figura 9 – Histogramas das estimativas de ρ para os dados gerados (da esquerda para a direita) com $n = 200, 300$ e 500.

Fonte: Elaborada pelo autor.

A Tabela 7 não mostra uma influência assintótica convincente do tamanho da amostra em relação ao parâmetro ρ, entretanto, a influência é evidente nos histogramas da Figura 9. A distribuição amostral do EMV é mais assimétrica à direita quando n decresce de 500 para 200, mas apresentando para $n = 300$ e $n = 500$ uma alta plausibilidade para as estimativas próximas do verdadeiro valor de ρ. Apesar do "drawback"do modelo proposto em relação ao parâmetro ρ, os demais parâmetros apresentam um baixo viés e desvio padrão assegurando um bom ajuste do MRDWLD, como foi observado na análise dos dados de HIV (ver Figura 10).

Concluindo, o MRWLD é uma excelente alternativa ao MRBNLD, após a inclusão do terceiro estágio ($p \sim \textbf{Beta}(\rho, 1)$, onde $p = \frac{1}{1+v}$) na representação estocástica da distribuição BN. Como a escolha da função de ligação é fundamental no sucesso inferencial, futuramente um estudo de simulação com a função logit aplicado a taxa de cura será realizado para comparar as vantagens e desvantagens destes dois modelos. A utilidade do modelo de regressão destrutivo Waring de longa duração é justificável, quando a fragilidade interna é separada da fragilidade externa. Se isto não ocorre, o modelo continua válido com somente uma fonte de variabilidade (sensibilidade), e comportando-se na fronteira do espaço paramétrico (RODRÍGUEZ-AVI *et al.*, 2009) como um modelo de regressão destrutivo BNeg II com longa duração.

5.2 Aplicação: Análise dos dados HIV

Considerando o conjunto de dados de sinusite (COLOSIMO; GIOLO, 2006), o modelo destrutivo Waring de longa duração consiste dos seguintes estágios:

- **Estágio de iniciação:** Seja $M_i, i = 1,...,103$ variáveis aleatórias, que representam o número de bactérias que competem entre si para a ocorrência da sinusite em 103 pacientes com HIV positivos e negativos. Além disso, vamos supor que estas variáveis discretas são independentes com distribuição Waring com parâmetros $a_i > 0$, $\rho > 2$. A taxa de cura do i-ésimo paciente é dada por

$$p_{0,i} = P[M_i = 0] = \frac{\rho}{a_i + \rho},\ \rho > 2. \tag{5.2}$$

- **Estágio de promoção:** Suponhamos que as variáveis não observáveis $Z_1,...,Z_{M_i}$ $(M_i \geq 1)$ (independentes dado M_i), que representam o tempo que cada bactéria leva para promover a sinusite (tempo de promoção da i-ésima bactéria), seguem a distribuição Weibull com taxa γ e parâmetro de forma α. A função de densidade e função de sobrevivência da distribuição Weibull são dadas por $f(t) = \alpha e^{\gamma t^{\alpha-1}} e^{-e^{\gamma t^{\alpha}}}$ e $S(t) = e^{-e^{\gamma t^{\alpha}}}$, respectivamente.

Segue de Rodrigues *et al.* (2009a) o seguinte função de sobrevivência de longa duração com a sua correspondente fdp:

$$S_p(t) = \frac{\rho}{a+\rho}\,{}_2F_1(a, 1, a+\rho+1; S(t)),$$

$$f_p(t) = f(t)\,\frac{a\rho}{(a+\rho)(a+\rho+1)}\,{}_2F_1(a+1, 2, a_i+\rho+2; S(t)).$$

Para formular o modelo de regressão destrutivo Waring de longa duração consideremos a seguinte covariada : Seja x_i uma variável que indica se o i-ésimo paciente é HIV positivo ou negativo:

$$x_i = \begin{cases} 1: \text{ HIV positivo,} \\ 0: \text{ HIV negativo.} \end{cases}$$

O modelo de regressão é finalizado com a função de ligação entre o parâmetro μ_i e a covariada x_i para o i-ésimo paciente expressa como:

$$\mu_i = \frac{a_i}{\rho - 1} = e^{\beta_0 + \beta_1 x_i} \Rightarrow a_i = (\rho - 1)\, e^{\beta_0 + \beta_1 x_i}.$$

Para os $n = 103$ pacientes em Colosimo e GIOLO (2006), a correspondente função de verosimilhança é dada por

$$
\begin{aligned}
L(\vartheta|D) &= \prod_{i=1}^{103}\left\{[f_p(t_i)]^{\delta_i}[S_p(t_i)]^{1-\delta_i}\right\} \\
&= \prod_{i=1}^{103}\left[f(t_i)\frac{a_i\rho}{(a_i+\rho)(a_i+\rho+1)}{}_2F_1(a_i+1,2,a_i+\rho+2;S(t_i))\right]^{\delta_i} \\
&\quad \left[\frac{\rho}{a_i+\rho}{}_2F_1(a_i,1,a_i+\rho+1;S(t_i))\right]^{1-\delta_i},
\end{aligned}
$$

onde $\vartheta = (\alpha,\gamma,\rho,\beta_0,\beta_1)$ e $D = \{(t_i,\delta_i), i = 1,\ldots,103\}$.

A EMV $\hat{\vartheta}$, o erro padrão (EP) e o p-valor foram obtidos numericamente pelo método L-BFGS-B (R Core Team, 2018), utilizando os códigos disponíveis no Apêndice C. Os resultados obtidos foram os seguintes:

Tabela 8 – Estimativas de máxima-verosimilhança baseadas no modelo de regressão destrutivo Waring de longa duração.

	EMV	EP	p-valor
$\hat{\alpha}$	1,354	0,486	0,005
$\hat{\gamma}$	-9,137	2,273	0,000
$\hat{\rho}$	2,261	3,497	0,518
$\hat{\beta}_0$	-0,677	1,627	0,678
$\hat{\beta}_1$	1,190	0,820	0,147

O EMV do parâmetro ρ apresenta uma alto EP em relação aos outros parâmetros. Esta propriedade indesejável foi confirmada em um estudo de simulação com dados de contagem realizado por Rodríguez-Avi *et al.* (2009).

Supondo que o tempo de promoção das bactérias para a ocorrência da sinusite segue uma distribuição de Weibull com parâmetros α e γ, a função de sobrevivência de longa duração para o paciente HIV negativo ou positivo é dada por:

$$
S_p(t) = \begin{cases} \dfrac{{}_2F_1\left(a_x,1,a_x+\rho+1;e^{-e^{\gamma}t^{\alpha}}\right)\rho}{(\rho-1)e^{\beta_0}+\rho} : \text{ se } x=0, \\ \dfrac{{}_2F_1\left(a_x,1,a_x+\rho+1;e^{-e^{\gamma}t^{\alpha}}\right)\rho}{(\rho-1)e^{\beta_0+\beta_1}+\rho} : \text{ se } x=1, \end{cases} \tag{5.3}
$$

onde $a_x = (\rho-1)e^{\beta_0+\beta_1 x}$.

A Figura 10 apresenta o gráfico da função de sobrevivência de longa duração em (5.3) para os pacientes HIV positivos e negativos, sugerindo um bom ajuste, quando comparados com os respectivos estimadores de Kaplan-Meier da Figura 2.

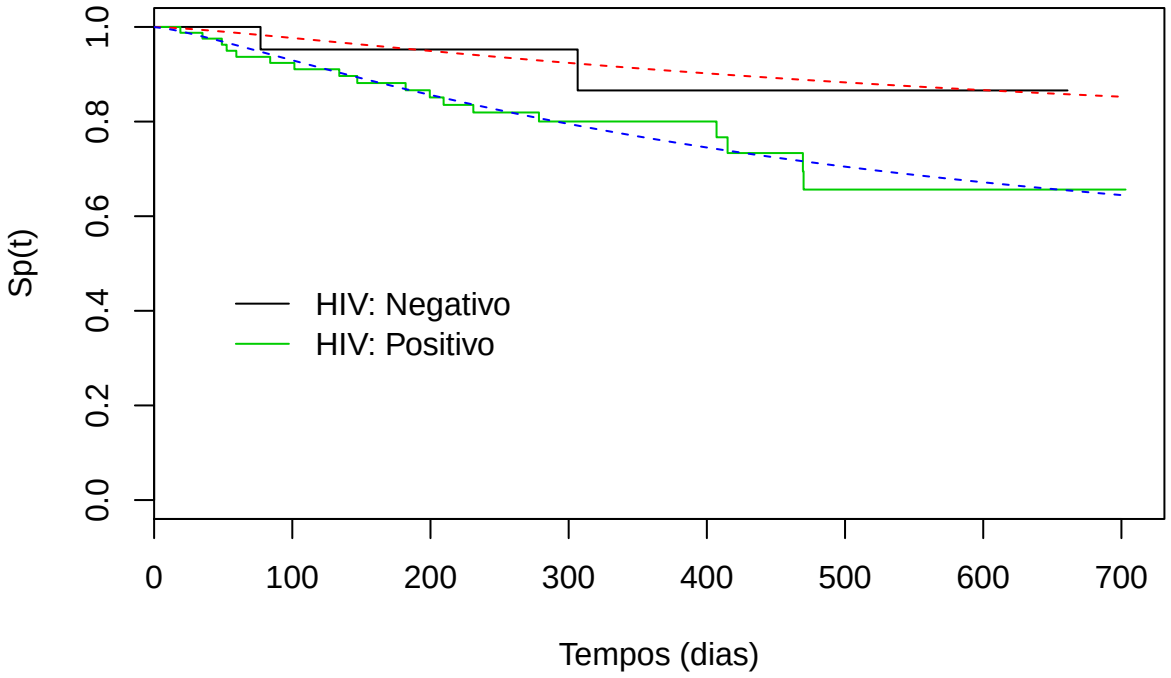

Figura 10 – Estimador de Kaplan-Meier e a função de sobrevivência baseado no MRDWLD para pacientes HIV positivos e negativos.

Fonte: Elaborada pelo autor.

As taxas de cura dos pacientes com HIV negativos e positivos são dadas por:

$$\begin{aligned} \hat{p}_1 &= 0,78 \text{ (HIV negativo)}, \\ \hat{p}_2 &= 0,52 \text{ (HIV positivo)}. \end{aligned}$$

Supondo que o número de bactérias é uma variável com distribuição Waring, as Tabela 9 e Tabela 10 mostram como a fragilidade interna é separada da externa acrescentando mais informações sobre o comportamento individual dos pacientes.

Tabela 9 – Decomposição da variância total da distribuição Waring para os pacientes HIV negativos.

Fonte de Variabilidade	**Variância**	**Taxa de variação**
Aleatório: Poisson	0,51	0,08
Fragilidade externa: Gama	3,92	0,58
Fragilidade interna: Beta	2,26	0,34
Total	6,69	1

As Figura 11 e Figura 12 apresentam os gráficos boxplot do comportamento das fontes de variabilidades do modelo de regressão destrutivo Waring para os pacientes HIV negativos $(x = 0)$ e positivos $(x = 1)$, respectivamente.

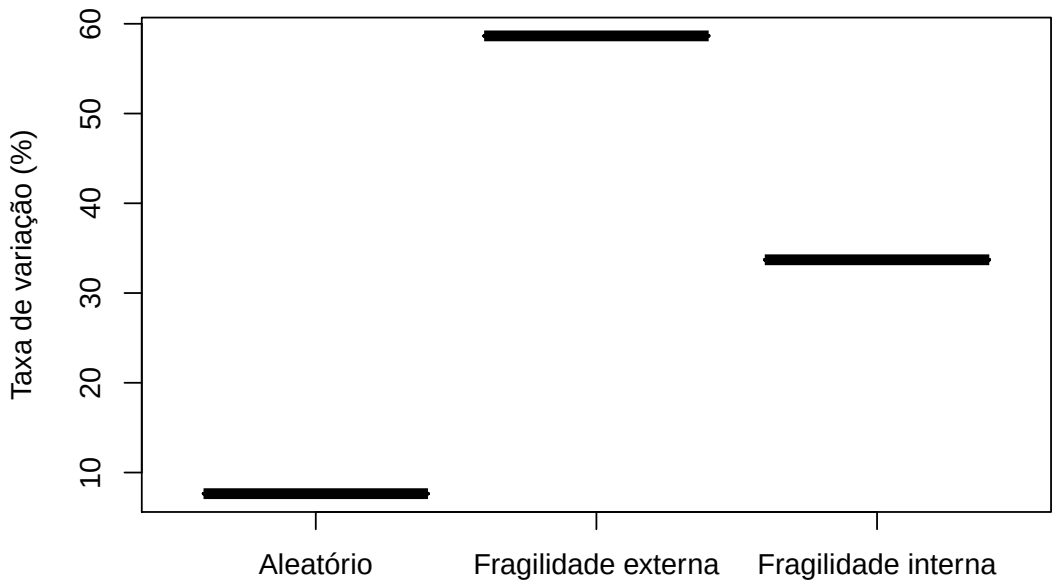

Figura 11 – Fontes de variabilidades do MRDWLD para os pacientes HIV negativos.

Fonte: Elaborada pelo autor.

Tabela 10 – Decomposição da variância total da distribuição Waring para os pacientes HIV positivos $(x = 1)$.

Fonte de Variabilidade	Variância	Taxa de variação
Aleatório: Poisson	1,67	0,04
Fragilidade externa: Gama	12,85	0,33
Fragilidade interna: Beta	24,24	0,63
Total	38,76	1

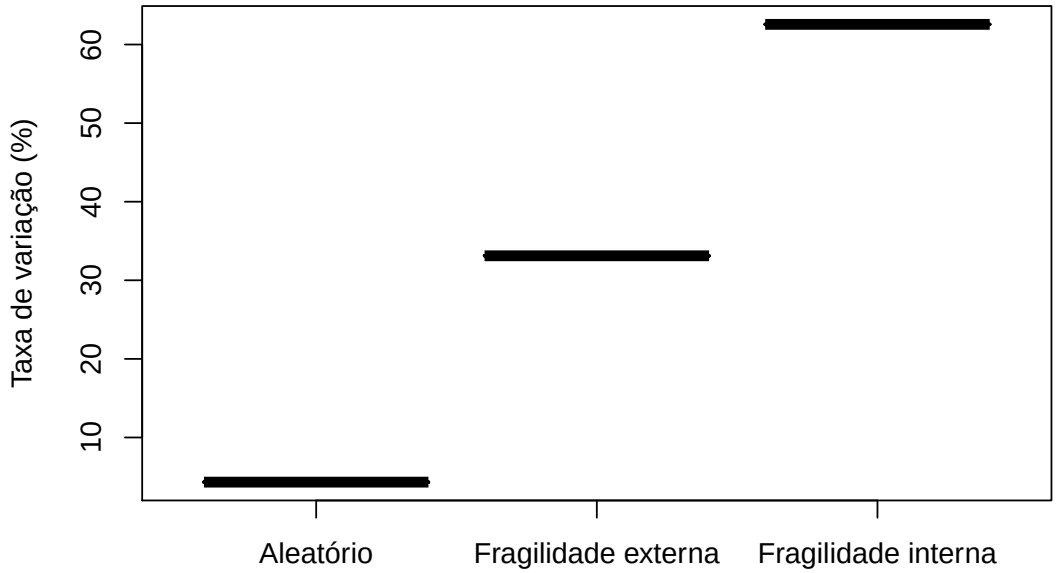

Figura 12 – Fontes de variabilidades para o MRDWLD para os pacientes HIV positivos.

Fonte: Elaborada pelo autor.

Comparando os resultados da Figura 12 com a Figura 11 verificamos que o fator HIV afeta drasticamente o sistema destrutivo aumentando a chance de ocorrência da sinusite. Neste caso específico, o tratamento deverá ser personalizado e focado no sistema destrutivo dos pacientes (imunoterapia). Este tipo de tratamento é personalizado com a finalidade de tornar o sistema destrutivo mais eficiente no reconhecimento e combate das bactérias responsáveis pela sinusite.

Baseado na Tabela 5 e com $\hat{\rho} = 2,26$, apresentamos a seguir uma comparação gráfica das fontes de variabilidades quando a média das bactérias (μ) cresce.

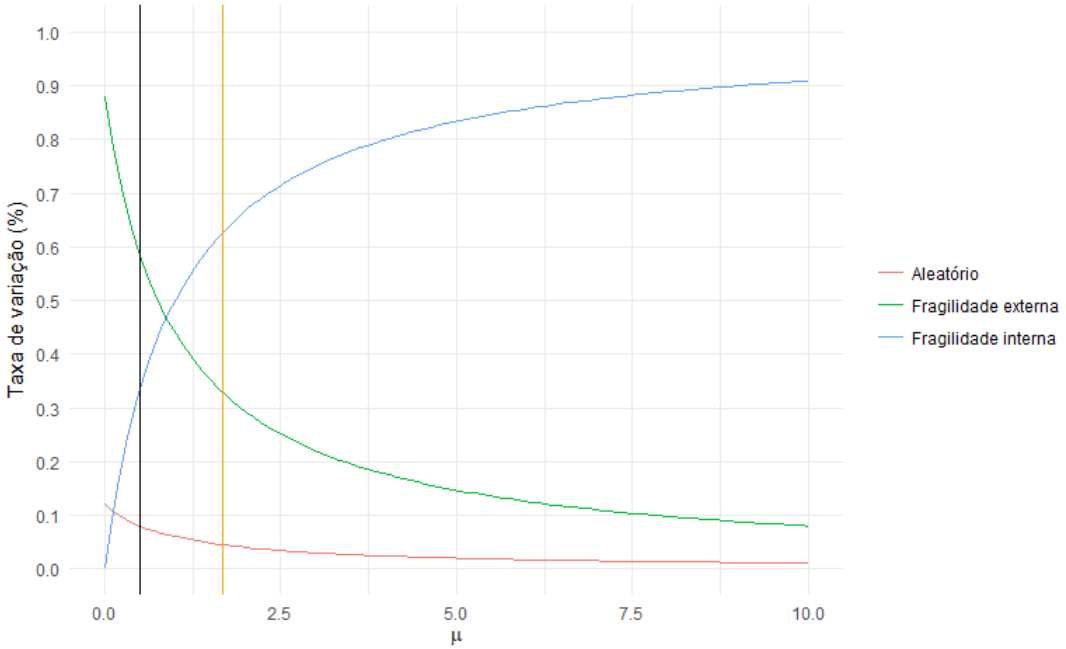

Figura 13 – Comparação das fontes de variabilidades do MRDWLD para $\hat{\rho} = 2,26$ (sem presença de covariáveis).

Fonte: Elaborada pelo autor.

A Figura 13 mostra que a medida que a média das bactérias aumenta a fonte de variabilidade devido ao mecanismo destrutivo (fragilidade interna) domina as demais fontes de variabilidades, isto é, o mecanismo destrutivo fica cada vez menos eficiente. Por outro lado, o efeito aleatório e a fragilidade externa decrescem com o aumento da média. Para valores grandes de μ, o tratamento deverá estimular o mecanismo destrutivo a reconhecer e destruir as bactérias. Acreditamos que este tipo de gráfico poderá ser útil para identificar qual fonte de dispersão é mais importante para cada paciente e propor um tratamento preventivo mais eficiente. Por exemplo, dada a estimativa $\hat{\mu}_i$ do i-ésimo paciente, podemos verificar através da Figura 13 qual será a fonte de dispersão mais importante e propor o tratamento adequado. Por exemplo, para os dados referente aos pacientes positivos e negativos temos o seguinte cenário: Para $\mu = 0,51$ (média das bactérias para os pacientes HIV negativos: eixo vertical preto) observamos na Figura 13 que a fragilidade externa é dominante em relação as fontes de variabilidades alternativas. Por outro lado, para $\mu = 1,67$ (média das bactérias para os pacientes HIV positivos: eixo vertical laranja), a fragilidade interna é a fonte de variabilidade dominante. Portanto, a presença do HIV causa um forte impacto na mecanismo defesa, isto é, a fragilidade interna é a maior responsável pela superdispersão das bactérias. O modelo proposto contesta fortemente a equidispersão das bactérias sugerida pelo modelo de regressão BN de longa duração, indicando que o mecanismo destrutivo é a principal causa da superdispersão e a necessidade de um tratamento personalizado.

A Tabela 11 apresenta os critérios de seleção AIC, BIC e MV para o **MRDWLD** e o modelo de regressão Binomial Negativo de longa duração (**MRBNLD**) :

Tabela 11 – Critérios de seleção: AIC, BIC e MV.

Critério	MRDWLD	MRBNLD
AIC	342,57	381,35
BIC	355,74	394,52
MV	-166,28	-185,67

Na Tabela 11, os critérios de seleção AIC, BIC e MV selecionam o **MRDWLD** como o mais adequado para os dados de HIV. Esta escolha é consistente com a comparação gráfica apresentada na Figura 10.

Para finalizar a análise dos dados é importante verificar se a presença do HIV tem uma influência significativa na ocorrência da sinusite. Em termos de hipótese estatística queremos verificar se a afirmação $H_0 : p_1 = p_2$ é significativa, isto é, o HIV influência a taxa de cura dos pacientes em relação a sinusite. Como $H_0 : p_1 = p_2$ é equivalente a $H_0 : \beta_1 = 0$, o procedimento de teste será baseado na normalidade assintótica marginal do emv do parâmetro β_1. Como o tamanho da amostra não é grande e com alta censura nos dados, as estimativas são tendenciosas e o p-valor não é confiável devido a normalidade assintótica insatisfatória. Para uma melhor aproximação normal, vamos utilizar a transformação paramétrica $\psi = \beta_1^{-1/3}$ proposta por Sprott e Kalbfleisch (1969). Utilizando o método delta, e a matriz de covariância observada de Fisher obtida numéricamente via R-Studio, temos o seguinte intervalo de confiança para ψ com nível de confiança de 95%:

$$\left(\hat{\psi} - 1.96\sqrt{\frac{(\hat{\psi})^8 \hat{Var}(\hat{\psi})}{9}}, \hat{\psi} + 1.96\sqrt{\frac{(\hat{\psi})^8 \hat{Var}(\hat{\psi})}{9}} \right), \tag{5.4}$$

onde $\hat{Var}(\hat{\psi})$ é a estimativa da variância assintótica do mle $\hat{\psi} = (\hat{\beta}_1)^{-1/3}$. Utilizando os resultados anteriores, o intervalo de confiança de 95% para β_1 é dado por $(0,22, 2,57)$. Este intervalo sugere, com nível de confiança de 95%, que as taxas de cura são significativamente diferentes. O tradicional p-valor não é uma medida de evidência apropriada para os dados HIV, devido a alta censura e amostra pequena, e com uma decisão totalmente fora da realidade, isto é, o HIV não tem influência na destruição das bactérias responsáveis pela sinusite. A transformação paramétrica e o método delta foram essenciais para obter um resultado consistente com dados e com as informações disponíveis sobre a influência do HIV no sistema imune do paciente.

5.3 Aplicação: Análise dos dados de melanoma

Nesta seção, a metodologia inferencial apresentada anteriormente será utilizada para analisar os dados de 205 pacientes, após a operação de remoção do tumor maligno (melanoma) durante o período de 1962 à 1977. Os dados estão disponíveis no pacote timereg em R Core Team (2018). O tempo observado (T) varia de 10 a 5565 dias (de 0,0274 a 15,25 anos, com média = 5,9 e desvio padrão = 3,1 anos), e refere-se ao tempo até a morte do paciente ou o tempo de censura. Pacientes mortos por outras causas, bem como pacientes ainda vivos no final do estudo, são observações censuradas (72%). Escolhemos como covariadas o status de ulceração (ausente, n = 115; presente, n = 90), a espessura do tumor (em mm, média = 2,92 e desvio padrão = 2,96) e o sexo do paciente (0: feminino, 1: masculino).

Para a análise estatística dos dados propomos o MRDWLD com decomposição da variância da mesma forma como foi utilizada no exemplo da sinusite, utilizando a mesma função de sobrevivência de longa duração e a correspondente função de densidade imprópria (RODRIGUES *et al.*, 2009a).

Finalmente, para formular o MRDWLD, a função de ligação associada ao i-ésimo paciente é definida como

$$\mu_i = \frac{a_i}{\rho - 1} = e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\beta_3 x_{3,i}} \Rightarrow a_i = (\rho - 1)\, e^{\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\beta_3 x_{3,i}},$$

com a seguinte covariada:

$$x_{1,i} = \begin{cases} 1: \text{ presente} \\ 0: \text{ ausente} \end{cases},$$

$$x_{2,i} = \text{Espessura do tumor},$$

$$x_{3,i} = \begin{cases} 1: \text{ masculino} \\ 0: \text{ feminino} \end{cases}.$$

Para os dados, $D = \{(t_i, \delta_i), i = 1, \ldots, 205\}$, utilizamos a mesma função de verosimilhança apresentada no exemplo de sinusite, $L(\vartheta|D)$, onde $\vartheta = (\alpha, \gamma, \rho, \beta_0, \beta_1, \beta_2, \beta_3)$.

A EMV $\hat{\vartheta}$, o erro padrão (EP) e o p-valor foram obtidos numericamente pelo método L-BFGS-B (R Core Team, 2018), utilizando os códigos disponíveis no Apêndice C. Os resultados obtidos foram os seguintes:

Tabela 12 – Estimativas de máxima verosimilhança baseado no MRDWLD.

	EMV	EP	p-valor
$\hat{\alpha}$	1,992	0,259	0,000
$\hat{\gamma}$	-3,733	0,490	0,000
$\hat{\rho}$	7,000	12,057	0,562
$\hat{\beta}_0$	-1,895	0,479	0,000
$\hat{\beta}_1$	1,475	0,363	0,000
$\hat{\beta}_2$	1,708	0,528	0,001
$\hat{\beta}_3$	0,668	0,339	0,049

Na Tabela 12 apresenta um alto EP do EMV do parâmetro ρ em relação aos outros parâmetros do MRDWLD, como nos dados sobre HIV. Observe que esta desvantagem do modelo proposto e as propriedades assintóticas do EMV foram avaliadas via Monte Carlo. A última coluna da Tabela 12 mostra que os coeficientes de regressão são significativos com impacto das covariadas consideradas no modelo, isto é, a presença e o tamanho do tumor são as covariadas mais influentes.

A seguir apresentamos o gráfico boxplot (Figura 14) das fontes de variabilidades do MRDWLD o para os pacientes com melanoma.

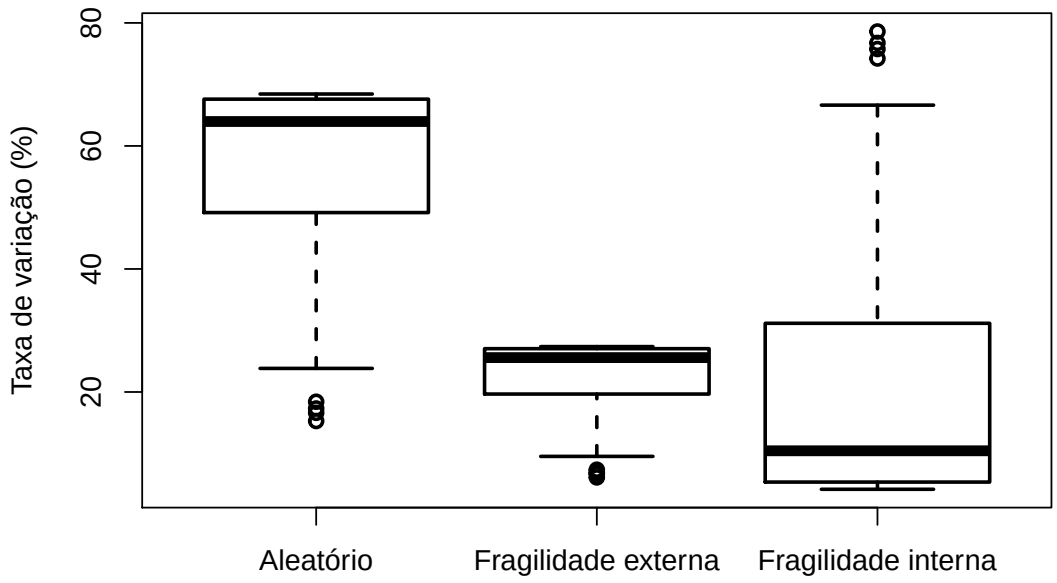

Figura 14 – Fontes de variabilidades baseado no MRDWLD para os dados de melanoma.

Fonte: Elaborada pelo autor.

Na Figura 14 observamos que a fonte aleatória domina a fragilidade interna e externa. A fragilidade interna sugere um fraco impacto do sistema destrutivo sobre a superdispersão das células malignas com acentuada tendência à equidispersão. Como $\hat{\rho} = 7$ é relativamente grande, a distribuição Waring aproxima-se da distribuição geométrica (BNeg II), isto é, $M \sim BN(1, \frac{1}{1+\mu_x})$ (RODRÍGUEZ-AVI *et al.*, 2009) com uma fonte variabilidade (sensibilidade), onde predomina a fragilidade externa sobre a interna. O modelo de regressão destrutivo geométrico em Rodrigues *et al.* (2011) e Rodrigues *et al.* (2012) não separa a fragilidade interna da externa, com um mecanismo destrutivo totalmente diferente baseado na distribuição binomial. A convergência da distribuição Waring para a distribuição geométrica, a decomposição da sensibilidade em fragilidade interna e externa e o protagonismo do paciente no tratamento são as principais vantagens do modelo de regressão destrutivo Waring sobre o modelo de regressão destrutivo geométrico.

A Figura 15, com $\hat{\rho} = 7$, apresenta a relação gráfica entre as fontes de variabilidades em função da média das células danificadas (μ). Este gráfico é útil para verificar para cada paciente a fonte de variabilidade dominante na superdispersão das células danificadas, e formular um tratamento preventivo mais adequado. Esta seria uma outra vantagem importante do MRDWLD, além de uma taxa de cura personalizada.

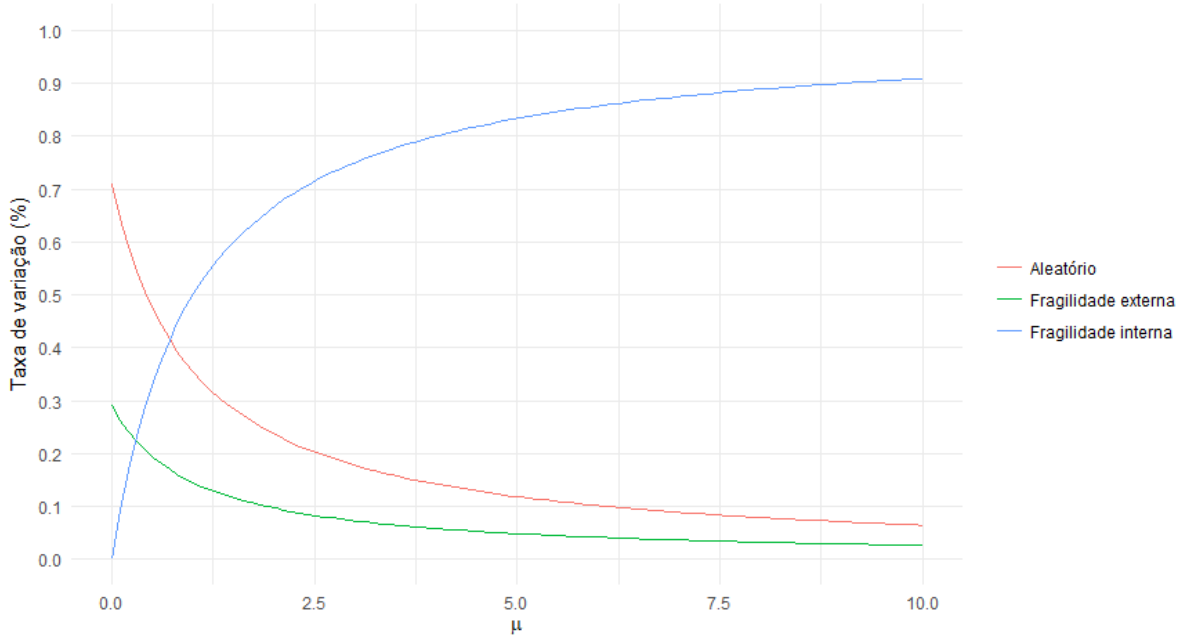

Figura 15 – Comparação das fontes de variabilidades do MRDWLD para $\hat{\rho} = 7$ (sem presença de covariáveis).

Fonte: Elaborada pelo autor.

A Tabela 13 apresenta a influência das covariadas sexo, ulceração e espessura do tumor na taxa de cura e as fragilidades internas de oito pacientes.

Tabela 13 – Taxa de cura e a fragilidade interna (FI) de oito pacientes com espessura de tumor mínimo e máximo.

Sexo	**Ulceração**	**Espessura do Tumor (mm)**	**Paciente**	p_0	**FI**
Masculino	**Presente**	0,081	**A**	0,44	0,60
		1,288	**B**	0,09	0,92
	Ausente	0,016	**C**	0,79	0,23
		1,466	**D**	0,24	0,78
Feminino	**Presente**	0,016	**E**	0,63	0,40
		1,742	**F**	0,08	0,93
	Ausente	0,010	**G**	0,88	0,13
		1,288	**H**	0,46	0,58

As colunas referentes a taxa de cura e a fragilidade interna mostram, que independentemente do sexo, a presença da ulceração com espessura máxima do tumor tem uma alta fragilidade interna e baixa taxa de cura, devido a ineficiência do mecanismo destrutivo. Para os dois pacientes, com menor taxa de cura e maior fragilidade interna (pacientes F e B), é recomendável um tratamento especifico para melhorar a eficiência do mecanismo destrutivo ou sistema imune.

CAPÍTULO

6

CONCLUSÕES FINAIS E PESQUISAS FUTURAS

Neste capítulo apresentamos as considerações mais importantes que foram discutidas durante o desenvolvimento deste trabalho, e algumas linhas de pesquisas que pretendemos desenvolver no futuro.

6.1 Conclusões finais

O número de fatores de riscos M é um efeito aleatório latente (fragilidade discreta), que expressa o comportamento heterogêneo dos pacientes em relação ao risco básico da população. Este comportamento está diretamente conectado ao fenômeno de superdispersão dos fatores de riscos e o mecanismo destrutivo. Na literatura sobre os modelos de longa duração em dois estágios, várias distribuições discretas com caudas pesadas ("J-shaped") têm sido utilizadas para explicar um possível excesso de variabilidade em relação a $E[M]$. Entretanto, este modelos não conseguem separar o efeito individual ou destrutivo (fragilidade interna) dos efeitos externos ou covariadas desconhecidas (fragilidade externa), que interferem na superdispersão do fatores de riscos e no mecanismo destrutivo dos pacientes. Motivado pela teoria de acidentes sugerimos a distribuição WG para o modelo de longa duração em dois estágios, devido a existência de um mecanismo inteligente (flexível) e personalizado. Esta distribuição envolve os parâmetros (ρ, k) da distribuição Beta tal que $E(p) = \frac{\rho}{\rho+k}$, onde ρ é o índice de precisão em relação a $E[p]$ e k o índice de precisão em relação a $1 - E[p]$. Em outras palavras, a distribuição WG é uma mistura da distribuição BN pela distribuição Beta, que representa o mecanismo destrutivo dos fatores de riscos, e interpreta a superdispersão através de três fontes de variabilidades:

- **Aleatória (equidispersão)**: Representada pela distribuição Poisson com parâmetro λ, supondo que o comportamento dos fatores de riscos é o mesmo (equidispersão) para todos

os pacientes da população.

- **Fragilidade externa (covariada desconhecida)**: Representado pela distribuição Gama dado (a_x, v). A mistura da distribuição Poisson com a distribuição Gama gera uma distribuição BN com parâmetros (a_x, p), onde $p = \frac{1}{1+v}$ e v o parâmetro de superdispersão.

- **Fragilidade interna (mecanismo destrutivo):** Depende da distribuição Beta que flexibiliza o mecanismo da BN adicionando os parâmetro k e ρ, que introduzem a precisão do mecanismo destrutivo definido como $\phi = \rho + k$.

Para avaliar a sobrevivência dos pacientes e formular planejamentos preventivos é fundamental estudar o impacto das fontes de variabilidades na superdispersão dos fatores de riscos. A destruição ou fragilidade interna é uma fonte personalizada de variabilidade e a mais importante, que não está presente nos modelos alternativos mais recentes em análise de sobrevivência de longa duração. Por simplicidade computacional utilizamos nas aplicações uma versão mais simples da WG $(k = 1)$, conhecida na literatura como distribuição Waring.

No estudo de simulação verificamos o alto desvio e vício do EMV do parâmetro ρ e a fraca consistência assintótica, mas com excelentes estimativas para os demais parâmetros, inclusive o coeficiente β_0. Lembramos que esta falha do modelo é mais explícita na distribuição WG para dados discretos (RODRÍGUEZ-AVI *et al.*, 2009).

Resumindo, o foco deste trabalho consistiu em formular um modelo de regressão destrutivo de longa duração em dois estágios com um mecanismo destrutivo interno dos fatores riscos competitivos e a decomposição da variância. A decomposição da variância é útil para analisar o impacto das fontes de variabilidades na superdispersão ou fragilidade dos pacientes. Utilizando a representação estocástica da distribuição Waring identificamos dois tipos de fragilidades não encontrados nos modelos de longa duração mais recentes: fragilidade externa e interna. Com este modelo de regressão destrutivo, dado (ρ, β), é possível através de μ identificar a fonte de variabilidade responsável pela superdispersão de um futuro paciente, e aplicar o tratamento adequado. As aplicações com dados HIV e melanoma mostraram como o modelo proposto pode explicar as características individuais mais importantes da doença, não focalizadas nos modelos de longa duração existentes na literatura. Um artigo com os principais resultados da dissertação foi elaborado para ser submetido em um periódico nacional ou internacional após a revisão do Prof. N. Balakrishnan-Canadá.

6.2 Pesquisas Futuras

- Para evitar o uso das propriedades assintóticas do EMV formular modelos bayesianos de regressão destrutivos WG de longa duração (k: desconhecido).

- Estudar o problema de identificabilidade do parâmetro ρ no contexto bayesiano e um estudo comparativo com o EMV baseado no modelo de regressão destrutivo Waring de longa duração ($k = 1$).
- Estudo de simulação com a função logit aplicado a taxa de cura para comparar as vantagens e desvantagens dos modelos destrutivos BN e Waring de longa duração.
- Formular um modelo mais flexível que o MDRWLD através da generalização da distribuição Beta (RODRÍGUEZ-AVI *et al.*, 2007).

REFERÊNCIAS

AJIFERUKE, I. A probabilistic model for the distribution of authorships. **Journal of the American Society for Information Science**, Wiley Online Library, v. 42, n. 4, p. 279–289, 1991. Citado na página 77.

BATES, G. E.; NEYMAN, J. **Contributions to the theory of accident proneness. 1. an optimistic model of the correlation between light and severe accidents**. [S.l.], 1952. Citado nas páginas 24 e 72.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor & Francis, v. 47, n. 259, p. 501–515, 1952. Citado na página 28.

CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. A new Bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Taylor & Francis, v. 94, n. 447, p. 909–919, 1999. Citado na página 23.

COLIN, A. C.; TRIVEDI, P. K. Econometric models based on count data: comparisons and applications of some estimators and test. **Journal of Applied Econometrics**, v. 1, n. Pt. 1, p. 29–53, 1986. Citado na página 47.

____. **Regression Analysis of Count Data, Econometric Society Monograph No. 30**. [S.l.]: Cambridge: Cambridge University Press, 1998. Citado na página 73.

COLOSIMO, E.; GIOLO, S. Análise de sobrevivência aplicada. 1ª edição. **São Paulo: Editora Edgard Blücher**, 2006. Citado nas páginas 32, 33, 34, 47 e 51.

COX, D. W. Regression models and life tables (with discussion). **J Royal Stat Soc Ser B**, v. 34, p. 187–220, 1972. Citado na página 28.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of applied statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citado na página 40.

GREENWOOD, M.; YULE, G. An enquiry into the nature of frequency distributions representative of multiple happenings, with special reference to multiple attacks of disease or repeated accidents. **Journal of the Royal Statistical Society**, v. 83, p. 255–279, 1920. Citado nas páginas 24 e 71.

IRWIN, J. O. The generalized Waring distribution applied to accident theory. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 131, n. 2, p. 205–225, 1968. Citado nas páginas 24, 40, 71, 75, 77, 80 e 81.

____. The generalized Waring distribution. part i. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 138, n. 1, p. 18–31, 1975. Citado nas páginas 23 e 43.

IRWIN, J. O. *et al.* The place of mathematics in medical and biological statistics. **Journal of the Royal Statistical Society**, v. 126, n. Pt. 1, p. 1–41, 1963. Citado nas páginas 41, 42, 74 e 76.

LEVENE, M.; FENNER, T.; LOIZOU, G.; WHEELDON, R. A stochastic model for the evolution of the web. **Computer Networks**, Elsevier, v. 39, n. 3, p. 277–287, 2002. Citado na página 77.

NEWBOLD, E. M. Practical applications of the statistics of repeated events' particularly to industrial accidents. **Journal of the Royal Statistical Society**, JSTOR, v. 90, n. 3, p. 487–547, 1927. Citado na página 80.

PENG, Y.; LORD, D.; ZOU, Y. Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis. **Accident Analysis & Prevention**, Elsevier, v. 73, p. 20–26, 2014. Citado nas páginas 24, 79 e 80.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado nas páginas 35, 47, 52 e 58.

RODRIGUES, J.; CANCHO, V.; CASTRO, M. D. Teoria unificada de análise de sobrevivência. **Associaçao Brasileira de Estatıstica**, v. 18, 2008. Citado nas páginas 27 e 29.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de; LOUZADA-NETO, F. On the unification of long-term survival models. **Statistics & Probability Letters**, Elsevier, v. 79, n. 6, p. 753–759, 2009. Citado nas páginas 27, 28, 29, 31, 34, 51 e 58.

RODRIGUES, J.; CASTRO, M.; CANCHO, V. G.; BALAKRISHNAN, N. COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. **Journal Statistical Planning Inference**, v. 139, p. 3605–3611, 2009. Citado na página 23.

_____. A Bayesian destructive weighted Poisson cure rate model and an application to a cutaneous melanoma data. **Statistical Methods in Medical Research**, v. 121, n. 16, p. 585–597, 2012. Citado na página 60.

RODRIGUES, J.; CASTRO, M. de; BALAKRISHNAN, N.; CANCHO, V. G. Destructive weighted Poisson cure rate models. **Lifetime Data Analysis**, Springer, v. 17, n. 3, p. 333–346, 2011. Citado na página 60.

RODRÍGUEZ-AVI, J.; CONDE-SÁNCHEZ, A.; SÁEZ-CASTILLO, A.; OLMO-JIMÉNEZ, M. A new generalization of the Waring distribution. **Computational Statistics & Data Analysis**, Elsevier, v. 51, n. 12, p. 6138–6150, 2007. Citado nas páginas 42, 49, 65, 72, 74, 76, 77, 78 e 79.

RODRÍGUEZ-AVI, J.; CONDE-SÁNCHEZ, A.; SÁEZ-CASTILLO, A.; OLMO-JIMÉNEZ, M.; MARTÍNEZ-RODRÍGUEZ, A. M. A Generalized Waring regression model for count data. **Computational Statistics & Data Analysis**, Elsevier, v. 53, n. 10, p. 3717–3725, 2009. Citado nas páginas 40, 41, 43, 45, 49, 50, 52, 60, 64, 74 e 78.

SPROTT, D.; KALBFLEISCH, J. D. Examples of likelihoods and comparison with point estimates and large sample approximations. **Journal of the American Statistical Association**, Taylor & Francis, v. 64, n. 326, p. 468–484, 1969. Citado na página 57.

TEŠÍTELOVÁ, M. On the role of nouns in lexical statistics. **Prague Studies in**, 1967. Citado na página 77.

WIENKE, A. **Frailty models in survival analysis**. [S.l.]: Chapman and Hall/CRC, 2010. Citado na página 30.

XEKALAKI, E. The univariate generalized Waring distribution in relation to accident theory: proneness, spells or contagion? **Biometrics**, JSTOR, p. 887–895, 1983. Citado na página 75.

____. On the distribution theory of over-dispersion. **Journal of Statistical Distributions and Applications**, Springer, v. 1, n. 1, p. 19, 2014. Citado na página 31.

YAKOVLEV, A. Y.; TSODIKOV, A. D.; BASS, L. A stochastic model of hormesis. **Mathematical Biosciences**, Elsevier, v. 116, n. 2, p. 197–219, 1993. Citado na página 23.

YULE, G. U. An introduction to the theory of statistics. **Bulletin of the American Mathematical Society**, v. 30, p. 465–466, 1924. Citado nas páginas 41, 44 e 76.

APÊNDICE

A

TEORIA DE ACIDENTES

A.1 Introdução

No Apêndice A apresentamos um resumo da teoria de acidentes que será crucial na formulação de um novo modelo de longa duração em dois estágios, onde a fragilidade do paciente esta diretamente associada ao fenômeno da superdispersão dos fatores de riscos responsáveis pela sobrevivência do paciente. A fragilidade na análise de sobrevivência tradicional é uma mistura do efeito ambiental (externo) com o efeito individual (interno), conhecida na teoria de acidentes como sensibilidade, para avaliar os riscos dos pacientes em relação a um risco básico na população. A escolha de um modelo de superdispersão para os fatores de riscos, que identifique separadamente o impacto dos efeitos na fragilidade ou sensibilidade, é realizada com sucesso na teoria de acidentes através da distribuição Waring Generalizada (WG) introduzida por Irwin (1968).

O objetivo da teoria de acidentes (GREENWOOD; YULE, 1920; IRWIN, 1968) é propor modelos probabilísticos para o número de acidentes com a finalidade de investigar as fontes de variabilidades aleatória e não aleatória responsáveis pela superdispersão dos dados.

Para entender a teoria de acidentes, o primeiro passo é deixar claro o significado da palavra chave "acidente". Acidente é um evento inesperado e indesejável que causa danos pessoais, materiais (danos ao patrimônio), danos financeiros e que ocorre de modo não intencional. Exemplos incluem colisões no tráfego, lesões na fábrica, ocorrência de uma bactéria de sinusite ou uma célula cancerígena, a morte de uma paciente, esquecer um compromisso, etc. A teoria de acidentes supõe que o número de acidentes (por exemplo, acidentes de tráfego em diferentes locais) segue uma distribuição de Poisson com o mesmo nível de predisposição (fator interno ou "proneness") e submissão (fator externo ou "liability") a um determinado acidente. Supondo que o número médio de acidentes segue uma distribuição gama, o número de acidentes é uma variável discreta com distribuição BN, onde sua variância expressa duas fontes de dispersão:

aleatorização e a sensibilidade ao acidente que consiste no efeito conjunto da predisposição e submissão. Por outro lado, não é realístico supor que os acidentes ocorrem devido aos mesmo fatores externos (submissão), por exemplo, o comportamento e a distração dos motoristas, tempo e condições de visibilidade. Com uma ilustração na área do esporte segue o seguinte exemplo (RODRÍGUEZ-AVI *et al.*, 2007)

Exemplo: Seja M o número de cartões amarelos apresentados a um determinado jogador de futebol (excluindo os goleiros) que participa de pelo menos um jogo de futebol. Esta variável está sujeita a 3 fontes de variabilidades:

- O número de jogos participados (aleatorização);
- A posição do jogador no campo: um defensor geralmente corre mais risco de ser penalizado do que um meio-campista, que por sua vez tem um risco maior do que um atacante (submissão ou fator externo);
- A personalidade do jogador de futebol fica evidente durante o jogo (predisposição ou fator interno).

A.2 Modelo regressão Binomial Negativa (BN)

Antes de formular o modelo de regressão Binomial Negativa vamos introduzir a distribuição Binomial Negativa através da representação estocástica apresentada em Bates e Neyman (1952).

- Distribuição Binomial Negativa e a decomposição da variância:

  Suponhamos que o número de acidentes seja uma variável aleatória de Poisson com parâmetro λ, onde a equidispersão é representada por

  $$Var(M) = E(M) = \lambda. \tag{A.1}$$

  Como já foi mencionado, se os acidentes apresentam superdispersão, uma forma de explicar este excesso de variabilidade é supor que $\lambda \sim Gama(a,v)$, isto é,

  $$f(\lambda;a,v) = \frac{1}{v^a\Gamma(a)}\lambda^{a-1}e^{-\frac{\lambda}{v}}, \lambda > 0, a, v > 0. \tag{A.2}$$

  Teorema A.1. De (A.1) e (A.2) a variável M tem uma distribuição binomial negativa, $M \sim BN(a,p)$, com função de probabilidade de massa (f.p.m.) dada por:

  $$f(m|a,p) = \frac{\Gamma(a+m)}{\Gamma(a)\,m!}p^a(1-p)^m, \quad m = 0,1,2,\ldots,\ a > 0, \tag{A.3}$$

  onde $p = (1+v)^{-1}$ e $v > 0$.

A justificativa de (A.3) é apresentada no Apêndice B.

De (A.3), segue a seguinte interpretação para a distribuição BN: 0 parâmetro a (supondo inteiro e positivo) pode ser interpretado como número de acidentes que foram evitados com probabilidade p e $M = m$ o número de acidentes não evitados. O efeito interno, p, e o efeito externo ou ambiental estão agindo simultaneamente na superdispersão do número M de acidentes. O processo BN termina quando a acidentes forem evitados, onde os acidentes não evitados $M = m$ estão sobre o impacto simultâneo dos efeitos individuais e ambientais. Para o modelo BN o número médio de acidentes não evitados é dado por

$$E[M] = E[E[M|\lambda]] = E[\lambda] = \frac{a}{\frac{1}{v}} = av = a(\frac{1-p}{p}) = \mu.$$

Como em Colin e Trivedi (1998), a variância total pode ser decomposta nas seguintes fontes de variabilidades:

$$\begin{aligned} Var[M] &= E[Var[M|\lambda]] + Var[E[M|\lambda]] \\ &= E[\lambda] + Var[\lambda] \\ &= \mu + av^2 \\ &= \mu + \frac{1}{a}\mu^2 \\ &= \mu(1 + \frac{\mu}{a}) \\ &= \mu(1+\theta), \end{aligned}$$

onde $\theta = \frac{\mu}{a}$ é o parâmetro de superdispersão e a o parâmetro de precisão. O primeiro termo μ representa à aleatorização e o segundo $\mu + av^2$ a sensibilidade.

Tabela 14 – Decomposição da variância total do modelo Binomial-Negativo.

Fonte de Variabilidade	**Variância**	**Taxa de Variação**
Aleatorização	μ	$\frac{a}{a+\mu}$
Sensibilidade	$\frac{1}{a}\mu^2$	$\frac{\mu}{a+\mu}$
Total	$\mu + \frac{1}{a}\mu^2$	1

- Função geradora de probabilidades da distribuição BN é dada por:

$$A_M(s) = \left[\frac{p}{1-(1-p)s}\right]^a. \tag{A.4}$$

A justificativa da equação (A.4) é apresentada no Apêndice B.

- Modelo de regressão Binomial Negativo:

  O modelo de regressão BN foi introduzido por Rodríguez-Avi *et al.* (2009) utilizando a seguinte função de ligação:

$$\log(\mu) = \beta_0 + x^T\beta,$$

  onde $\beta^T = (\beta_1, \ldots, \beta_p)$ e $x^T = (x_1, \ldots, x_p)$ é o vetor de covariadas.

  A ligação entre o parâmetro p_i e a covariada x_i do i-ésimo acidente é obtida como:

$$\mu_i = av_i = e^{\beta_0+\beta x_i^T} \Rightarrow v_i = \frac{e^{\beta_0+\beta x_i^T}}{a},$$

  logo segue a função de ligação:

$$p_i = (1+v_i)^{-1} \Rightarrow p_i = \left[1 + \frac{e^{\beta_0+\beta x_i^T}}{a}\right]^{-1},$$

  onde $i = 1, ..., n$.

A.3 Modelo de regressão Waring Generalizada (WG) e decomposição da variância

Nesta seção vamos definir a distribuição WG e as suas respectivas propriedades.

- Distribuição Waring generalizada :

  A distribuição Waring (RODRÍGUEZ-AVI *et al.*, 2007) é obtida da série Waring definida como

$$\frac{1}{x-a} = \sum_{m=0}^{\infty} \frac{(a)_m}{(x)_{m+1}}, \quad m = 0, 1, \ldots, \tag{A.5}$$

  onde $(\alpha)_k = \alpha(\alpha+1)\ldots(\alpha+k-1)$; se $\alpha > 0$ segue que $(\alpha)_k = \Gamma(\alpha+k)/\Gamma(\alpha)$. Se $\rho = x-a$, a função de probabilidade (f.p.) segue diretamente de (A.5) :

$$p_m = PM = m\rho \frac{(a)_m}{(a+\rho)_{m+1}}, \quad m = 0, 1, \ldots, \tag{A.6}$$

  onde $a, \rho > 0$. A série (A.5) pode ser generalizada da seguinte forma

$$\frac{1}{(x-a)_k} = \sum_{m=0}^{\infty} \frac{(a)_m (k)_m}{(x)_{m+k}} \frac{1}{m!}, \quad m = 0, 1, \ldots$$

  com $k > 0$ (IRWIN *et al.*, 1963).

Teorema A.2. A distribuição de WG é obtida de (A.6) com $\rho = x - a$, e expressa como

$$p_m = \frac{(\rho)_k}{(a+\rho)_k} \frac{(a)_m (k)_m}{(a+k+\rho)_m} \frac{1}{m!}, \quad m = 0, 1, \ldots. \tag{A.7}$$

A justificativa da equação (A.7) é apresentada no Apêndice B.

- Representação hierárquica da distribuição WG:

  A seguir apresentamos a representação estocástica da distribuição WG que será útil para obter a média, que representará a variabilidade devido a aleatorização, e a variância total que será fundamental para identificar as fontes de variabilidades devido a predisposição e submissão (IRWIN, 1968). Como foi comentado anteriormente, a distribuição BN considera a sensibilidade como um tipo superdispersão que não separa o efeito ambiental ou externo (submissão) do efeito individual ou interno (predisposição). Estas fontes são essenciais na previsão ou prevenção de acidentes. Para obter a separação das fontes de variabilidades internas e externas, Irwin (1968) introduziu o seguinte modelo hierárquico em três estágios:

  - $M|\lambda \sim \text{Poisson}(\lambda)$,
  - $\lambda|a, v \sim \text{Gama}(a, v)$, função de densidade de probabilidade (f.p.) dada por

    $$\frac{\lambda^{a-1} e^{-\lambda/v}}{\Gamma(a)\, v^a},$$

  - $p|\rho, k \sim \text{Beta}(\rho, k)$, fdp é dada por

    $$\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)} p^{\rho-1} (1-p)^{k-1},$$

    onde $p = \frac{1}{1+v}$.

Teorema A.3. (IRWIN, 1968; XEKALAKI, 1983) Sob a representação hierárquica em três estágios, a variável discreta M tem uma distribuição WG , representada por $M \sim WG(a, \rho, k)$, e dada por

$$p_m = P[M = m] = \frac{(\rho)_k}{(a+\rho)_k} \frac{(a)_m (k)_m}{(a+k+\rho)_m} \frac{1}{m!}, \quad m = 0, 1, \ldots \tag{A.8}$$

A prova de (A.8) pode ser encontrada no Apêndice B.

Como ilustração apresentamos o gráfico da representação hierárquica ou estocástica da distribuição WG, para $n = 2$.

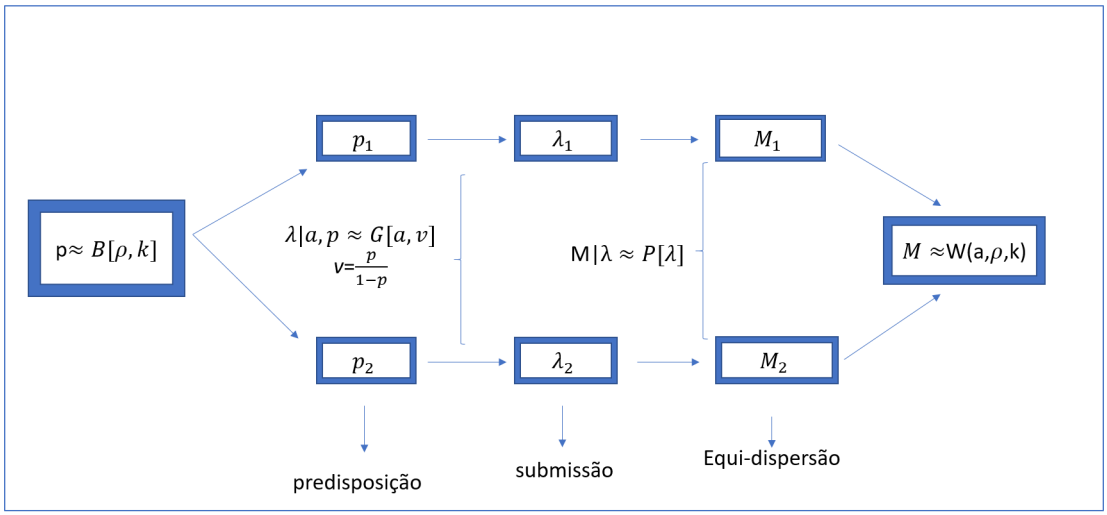

Figura 16 – Representação hierárquica ou estocástica da distribuição WG, para $n = 2$

Fonte: Elaborada pelo autor.

O par (ρ, k) caracteriza o impacto do efeito individual na dispersão dos M acidentes não evitados (separado do efeito ambiental). Do ponto de vista individual, dado p, temos uma forte correlação negativa entre ρ e k. As seguintes restrições paramétricas implicam em distribuições WG mais simples:

- $k = 1$, modelo Waring (IRWIN *et al.*, 1963);
- $a = k = 1$, modelo Yule (YULE, 1924).

- Propriedades da distribuição WG:

  A seguir apresentamos algumas propriedades da distribuição WG, que podem ser encontradas em Rodríguez-Avi *et al.* (2007).

  - A distribuição WG é uma distribuição com superdispersão, ou seja, $E(M) < Var(M)$. Consequentemente, a cauda dessa distribuição pode ser extremamente longa (efeito cauda pesada).
  - $E\left(M^k\right) < \infty$ se e somente se $\rho > k$. Portanto, a distribuição pode ter variância infinita.
  - A função de probabilidade é unimodal com moda diferente de zero.
  - Média da distribuição WG: Para o modelo WG temos que a média é dada por :

$$E(M) = \frac{ak}{\rho - 1} = \mu. \tag{A.9}$$

    A justificativa da equação em (A.9) está disponível no Apêndice B.

  - Variância da distribuição WG: A variância é dada por

$$\sigma^2 = Var(M) \quad = \mu + \frac{(k+1)}{(\rho-2)}\mu + \frac{(k+\rho-1)}{(\rho-2)}\frac{\mu^2}{k}. \tag{A.10}$$

para $a, k > 0$ e $\rho > 2$ (se $\rho < 2$, a distribuição tem variância infinita). Justificativa ver Apêndice B.

O primeiro termo representa a variabilidade devido à aleatorização do acidente (Poisson), o segundo termo expressa a submissão (fator externo: λ) e o último termo a predisposição (fator interno: v). O efeito externo ou ambiental (submissão) está representado pela distribuição gama dado v do segundo estágio e a predisposição (efeito interno) pela variável $v = (1-p)/p$, cuja distribuição é obtida do terceiro estágio dada por

$$f(v) = \frac{\Gamma(k+\rho)}{\Gamma(k)\Gamma(\rho)} v^{k-1}(1+v)^{-(k+\rho)}, v > 0.$$

Estas fontes de variabilidades estão resumidas na Tabela 15. A Tabela 15 será útil

Tabela 15 – Decomposição da variância total da distribuição WG.

Fonte de Variabilidade	**Variância**	**Taxa de variação**
Aleatorização	$\sigma_1^2 = \mu$	$\frac{\rho-2}{k+\rho-1}\frac{k}{k+\mu}$
Submissão	$\sigma_2^2 = \frac{(k+1)}{(\rho-2)}\mu$	$\frac{k+1}{k+\rho-1}\frac{k}{k+\mu}$
Predisposição	$\sigma_3^2 = \frac{(k+\rho-1)}{(\rho-2)}\frac{\mu^2}{k}$	$\frac{\mu}{k+\mu}$
Total	$\sigma^2 = \frac{k+\rho-1}{\rho-2}\left(\mu + \frac{1}{k}\mu^2\right)$	1

para avaliar as relações entre as fontes de variabilidades e o impacto na ocorrência de acidentes ou em áreas como lexicologia (TEŠÍTELOVÁ, 1967), número de autores de artigos científicos (AJIFERUKE, 1991), número de *links* da *World Web* (LEVENE *et al.*, 2002) e esportes (RODRÍGUEZ-AVI *et al.*, 2007). A Tabela 15 mostra que a taxa de variabilidade devido à aleatorização e submissão decresce quando μ cresce, enquanto que a taxa de variabilidade devido a predisposição aumenta.

– A distribuição WG converge para a distribuição BN I (IRWIN, 1968) como segue:

Para $k, \rho \to \infty$, temos que

$$\begin{aligned} p_m &\propto \frac{(a)_m}{m!}\frac{(\theta(\rho-1))_m}{(a+(1+\theta)\rho-\theta)_m} \\ &= \frac{(a)_m}{m!}\frac{\theta^m\rho^m + 0\left(\rho^{(r-1)}\right)}{(1+\theta)^m\rho^m + 0\left(\rho^{(r-1)}\right)} \to \frac{(a)_m}{m!}\left(\frac{\theta}{1+\theta}\right)^m, \end{aligned}$$

onde $0\left(\rho^{(r-1)}\right)$ é uma função infinitesimal de ordem r e $\theta = \frac{k}{\rho-1}$. Observe que a distribuição limite é uma $BN\left(a, \frac{\theta}{1+\theta}\right)$.

– Para $k = 1$ a distribuição de WG coincide com a distribuição de Waring.

– Função geradora de probabilidades da distribuição WG:

$$A_M(s;x) = \frac{(\rho)_k}{(a+\rho)_k} {}_2F_1(a,k;a+k+\rho;z), \quad m = 0,1,..., \tag{A.11}$$

onde ${}_2F_1(a,k;a+k+\rho;s)$ é a função hipergeométrica gaussiana e $a = \frac{\mu_x(\rho-1)}{k}$. A justificativa da equação em (A.11) está disponível no Apêndice B.

– Se $\rho = a = k = \alpha$, temos que $\lim_{\alpha\to\infty} \sigma_i^2(\alpha) = \sigma_i^2$, onde $\sigma_1^2 = 1/2, \sigma_2^2 = 1/4, \sigma_3^2 = 1/4$. Para grande valores de α ou μ temos um comportamento uniforme das fontes de variabilidades, devido $\rho = k$ sugerir $p = 1/2$.

- Modelo de regressão WG (Rodríguez-Avi *et al.* (2009)):

  O modelo de regressão WG foi introduzido por Rodríguez-Avi *et al.* (2009), utilizando a seguinte função de ligação:

$$\log(\mu) = \beta_0 + x^T\beta,$$

onde $\beta^T = (\beta_1,\ldots,\beta_p)$ e $x^T = (x_1,\ldots,x_p)$ é o vetor de covariadas. A predisposição e submissão não foram consideradas como covariáveis porque são não observáveis. Como $\mu_x = E[M;x] = \frac{a_x k}{\rho-1}$ temos que $k > 0$ e $\rho > 2$. Para garantir estas suposições, (RODRÍGUEZ-AVI *et al.*, 2009) consideraram a seguinte reparametrização:

$$\beta_0 = \log(k),\ \rho_0 = \log(\rho-1),$$

onde $\beta_0 \in R$ e $\rho_0 \in R$.

Infelizmente, o modelo de WG não é identificável nos parâmetros a e k complicando a identificação das fontes de variabilidades, sendo necessário a introdução de variáveis regressoras ((RODRÍGUEZ-AVI *et al.*, 2009), (RODRÍGUEZ-AVI *et al.*, 2007)).

- Método dos Momentos Fatoriais:

  Teorema A.4. Seja M o número de acidentes de um indivíduo, então

$$M_{(r)} = M(M-1)\cdots(M-r+1) = \frac{M!}{(M-r)!}$$

com $a_{(m+r)} = a_{(r)}(a+r)_{(m)}$ e $k_{(m+r)} = k_{(r)}(k+r)_{(m)}$ temos o r-ésimo momento fatorial da distribuição de WG dada por:

$$\mu_{[r]} = E[M_{[r]}] = \frac{(a)_r(k)_r}{(\rho-1)(\rho-2)\cdots(\rho-r)}, \rho > r. \tag{A.12}$$

Prova da equação (A.12), ver Apêndice B.

Para formular o método dos métodos fatoriais vamos introduzir os seguintes momentos:

– **Momento fatorial amostral de ordem r:**

$$m_{(r)} = \frac{1}{n}\sum_{i=1}^{j} n_i\,(n_i-1)\,(n_i-2)\cdots(n_i-r+1), r = 1,2,3,... \tag{A.13}$$

onde $n_i, i = 1,2,\ldots,j$ é a frequência observada da i-ésima classe tal que $\sum_{i=1}^{j} = n$, sendo n o tamanho da amostra.

– **Momento fatorial populacional de ordem r:**

$$\mu_{[r]} = \frac{(a)_r(k)_r}{(\rho-1)(\rho-2)\cdots(\rho-r)}.$$

O princípio do método dos momentos fatoriais, para obter os estimadores dos parâmetros envolvidos na distribuição de WG, consiste em igualar os momentos fatoriais de ordem $r = 3$ com os correspondentes momentos fatoriais amostrais, isto é,

$$\begin{aligned}\mu_{(1)} &= m_{(1)},\\ \mu_{(2)} &= m_{(2)},\\ \mu_{(3)} &= m_{(3)}.\end{aligned} \tag{A.14}$$

De (A.12), (A.13) e (A.14) temos que:

$$\begin{aligned}\frac{ak}{\rho-1} &= \frac{1}{n}\sum_{i=1}^{j} n_i,\\ \frac{a\,(a-1)\,k\,(k-1)}{(\rho-1)\,(\rho-2)} &= \frac{1}{n}\sum_{i=1}^{j} n_i\,(n_i-1),\\ \frac{a\,(a-1)\,(a-2)\,k\,(k-1)\,(k-2)}{(\rho-1)\,(\rho-2)\,(\rho-3)} &= \frac{1}{n}\sum_{i=1}^{j} n_i\,(n_i-1)\,(n_i-2)\,.\end{aligned} \tag{A.15}$$

As estimativas dos parâmetros (k, ρ, a) são obtidas resolvendo o sistema de equações em (A.15).

- Método de Máxima-Verossimilhança:

A estimação dos parâmetros do modelo WG pode ser obtida pelo método da máxima verossimilhança (MV). O primeiro passo é formular a função verossimilhança (RODRÍGUEZ-AVI *et al.*, 2007; PENG; LORD; ZOU, 2014) dada por:

$$\prod_{i=1}^{n} f\,(m_i|k,\rho,a) = \prod_{i=1}^{n} \frac{\Gamma(a+\rho)\,\Gamma(k+\rho)}{\Gamma(a+k+\rho)\,\Gamma(\rho)}\,\frac{(a)_{m_i}(k)_{m_i}}{(\alpha+k+\rho)_{m_i}}\,\frac{1}{m_i!}, \tag{A.16}$$

onde n é o tamanho da amostra e m_i o número de acidentes do i-ésimo indivíduo.

O logaritmo da equação em (A.16) é dado por

$$\log \prod_{i=1}^{n} f(m_i|k,\rho,a) = \sum_{i=1}^{n} \log \left[\frac{\Gamma(a+\rho)}{\Gamma(a+k+\rho)\Gamma(\rho)} \frac{(a)_{m_i}(k)_{m_i}}{(a+k+\rho)_{m_i}} \frac{1}{m_i!} \right]. \quad \text{(A.17)}$$

Para mais detalhes e uma aplicação para dados com acidentes utilizando a expressão em (A.17), via *software R*, ver Peng, Lord e Zou (2014).

A.4 Aplicação

A Tabela 16 apresenta as frequências observadas de acidentes entre homens em uma fábrica de sabão, que podem ser encontradas em Newbold (1927). O período de exposição foi de 5 meses. Irwin (1968) ,via método dos momentos fatoriais, fez um estudo comparativo entre o modelos WG e BN que estão resumidos na Tabela 16:

Tabela 16 – Acidentes com homens em uma fábrica de sabão (exposição de 5 meses)

Observação		**Esperado**	
m	n_i	**Binomial Negativa**	**Waring Generalizada**
0	239	251	240
1	98	93	105
2	57	46	49
3	33	25	24
4	9	14	12
5	2	8	7
6	2	4	4
7-13	7	6	6
Total	447	447	447
		$\chi^2_5 = 13,7$	$\chi^2_4 = 10,6$

As frequências esperadas para os modelos BN e WG foram obtidas via método dos momentos fatoriais. As frequências observadas e esperadas para sete ou mais acidentes foram combinadas em um grupo. Logo, para a distribuição BN temos oito grupos de acidentes com cinco graus de liberdade e $\chi^2_5 = 13,7$. Para a distribuição WG também temos oito grupos de acidentes com 4 graus de liberdade e $\chi^2_4 = 10,6$. Pelo método do qui-quadrado a distribuição WG é o melhor modelo com $\chi^2_4 = 10,6$. A seguir apresentamos um gráfico para comparar os dois modelos de uma forma mais simples.

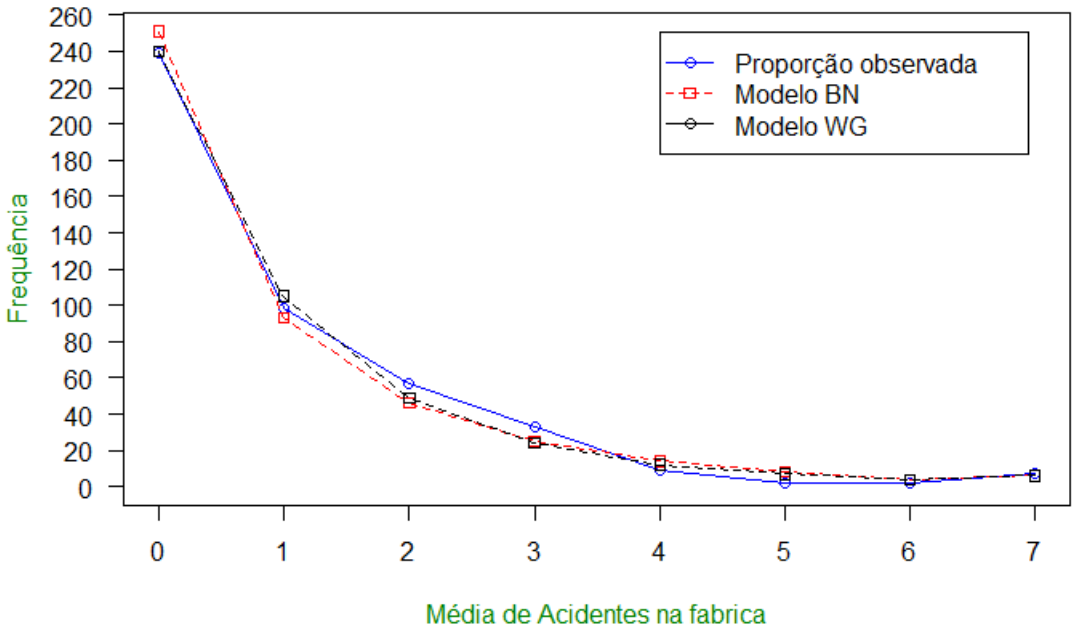

Figura 17 – Comparação dos modelos BN e WG com os dados observados

Fonte: Elaborada pelo autor.

A Figura 17 confirma graficamente que o modelo WG é o modelo que melhor ajusta aos dados observados.

Os parâmetros da distribuição WG foram estimados utilizando método dos momentos fatoriais com os seguintes resultados: $\hat{\rho} = 7,55446$, $\hat{ak} = 6,40784$ e $\hat{a} + \hat{k} = 7,10748$ de onde $\hat{a} = 6,04798$ e $\hat{k} = 1,05950$ (ver Irwin (1968)).

Tabela 17 – Partição da variância da distribuição WG

Fonte de Variabilidade	**Variância**	**Taxa de variação**
Aleatorização	0,9776	0,379
Submissão	0,3625	0,141
Predisposição	1,2366	0,480
Total	2,5767	1

Na Tabela 17 verificamos que 37,9% da dispersão dos acidentes na fábrica foram por motivos aleatórios, 14,1% da dispersão dos acidentes foram por fatores externos (submisssão), e 48% da dispersão dos acidentes foram por fatores internos (predisposição ao acidente). Esta análise dos efeitos aleatórios e não aleatórios foram cruciais para o planejamento preventivo de acidentes na fábrica.

Com os resultados da Tabela 17 temos o seguinte gráfico das fontes de variabilidades:

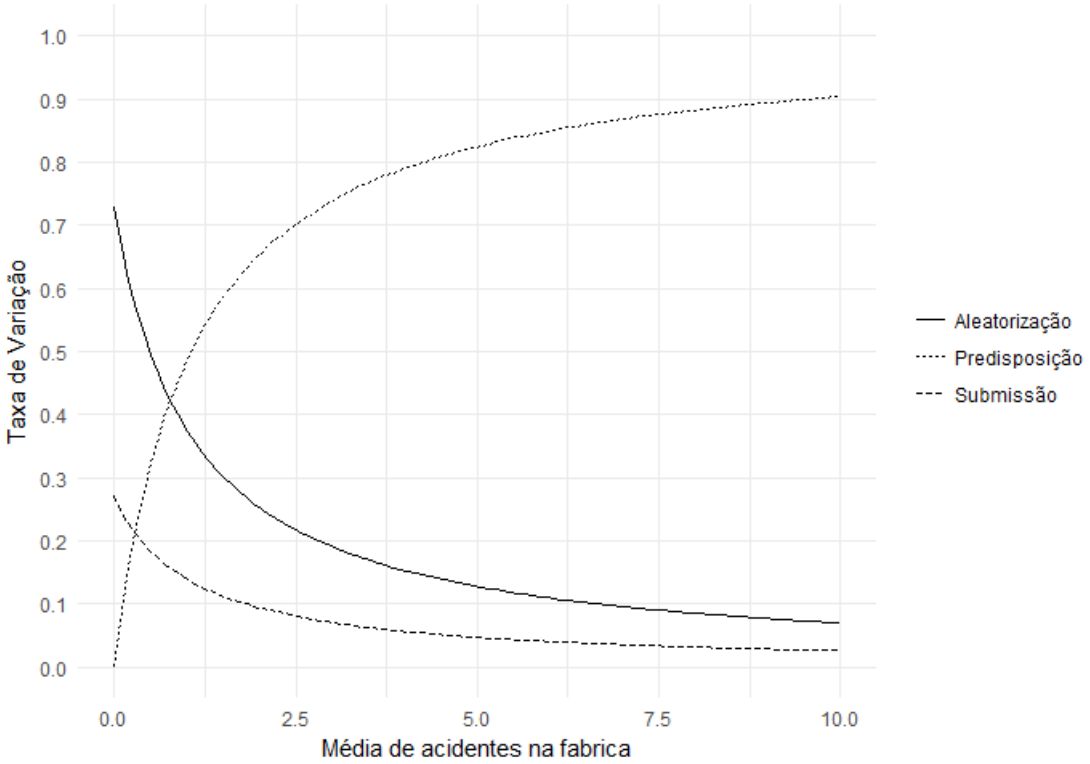

Figura 18 – Comparação das fontes de variabilidades do modelo WG

Fonte: Elaborada pelo autor.

A Figura 18 apresenta o comportamento das fontes de variabilidades com respeito a média dos acidentes na fábrica. Podemos observar que a medida que a média dos acidentes na fábrica aumenta, a taxa de variação devido ao fator interno ("Proneness") aumenta exponencialmente. Por outro lado, a variabilidade devido ao efeito aleatório e ambiental descresse exponencialmente. Concluímos que o fator interno é responsável pelo grande número de acidentes. Para o sucesso de um planejamento preventivo dos acidentes na fabrica é essencial que os aspectos pessoais dos trabalhadores sejam prioritários.

APÊNDICE

B

JUSTIFICATIVAS DOS PRINCIPAIS RESULTADOS DA DISSERTAÇÃO

- Equação A.3:

  Como,

$$p(m,\lambda) = f(\lambda)\,P(M=m|\lambda),$$

  temos que,

$$\begin{aligned}
P[M=m|a,v] &= \int_0^\infty p(m,\lambda)\,d\lambda \\
&= \int_0^\infty f(\lambda)\,P(M=m|\lambda)\,d\lambda \\
&= \int_0^\infty \frac{e^{-\lambda}\lambda^m}{m!}\cdot\frac{1}{v^a\Gamma(a)}\lambda^{a-1}e^{-\frac{\lambda}{v}}d\lambda \\
&= \frac{1}{v^a\Gamma(a)m!}\int_0^\infty \lambda^{(a+m)-1}e^{-\lambda\left(1+\frac{1}{v}\right)}d\lambda \\
&= \frac{1}{v^a\Gamma(a)m!}\cdot\frac{\Gamma(a+m)}{\left(1+\frac{1}{v}\right)^{a+m}}\int_0^\infty \frac{\left(1+\frac{1}{v}\right)^{a+m}}{\Gamma(a+m)}\cdot\lambda^{(a+m)-1}e^{-\lambda\left(1+\frac{1}{v}\right)}d\lambda \\
&= \frac{1}{v^a\Gamma(a)m!}\cdot\frac{\Gamma(a+m)}{\left(1+\frac{1}{v}\right)^{a+m}}\cdot 1 \\
&= \frac{\Gamma(a+m)}{\Gamma(a)m!}\cdot\frac{1}{v^a}\cdot\left(\frac{1}{1+\frac{1}{v}}\right)^a\cdot\left(\frac{1}{1+\frac{1}{v}}\right)^m \\
&= \frac{\Gamma(a+m)}{\Gamma(a)m!}\cdot\left(\frac{1}{v}\right)^a\cdot\left(\frac{v}{1+v}\right)^a\cdot\left(\frac{v}{1+v}\right)^m \\
&= \frac{\Gamma(a+m)}{\Gamma(a)m!}\cdot\left(\frac{1}{1+v}\right)^a\cdot\left(\frac{v}{1+v}\right)^m.
\end{aligned}$$

  Logo,

$$P[M=m|a,v] = \frac{\Gamma(a+m)}{\Gamma(a)m!}\left(\frac{1}{1+v}\right)^a\left(\frac{v}{1+v}\right)^m, m=0,1,2,\ldots;a,v>0,$$

- Equação A.4:

$$\begin{aligned}
A_M(s) &= E\left[s^M\right] \\
&= E_\lambda\left[E\left[s^M|\lambda\right]\right] \\
&= E_\lambda\left[e^{-\lambda(1-s)}\right] \\
&= \int_0^\infty e^{-\lambda(1-s)}\frac{1}{v^a\Gamma(a)}\lambda^{a-1}e^{-\frac{\lambda}{v}}d\lambda \\
&= \frac{1}{v^a\Gamma(a)}\int_0^\infty \lambda^{a-1}e^{-\lambda\left[(1-s)+\frac{1}{v}\right]}d\lambda \\
&= \frac{1}{v^a\Gamma(a)}\cdot\frac{\Gamma(a)}{\left[(1-s)+\frac{1}{v}\right]^a}\int_0^\infty \frac{\left[(1-s)+\frac{1}{v}\right]^a}{\Gamma(a)}\lambda^{a-1}e^{-\lambda\left[(1-s)+\frac{1}{v}\right]}d\lambda \\
&= \frac{1}{v^a\Gamma(a)}\cdot\frac{\Gamma(a)}{\left[(1-s)+\frac{1}{v}\right]^a} \\
&= \frac{1}{v^a}\cdot\frac{1}{\left[(1-s)+\frac{1}{v}\right]^a} \\
&= \left[\frac{1}{1+v-vs}\right]^a \\
&= \left[\frac{1}{1+\frac{1-p}{p}-\left(\frac{1-p}{p}\right)s}\right]^a \\
&= \left[\frac{p}{p+1-p-s-ps}\right]^a \\
&= \left[\frac{p}{1-(1-p)s}\right]^a,
\end{aligned}$$

onde $v=\dfrac{1-p}{p}$. Portanto,

$$A_M(s)=\left[\frac{p}{1-(1-p)s}\right]^a.$$

- Equação A.7: Temos que,

$$p_m=\frac{(\rho)_k(a)_m(k)_m}{(a+\rho)_{m+k}}\frac{1}{m!},$$

onde

$$\begin{aligned}
(a+\rho)_{m+k} &= \frac{\Gamma(a+\rho+m+k)}{\Gamma(a+\rho)} \\
&= \frac{\Gamma(a+\rho+m+k)}{\Gamma(a+\rho)}\cdot\frac{\Gamma(a+\rho+k)}{\Gamma(a+\rho+k)} \\
&= \frac{\Gamma(a+\rho+k)}{\Gamma(a+\rho)}\cdot\frac{\Gamma(a+\rho+m+k)}{\Gamma(a+\rho+k)} \\
&= (a+\rho)_k(a+\rho+k)_m.
\end{aligned}$$

Do resultado acima a distribuição WG é dada por

$$p_m = \frac{(\rho)_k}{(a+\rho)_k}\frac{(a)_m(k)_m}{(a+k+\rho)_m}\frac{1}{m!}, \quad m = 0,1,...$$

- Equação A.8 (função de probabilidade da distribuição WG obtida através da representação hierárquica):

$$p_m = \frac{(\rho)_k}{(a+\rho)_k}\frac{(a)_m(k)_m}{(a+k+\rho)_m}\frac{1}{m!}, \quad m = 0,1,\ldots$$

Temos que,

$$\begin{aligned}
p_m &= \int_0^1 \frac{\Gamma(a+m)}{\Gamma(a)m!}p^a(1-p)^m\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}p^{\rho-1}(1-p)^{k-1}dp \\
&= \frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(a+m)}{\Gamma(a)m!}\int_0^1 p^{a+\rho-1}(1-p)^{m+k-1}dp \\
&= \frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(a+m)}{\Gamma(a)m!}\frac{\Gamma(a+\rho)\Gamma(k+m)}{\Gamma(a+\rho+k+m)}\int_0^1 \frac{\Gamma(a+\rho+k+m)}{\Gamma(a+\rho)\Gamma(k+m)}p^{a+\rho-1}(1-p)^{k+m-1}dp \\
&= \frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(a+m)}{\Gamma(a)m!}\frac{\Gamma(a+\rho)\Gamma(k+m)}{\Gamma(a+\rho+k+m)} \\
&= \frac{\Gamma(\rho+k)}{\Gamma(\rho)}\frac{\Gamma(a+m)}{\Gamma(a)}\frac{\Gamma(k+m)}{\Gamma(k)}\frac{1}{\frac{\Gamma(a+\rho+k+m)}{\Gamma(a+\rho)}}\frac{1}{m!} \\
&= \frac{(\rho)_k(a)_m(k)_a}{(a+\rho)_{m+k}}\frac{1}{m!}.
\end{aligned}$$

Como $(a+\rho)_{m+k} = (a+\rho)_k(a+\rho+k)_m$, temos que:

$$p_m = \frac{(\rho)_k}{(a+\rho)_k}\frac{(a)_n(k)_m}{(a+k+\rho)_m}\frac{1}{m!}, \quad m = 0,1,...$$

- Equação A.9: Temos que,

$$E(Y|x) = \frac{ak}{\rho-1}.$$

Portanto,

$$E(Y|x,p) = \mu_x = av = a\frac{(1-p)}{p}, \quad 0 < p < 1.$$

$$
\begin{aligned}
E(Y|x) &= E_p[E(Y|x,p)] \\
&= E_p\left[a\frac{(1-p)}{p}\right] \\
&= aE_p\left[\frac{(1-p)}{p}\right] \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\int_0^1 \frac{(1-p)}{p}p^{\rho-1}(1-p)^{k-1}dp\right) \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\int_0^1 p^{\rho-1-1}(1-p)^{k+1-1}dp\right) \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(\rho-1)\Gamma(k+1)}{\Gamma(\rho-1+k+1)}\frac{\Gamma(\rho-1+k+1)}{\Gamma(\rho-1)\Gamma(k+1)}\int_0^1 p^{\rho-1-1}(1-p)^{k+1-1}dp\right) \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(\rho-1)\Gamma(k+1)}{\Gamma(\rho+k)}\right) \\
&= a\left(\frac{\Gamma(\rho-1)\Gamma(k+1)}{\Gamma(\rho)\Gamma(k)}\right) \\
&= a\left(\frac{k!\,(\rho-2)!}{(k-1)!\,(\rho-1)!}\right) \\
&= \frac{ak}{\rho-1}.
\end{aligned}
$$

Portanto,

$$E(Y|x) = \frac{ak}{\rho-1}.$$

- Equação A.10:

$$Var(M) \quad = \mu_x + \frac{(k+1)}{(\rho-2)}\mu_x + \frac{(k+\rho-1)}{(\rho-2)}\frac{\mu_x^2}{k}.$$

Temos que,

$$
\begin{aligned}
Var(Y|x) &= E_p[Var(Y|x,p)] + Var_p[E(Y|x,p)], \\
Var(Y|x) &= E_p\left[a\frac{(1-p)}{p} + a\left(\frac{(1-p)}{p}\right)^2\right] + Var_p\left[a\frac{(1-p)}{p}\right], \\
Var(Y|x) &= E_p\left[a\frac{(1-p)}{p}\right] + E_p\left[a\left(\frac{(1-p)}{p}\right)^2\right] + Var_p\left[a\frac{(1-p)}{p}\right]. \qquad \text{(B.1)}
\end{aligned}
$$

A primeira componente da expressão acima é dada por $E_p\left[a\frac{(1-p)}{p}\right] = \frac{ak}{\rho-1}$. As duas

últimas componentes de (B.1) são obtidas, respectivamente, como segue :

$$
\begin{aligned}
E_p\left[a\left(\frac{(1-p)}{p}\right)^2\right] &= aE_p\left[\frac{(1-p)^2}{p^2}\right] \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\int_0^1 \frac{(1-p)^2}{p^2}p^{\rho-1}(1-p)^{k-1}dp\right) \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\int_0^1 p^{\rho-2-1}(1-p)^{k+2-1}dp\right) \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(\rho-2)\Gamma(k+2)}{\Gamma(\rho+k)}\frac{\Gamma(\rho+k)}{\Gamma(\rho-2)\Gamma(k+2)}\int_0^1 p^{\rho-2-1}(1-p)^{k+2-1}d\right. \\
&= a\left(\frac{\Gamma(\rho+k)}{\Gamma(\rho)\Gamma(k)}\frac{\Gamma(\rho-2)\Gamma(k+2)}{\Gamma(\rho+k)}\right) \\
&= a\left(\frac{\Gamma(\rho-2)\Gamma(k+2)}{\Gamma(\rho)\Gamma(k)}\right) \\
&= a\left(\frac{(\rho-3)!\,(k+1)!}{(\rho-1)!\,(k-1)!}\right) \\
&= \frac{ak(k+1)}{(\rho-1)(\rho-2)}.
\end{aligned}
$$

Logo,

$$
E_p\left[a\left(\frac{(1-p)}{p}\right)^2\right] = \frac{ak(k+1)}{(\rho-1)(\rho-2)}, \tag{B.2}
$$

$$
Var_p\left[a\frac{(1-p)}{p}\right] = E\left[a^2\left(\frac{(1-p)}{p}\right)^2\right] - \left\{E\left[a\frac{(1-p)}{p}\right]\right\}^2,
$$

$$
\begin{aligned}
E\left[a^2\left(\frac{(1-p)}{p}\right)^2\right] &= a^2E_p\left[\frac{(1-p)^2}{p^2}\right] \\
&= a^2\frac{k(k+1)}{(\rho-1)(\rho-2)}
\end{aligned}
$$

e

$$
\begin{aligned}
Var_p\left[a\frac{(1-p)}{p}\right] &= E\left[a^2\left(\frac{(1-p)}{p}\right)^2\right] - \left\{E\left[a\frac{(1-p)}{p}\right]\right\}^2 \\
&= \frac{a^2k(k+1)}{(\rho-1)(\rho-2)} + \left(\frac{ak}{\rho-1}\right)^2 \\
&= \frac{a^2k(k+1)(\rho-1)}{(\rho-1)^2(\rho-2)} - \frac{a^2k^2(\rho-2)}{(\rho-1)^2(\rho-2)} \\
&= \frac{a^2k[(k+1)(\rho-1)-k(\rho-2)]}{(\rho-1)^2(\rho-2)} \\
&= \frac{a^2k(k+\rho-1)}{(\rho-1)^2(\rho-2)},
\end{aligned}
$$

$$Var_p\left[a\frac{(1-p)}{p}\right] = \frac{a^2k(k+\rho-1)}{(\rho-1)^2(\rho-2)}. \tag{B.3}$$

Finalmente, substituindo (A.10), (B.2) e (B.3) em (B.1) temos o resultado final:

$$\begin{aligned} Var(M) &= \frac{ak}{\rho-1} + \frac{ak(k+1)}{(\rho-1)(\rho-2)} + \frac{a^2k(k+\rho-1)}{(\rho-1)^2(\rho-2)} \\ &= \mu_x + \frac{(k+1)}{(\rho-2)}\mu_x + \frac{(k+\rho-1)}{(\rho-2)}\frac{\mu_x^2}{k}. \end{aligned}$$

- Equação A.11:

$$\begin{aligned} A_M(s) &= \sum_{m=0}^{\infty} P[M=m]s^m \\ &= \sum_{m=0}^{\infty} \frac{(\rho)_k}{(a+\rho)_k}\frac{(a)_m(k)_m}{(a+k+\rho)_m}\frac{1}{m!}s^m \\ &= \frac{(\rho)_k}{(a+\rho)_k}\sum_{m=0}^{\infty}\frac{(a)_m(k)_m}{(a+k+\rho)_m}\frac{s^m}{m!} \\ &= \frac{(\rho)_k}{(a+\rho)_k}{}_2F_1(a,k;a+k+\rho;s). \end{aligned}$$

Portanto, segue que

$$A_M(s) = \frac{(\rho)_k}{(a+\rho)_k}{}_2F_1(a,k;a+k+\rho;s).$$

- Equação A.12:

$$\mu_{[r]} = E[M_{[r]}] = \frac{(a)_r(k)_r}{(\rho-1)(\rho-2)\cdots(\rho-r)}, \rho > r.$$

Temos que:

$$\begin{aligned} \mu_{(r)} &= \sum_{j=r}^{\infty}\frac{1}{(j-r)!}\cdot\frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)}\cdot\frac{a_{(j)}k_{(j)}}{(a+k+\rho)_{(j)}} \\ &= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)}\sum_{j=r}^{\infty}\frac{a_{(j)}k_{(j)}}{(a+k+\rho)_{(j)}(j-r)!}. \end{aligned}$$

Seja $m = j - r \rightarrow j = m + r$, então

$$\begin{aligned}
\mu_{(r)} &= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)} \sum_{m=0}^{\infty} \frac{a_{(m+r)}k_{(m+r)}}{(a+k+\rho)_{(m+r)}(m)!} \\
&= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)} \sum_{m=0}^{\infty} \frac{a_{(r)}(a+r)_{(m)}k_{(r)}(k+r)_{(m)}}{(a+k+\rho)_{(r)}(a+k+\rho+r)_{(m)}(m)!} \\
&= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)} \cdot \frac{a_{(r)}k_{(r)}}{(a+k+\rho)_{(r)}} \sum_{m=0}^{\infty} \frac{(a+r)_{(m)}(k+r)_{(m)}}{(a+k+\rho+r)_{(m)}(m)!} \\
&= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)} \cdot \frac{a_{(r)}k_{(r)}}{(a+k+\rho)_{(r)}} \sum_{m=0}^{\infty} \frac{(a+r)_{(m)}(k+r)_{(m)}}{(a+r+k+r+\rho-r)_{(m)}(m)!} \\
&= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)} \cdot \frac{a_{(r)}k_{(r)}}{(a+k+\rho)_{(r)}} \cdot \frac{\Gamma(\rho-r)\Gamma(a+r+k+r+\rho-r)}{\Gamma(k+r+\rho-r)\Gamma(a+r+\rho-r)} \\
&= \frac{\Gamma(k+\rho)\Gamma(a+\rho)}{\Gamma(\rho)\Gamma(a+k+\rho)} \cdot \frac{a_{(r)}k_{(r)}}{\frac{\Gamma(a+k+\rho+r)}{\Gamma(a+k+\rho)}} \cdot \frac{\Gamma(\rho-r)\Gamma(a+k+\rho+r)}{\Gamma(k+\rho)\Gamma(a+\rho)} \\
&= \frac{a_{(r)}k_{(r)}\Gamma(\rho-r)}{\Gamma(\rho)} \\
&= \frac{a_{(r)}k_{(r)}}{(\rho-1)(\rho-2)\cdots(\rho-r)}.
\end{aligned}$$

Portanto,

$$\mu_{[r]} = \frac{(a)_r(k)_r}{(\rho-1)(\rho-2)\cdots(\rho-r)}, \rho > r.$$

- Equação (2.4)

$$\begin{aligned}
S_p(t) &= p_0 + \sum_{m=1}^{\infty} p_m \{S(t)\}^m \\
&= p_0 + (1-p_0) \frac{\sum_{m=1}^{\infty} p_m \{S(t)\}^m}{1-p_0} \\
&= p_0 + (1-p_0) S_p^*(t).
\end{aligned}$$

- Equação (2.2):

$$
\begin{aligned}
S_p(t) &= P[T \geq t] \\
&= P[\{M=0\} \cup \{Y \geq t, \quad M \geq 1\}] \\
&= P\left[\{M=0\} \cup \left\{\min_{1\leq i\leq M}\{Z_i\} \geq t\right\}\right] \\
&= P[M=0] + P[Z_1 \geq t, \quad Z_2 \geq t, \ldots, \quad Z_M \geq t] \\
&= p_0 + \sum_{m=1}^{\infty} P[M=m]\, P[Z_1 \geq t, \quad Z_2 \geq t, \ldots, \quad Z_M \geq t | M=m] \\
&= p_0 + \sum_{m=1}^{\infty} P[M=m]\, P[Z_1 \geq t, \quad Z_2 \geq t, \ldots, \quad Z_m \geq t] \\
&= p_0 + \sum_{m=1}^{\infty} p_m \prod_{i=1}^{m} P[Z_i \geq t] \\
&= p_0 + \sum_{m=1}^{\infty} p_m \prod_{i=1}^{m} S(t) \\
&= p_0 + \sum_{m=1}^{\infty} p_m \{S(t)\}^m \\
&= \sum_{m=0}^{\infty} p_m \{S(t)\}^m.
\end{aligned}
$$

Portanto, segue que

$$S_p(t) = A_M(S(t)).$$

APÊNDICE

C

PROGRAMAS EM RSTUDIO

C.1 Códigos da aplicação do MRDWLD: HIV.

```
rm(list = ls(all=TRUE))
library(BMS)
require(survival)

#################### Dados ########################
dados <- read.table("AIDS.txt", header = TRUE, sep="")
y=c(dados$tf)
status=c(dados$cens)
x1=c(dados$gr)

############################
###### Kaplan-Meier #######
############################

ekm_g=survfit(Surv(y,status)~x1)
plot(ekm_g,xlab= "Tempos (dias)",ylab = "Sp(t)",
     mark.time = F,col=c(1,3),lty = c(1,1))
legend(40,0.5,lty = c(1,1),c("HIV: Negativo","HIV: Positivo"),
       lwd = 1,bty = "n",col=c(1,3))

########################################################
########## função de maxima verosimilhança ############
########################################################
```

```
log_vero <-function(vpar,y,X,status)
{
  alpha=(vpar[1])
  lambda=(vpar[2])
  rho =vpar[3]
  beta = vpar[-c(1:3)]
  linear=c(X%*%beta)
  a1=(rho-1)*exp(linear)
  n=length(y)
  fp=numeric(n)
  Sp=numeric(n)
  S_0=exp(-exp(lambda)*y^alpha)
  f_0=exp(lambda)*y^(alpha-1)*alpha*S_0
  log_lik=0
  for(k in 1:n){
    a=a1[k]
    fp[k]=f_0[k]*f21hyper(a+1,2,a+rho+2,S_0[k])*rho*a/((a+rho+1)*(a+rho))
    Sp[k]=f21hyper(a,1,a+rho+1,S_0[k])*rho/(a+rho)
    log_lik=log_lik+status[k]*log(fp[k])+(1-status[k])*log(Sp[k])
  }
  log_lik
}

X=model.matrix(~x1)
nc=ncol(X)
vpar=c(0.5,2,3,rep(-0.01,nc))
log_vero(vpar,y,X,status)
fit3 = optim(vpar,fn=log_vero,y=y,X=X,status=status, control=list(fnscale=-1),
       method="L-BFGS-B",hessian=T,lower = c(0.01,-Inf,2.001,
       rep(-Inf,nc)),upper = c(rep(length(vpar))))

EMV=fit3$par
EMV
############## criterios #################
AICW=2*5-2*log_vero(EMV,y,X,status)
AICW

n=length(y)
BICW=log(n)*5-2*log_vero(EMV,y,X,status)
```

```
BICW

log_vero(EMV,y,X,status)
##########################################

alfa=fit3$par[1]
lamda=fit3$par[2]
ro=fit3$par[3]
beta0=fit3$par[4]
beta1=fit3$par[5]
beta1
############ taxas de cura ###############
#######x=0#######
mu1=exp(beta0)
mu1
a1=(ro-1)*mu1
a1
p1=ro/(a1+ro)
p1
#######x=1#######
mu2=exp(beta0+beta1)
mu2
a2=(ro-1)*mu2
a2
p2=ro/(a2+ro)
p2

############################################################
### curvas da função de sobrevivência de longa duração ###
############################################################

Sp=function(a,rho,lambda,alpha,y)
{
  n=length(y)
  Sp=numeric(n)
  S_0=exp(-exp(lambda)*y^alpha)
  for(k in 1:n){
    Sp[k]=f21hyper(a,1,a+rho+1,S_0[k])*rho/(a+rho)
  }
```

```
  return(Sp)
}
curve(Sp(0.64,2.26,-9.14,1.35,x),0,700,ylim=c(0,1),add=T, lty=2, col="red")
curve(Sp(2.11,2.26,-9.14,1.35,x),0,700,ylim=c(0,1),add=T, lty=2, col="blue")


MVC=-solve(fit3$hessian)
EP=sqrt(diag(MVC))
Est=EMV/EP

p_valor=2*(1-pnorm(abs(Est)))
Saida=cbind(EMV, EP,abs(Est), p_valor)
rownames(Saida)=c("alpha","lambda", "rho",paste("beta_", 0:(nc-1), sep = ""))
print(Saida, digits = 2)
library(xtable)
xtable(Saida, digits=3)

##############################
########### Sem HIV ##########
##############################

Z2=X[1:21,1:2]

VAR=function(vpar,Z2){
  alpha=(vpar[1])
  lambda=(vpar[2])
  rho =vpar[3]
  beta = vpar[-c(1:3)]
  mu=exp(Z2%*%beta)
  mu
  sub=2*mu/(rho-2)
  pred=rho*mu^2/(rho-2)
  vt=mu+sub+pred
  sal=cbind(mu=mu,sub=sub,pred=pred,vt=vt)
  sal
}
vpa=VAR(EMV,Z2)
vv=vpa[,-4]/vpa[,4]*100
colnames(vv)=c("Aleatório","Fragilidade externa","Fragilidade interna")
```

```
boxplot(vv,ylab="Taxa de variação (%)",lwd=2)
apply(vv,2,mean)

##############################
############Com HIV###########
##############################

Z1=X[22:103,1:2]

VAR=function(vpar,Z1){
  alpha=(vpar[1])
  lambda=(vpar[2])
  rho =vpar[3]
  beta = vpar[-c(1:3)]
  mu=exp(Z1%*%beta)
  mu
  sub=2*mu/(rho-2)
  pred=rho*mu^2/(rho-2)
  vt=mu+sub+pred
  sal=cbind(mu=mu,sub=sub,pred=pred,vt=vt)
  sal
}
vpa=VAR(EMV,Z1)
vv=vpa[,-4]/vpa[,4]*100
colnames(vv)=c("Aleatório","Fragilidade externa","Fragilidade interna")
boxplot(vv,ylab="Taxa de variação (%)",lwd=2)
apply(vv,2,mean)
```

C.2 Códigos da aplicação do MRDWLD: Melanoma.

```
rm(list = ls(all=TRUE))
library(BMS)
Sp=function(a,rho,lambda,alpha,y)
{
n=length(y)
Sp=numeric(n)
S_0=exp(-exp(lambda)*y^alpha)
for(k in 1:n){
Sp[k]=f21hyper(a,1,a+rho+1,S_0[k])*rho/(a+rho)
```

```
}
return(Sp)
}
fp=function(a,rho,lambda,alpha,y)
{
  n=length(y)
  fp=numeric(n)
  S_0=exp(-exp(lambda)*y^alpha)
  f_0=exp(lambda)*y^(alpha-1)*alpha*S_0
  for(k in 1:n){
    fp[k]=f_0[k]*f21hyper(a+1,2,a+rho+2,S_0[k])*rho*a/((a+rho+1)*(a+rho))
  }
  return(fp)
}
####### Figuras #########
#####a=2#####

curve(Sp(2,1,-3,2,x),0,15,ylab="Função de Sobrevivência",
      xlab="Tempo",ylim=c(0,1),lty=1,col=2,main = "a=2")
curve(Sp(2,2,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=3)
curve(Sp(2,3,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=4)
curve(Sp(2,4,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=5)
curve(Sp(2,5,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=6)
legend(-0.1,0.38,c(expression(rho==1),expression(rho==2),
      expression(rho==3),expression(rho==4),expression(rho==5)),
      col=2:6,lty=c(1,1,1,1,1),bty='n')

curve(fp(2,1,-3,2,x),0,10,ylab="Função de Densidade",
      xlab="Tempo",ylim=c(0,0.35),lty=1,col=2,main = "a=2")
curve(fp(2,2,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=3)
curve(fp(2,3,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=4)
curve(fp(2,4,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=5)
curve(fp(2,5,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=6)
legend(5,0.3,c(expression(rho==1),expression(rho==2),
      expression(rho==3),expression(rho==4),expression(rho==5)),
      col=2:6,lty=c(1,1,1,1,1),bty='n')

######a=5#####
```

```
curve(Sp(5,1,-3,2,x),0,15,ylab="Função de Sobrevivência",
      xlab="Tempo",ylim=c(0,1),lty=1,col=2,main = "a=5")
curve(Sp(5,2,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=3)
curve(Sp(5,3,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=4)
curve(Sp(5,4,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=5)
curve(Sp(5,5,-3,2,x),0,15,ylim=c(0,1),add=T, lty=1, col=6)
legend(8,1,c(expression(rho==1),expression(rho==2),expression(rho==3),
      expression(rho==4),expression(rho==5)),col=2:6,lty=c(1,1,1,1,1),bty='n')

curve(fp(5,1,-3,2,x),0,10,ylab="Função de Densidade",
      xlab="Tempo",ylim=c(0,0.35),lty=1,col=2,main = "a=5")
curve(fp(5,2,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=3)
curve(fp(5,3,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=4)
curve(fp(5,4,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=5)
curve(fp(5,5,-3,2,x),0,10,ylim=c(0,1),add=T, lty=1,col=6)
legend(6,0.3,c(expression(rho==1),expression(rho==2),expression(rho==3),
      expression(rho==4),expression(rho==5)),col=2:6,lty=c(1,1,1,1,1),bty='n')


############################
##Log-verossilhança
############################
log_vero <-function(vpar,y,X,status)
{
  alpha=(vpar[1])
  lambda=(vpar[2])
  rho =vpar[3]
  beta = vpar[-c(1:3)]
  linear=c(X%*%beta)
  a1=(rho-1)*exp(linear)
  n=length(y)
  fp=numeric(n)
  Sp=numeric(n)
  S_0=exp(-exp(lambda)*y^alpha)
  f_0=exp(lambda)*y^(alpha-1)*alpha*S_0
  log_lik=0
  for(k in 1:n){
    a=a1[k]
    fp[k]=f_0[k]*f21hyper(a+1,2,a+rho+2,S_0[k])*rho*a/((a+rho+1)*(a+rho))
```

```
    Sp[k]=f21hyper(a,1,a+rho+1,S_0[k])*rho/(a+rho)
    log_lik=log_lik+status[k]*log(fp[k])+(1-status[k])*log(Sp[k])
  }
  log_lik
}
library(timereg)
data(melanoma)
names(melanoma)
y=melanoma$days/365
status=ifelse(melanoma$status==1,1,0)
x1=(melanoma$ulc)
x2=melanoma$thick/1000
x3=melanoma$sex
X=model.matrix(~x1+x2+x3)
nc=ncol(X)
vpar=c(0.5,2,3,rep(-0.01,nc))
log_vero(vpar,y,X,status)
fit1 = optim(vpar,fn=log_vero,y=y,X=X,status=status, control=list(fnscale=-1),
      method="L-BFGS-B",hessian=T,lower = c(0.01,-Inf,2.001,  rep(-Inf,nc)),
      upper = c(rep(length(vpar))))

EMV=fit1$par
EMV
MVC=-solve(fit1$hessian)
EP=sqrt(diag(MVC))
Est=EMV/EP
IC0=EMV-1.96*EP
IC1=EMV+1.96*EP
p_valor=2*(1-pnorm(abs(Est)))
Saida=cbind(EMV, EP,IC0,IC1, p_valor)
rownames(Saida)=c("alpha","lambda", "rho",paste("beta_", 0:(nc-1), sep = ""))
print(Saida, digits = 2)
library(xtable)
xtable(Saida, digits=3)
####################################
## Descomposição da variabilidade
####################################
VAR=function(vpar,X){
alpha=(vpar[1])
```

```
lambda=(vpar[2])
rho =vpar[3]
beta = vpar[-c(1:3)]
mu=exp(X%*%beta)
sub=2*mu/(rho-2)
pred=2*mu^2/(rho-2)
vt=mu+sub+pred
sal=cbind(mu=mu,sub=sub,pred=pred,vt=vt)
sal
}
vpa=VAR(EMV,X)
vpa
vv=vpa[,-4]/vpa[,4]*100
colnames(vv)=c("Aleatório","Fragilidade externa","Fragilidade interna")
boxplot(vv,ylab="Taxa de variação (%)",lwd=2)
apply(vv,2,mean)
```

C.3 Códigos da aplicação do MRBNLD: HIV.

```
rm(list = ls(all=TRUE))
library(BMS)
require(survival)

#################### Dados ########################

dados <- read.table("AIDS.txt", header = TRUE, sep="")
y=c(dados$tf)
status=c(dados$cens)
x1=c(dados$gr)

############################
###### Kaplan-Meier #######
############################

ekm_g=survfit(Surv(y,status)~x1)
plot(ekm_g,xlab = "Tempos (dias)",ylab = "Sp(t)",
     mark.time = F,col=c(1,3),lty = c(1,1))
legend(40,0.5,lty = c(1,1),c("HIV: Negativo","HIV: Positivo"),
       lwd = 1,bty = "n",col=c(1,3))
```

```
#########################################################
########## função de maxima verosimilhança ############
#########################################################

log_vero <-function(vpar,y,X,status)
{
  alpha=(vpar[1])
  lambda=(vpar[2])
  a=vpar[3]
  beta = vpar[-c(1:3)]
  linear=c(X%*%beta)
  p=1/((1+(exp(linear)/a)))
  S_0=exp(-exp(lambda)*y^alpha)
  f_0=exp(lambda)*y^(alpha-1)*alpha*S_0

fp=a*f_0*((p/(1-(1-p)*S_0))^a)*(1-p/(1-(1-p)*S_0))
Sp=(p/(1-(1-p)*S_0))^a
    log_lik=sum(status*log(fp)+(1-status)*log(Sp))
log_lik
}

X=model.matrix(~x1)
nc=ncol(X)
vpar=c(1,-1,1,rep(-0.01,nc))
log_vero(vpar,y,X,status)
fit2=optim(vpar,fn=log_vero,y=y,X=X,status=status, control=list(fnscale=-1),
           method="L-BFGS-B",hessian=T,lower = c(0.01,-Inf,0.01,
            rep(-Inf,nc)),upper = c(rep(length(vpar))))

EMV=fit2$par
EMV
############## criterios #################
AICBN=2*5-2*log_vero(EMV,y,X,status)
AICBN

n=length(y)
BICBN=log(n)*5-2*log_vero(EMV,y,X,status)
BICBN
```

```
log_vero(EMV,y,X,status)
########################################
alfa=fit2$par[1]
lamda=fit2$par[2]
a=fit2$par[3]
beta0=fit2$par[4]
beta1=fit2$par[5]

############ taxas de cura ###############
#######x=0#######
mu1=exp(beta0)
mu1
p1=(1+mu1/a)^-1
p1
t1=p1^a
t1
#######x=1#######
mu2=exp(beta0+beta1)
mu2
p2=(1+mu2/a)^-1
p2
t2=p2^a
t2

############################################################
### curvas da função de sobrevivência de longa duração ###
############################################################

Sp=function(a,p,lambda,alpha,y)
{
  n=length(y)
  Sp=numeric(n)
  S_0=exp(-exp(lambda)*y^alpha)
  for(k in 1:n){
    Sp[k]=(p/(1-(1-p)*S_0[k]))^a
  }
  return(Sp)
}
```

```
curve(Sp(5,0.98,-4.71,0.93,x),0,700,ylim=c(0,1),add=T, lty=2, col="red")
curve(Sp(5,0.94,-4.71,0.93,x),0,700,ylim=c(0,1),add=T, lty=2, col="blue")


MVC=-solve(fit2$hessian)
EP=sqrt(diag(MVC))
Est=EMV/EP
p_valor=2*(1-pnorm(abs(Est)))
Saida=cbind(EMV, EP,abs(Est), p_valor)
rownames(Saida)=c("alpha","lambda", "a",paste("beta_", 0:(nc-1), sep = ""))
print(Saida, digits = 2)
library(xtable)
xtable(Saida, digits=3)

##############################
########### Sem HIV ##########
##############################

Z2=X[1:21,1:2]

VAR=function(vpar,Z2){
  alpha=(vpar[1])
  lambda=(vpar[2])
  a =vpar[3]
  beta = vpar[-c(1:3)]
  mu=exp(Z2%*%beta)
  mu
  frag=mu^2/a
  vt=mu+frag
  sal=cbind(mu=mu,frag=frag,vt=vt)
  sal
}
vpa=VAR(EMV,Z2)
vv=vpa[,-3]/vpa[,3]*100
colnames(vv)=c("Aleatório","Sensibilidade")
boxplot(vv,ylab="Taxa de Variabilidade",lwd=2)
apply(vv,2,mean)
```

```
##############################
############Com HIV###########
##############################
par(mfrow=c(1,2))

Z1=X[22:103,1:2]

VAR=function(vpar,Z1){
  alpha=(vpar[1])
  lambda=(vpar[2])
  a =vpar[3]
  beta = vpar[-c(1:3)]
  mu=exp(Z1%*%beta)
  mu
  frag=mu^2/a
  vt=mu+frag
  sal=cbind(mu=mu,frag=frag,vt=vt)
  sal
}
vpa=VAR(EMV,Z1)
vv=vpa[,-3]/vpa[,3]*100
colnames(vv)=c("Aleatório ","Sensibilidade")
boxplot(vv,ylab="Taxa de Variabilidade",lwd=2)
apply(vv,2,mean)
```

C.4 Códigos da simulação : MRDWLD

```
rm(list = ls(all=TRUE))
library(orthopolynom)
library(BMS)

rwg=function(n,rho,a){
  p=rbeta(n,rho,1)
  lamb=rgamma(n,shape=a,scale=(1-p)/p)
  M=rpois(n,lamb)
  M
}
############################
```

```
PC=function(L,S,theta0){

  Lf=rep(0,length(theta0))
  Uf=Lf
  Cf=Lf
  for(k in 1:length(theta0)){
    Lf[k]=mean(ifelse(L[,k]>theta0[k],1,0))
    Uf[k]=mean(ifelse(S[,k]<theta0[k],1,0))
    Cf[k]=1-Lf[k]-Uf[k]
  }
  saida=list(Lf=Lf,Uf=Uf,Cf=Cf)
  saida
}

EQM=function(MP, thet){
  r=length(thet)
  RR=nrow(MP)
  eq=matrix(0,RR,r)
  for (j in 1:r){
    eq[,j]=(MP[,j]-thet[j])^2
  }
  saida=sqrt(apply(eq,2,mean))
  saida
}

## vies
VIES=function(MP, thet){
  r=length(thet)
  RR=nrow(MP)
  eq=matrix(0,RR,r)
  for (j in 1:r){
    eq[,j]=(MP[,j]-thet[j])
  }
  saida=(apply(eq,2,mean))
  saida
}

############################
##Log-verossilhança
```

```
##########################
log_veroW <-function(vpar,y,X,status)
{
  lambda=(vpar[1])
  rho =vpar[2]
  beta = vpar[-c(1:2)]
  linear=c(X%*%beta)
  a1=(rho-1)*exp(linear)
  n=length(y)
  fp=numeric(n)
  Sp=numeric(n)
  S_0=exp(-lambda*y)
  f_0=(lambda)*S_0
  log_lik=0
  for(k in 1:n){
    a=a1[k]
    fp[k]=f_0[k]*f21hyper(a+1,2,a+rho+2,S_0[k])*rho*a/((a+rho+1)*(a+rho))
    Sp[k]=f21hyper(a,1,a+rho+1,S_0[k])*rho/(a+rho)
    log_lik=log_lik+status[k]*log(fp[k])+(1-status[k])*log(Sp[k])
  }
  log_lik
}
N=numeric()
te=numeric()
y=numeric()
ce = numeric()
status = numeric()
n=500
rho0=4
beta00=2
beta10=-1
beta11=-1
x1=rbinom(n,1,0.5)
x2=runif(n,0,1)
mu0=exp(beta00+beta10*x1+beta11*x2)
a0=(rho0-1)*mu0
lamb0=1
p0=rho0/(a0+rho0)
M1 = NULL
```

```
MI=NULL
MS=NULL
iNREP =1000
irep = ntent = 0
set.seed(135790)
options(show.error.messages = F)
while (irep < iNREP)
{
  ntent <- ntent + 1

  for(i in 1:n){
    N[i] =  rwg(1, rho0,a0[i])
    if (N[i]==0) te[i]=Inf
    else te[i]=  min(rexp(N[i], lamb0))
    ce[i] = runif(1,2,5) #
    y[i] = min(te[i], ce[i])
    status[i] = ifelse(te[i] < ce[i], 1, 0)
  }
  X=model.matrix(~1+x1+x2)
  nc=ncol(X)
  vpar=c(lamb0,rho0,beta00,beta10,beta11)
  log_veroW(vpar,y,X,status)
  fit = try(optim(vpar,fn=log_veroW,y=y,X=X,status=status, control=list(fnscale=-1),
                  method="L-BFGS-B",hessian=T,lower = c(0.001,2.001,rep(-Inf,nc)),
                  upper = c(Inf,Inf, rep(Inf,nc))))
  if (class(fit) != "try-error" && fit$convergence == 0 && fit$par[2]< 12 ){
    cov1=-fit$hessian
    va=diag(solve(cov1))
    if(all(va>0)){
      EP=sqrt(va)
      I=fit$par-1.96*EP
      S=fit$par+1.96*EP
      MI=rbind(MI,I)
      MS=rbind(MS,S)
      M1=rbind(M1,fit$par)
      irep=irep+1
      print(irep)
    }
  }
```

```
}

options(show.error.messages = TRUE)
cat("\n Numero de amostras geradas = ", ntent,"\n")
cat("\n Numero de amostras geradas = ", n,"\n")
med=apply(M1,2,mean)
desv_pa=apply(M1,2,sd)
QM=EQM(M1,vpar)
VIE=VIES(M1,vpar)
Pc=PC(MI,MS,vpar)

saida=cbind(med,desv_pa,VIE,QM,Pc$Cf)
colnames(saida)=c("Mean","Standard Desv","VIES", "RMS","CP")
rownames(saida)=c("lambda","rho","beta_0","beta_1","beta_2")
print(saida, digits = 3)
library(xtable)
xtable(t(saida), digits=3)

######## Histograma #######
rho=M1[1:1000,2]
hist(rho,freq=F,main = NULL , xlab = expression(rho), ylab = "Frequência")
d.rv=density(rho)
d.rv.max = d.rv$x[which.max(d.rv$y)]
lines(density(rho), col="black",lty=5)
abline(v=rho0,col="red", lwd=1)
```

C.5 Códigos da simulação: MRBNLD

```
rm(list = ls(all=TRUE))
library(orthopolynom)
library(BMS)

rbn=function(n,a,nu){
  lamb=rgamma(n,shape=a,rate =nu)
  M=rpois(n,lamb)
  M
}
############################
PC=function(L,S,theta0){
```

```
  Lf=rep(0,length(theta0))
  Uf=Lf
  Cf=Lf
  for(k in 1:length(theta0)){
    Lf[k]=mean(ifelse(L[,k]>theta0[k],1,0))
    Uf[k]=mean(ifelse(S[,k]<theta0[k],1,0))
    Cf[k]=1-Lf[k]-Uf[k]
  }
  saida=list(Lf=Lf,Uf=Uf,Cf=Cf)
  saida
}

EQM=function(MP, thet){
  r=length(thet)
  RR=nrow(MP)
  eq=matrix(0,RR,r)
  for (j in 1:r){
    eq[,j]=(MP[,j]-thet[j])^2
  }
  saida=sqrt(apply(eq,2,mean))
  saida
}

## vies
VIES=function(MP, thet){
  r=length(thet)
  RR=nrow(MP)
  eq=matrix(0,RR,r)
  for (j in 1:r){
    eq[,j]=(MP[,j]-thet[j])
  }
  saida=(apply(eq,2,mean))
  saida
}

############################
##Log-verossilhança
############################
```

```
log_veroW <-function(vpar,y,X,status)
{
  lambda=(vpar[1])
  nu =vpar[2]
  beta = vpar[-c(1:2)]
  mu=exp(X%*%beta)
  a=mu/nu

  S_0=exp(-lambda*y)
  f_0=(lambda)*S_0
  Sp=(1+nu*(1-S_0))^(-a)
  fp=mu*f_0*(1+nu*(1-S_0))^(-a-1)
  log_lik=sum(status*log(fp)+(1-status)*log(Sp))

  log_lik
}
N=numeric()
te=numeric()
y=numeric()
ce = numeric()
status = numeric()
n=500
nu0=4
beta00=2
beta10=-1
beta11=-1
x1=rbinom(n,1,0.5)
x2=runif(n,0,1)
mu0=exp(beta00+beta10*x1+beta11*x2)
a0=(mu0/nu0)
lamb0=1
M1 = NULL
MI=NULL
MS=NULL
iNREP =1000
irep = ntent = 0
set.seed(135790)
options(show.error.messages = F)
while (irep < iNREP)
```

```
{
  ntent <- ntent + 1

  for(i in 1:n){

    N[i] =  rbn(1,a0[i],nu0)
    if (N[i]==0) te[i]=Inf
    else te[i]=  min(rexp(N[i],lamb0))
    ce[i] = runif(1,2,5)
    y[i] = min(te[i], ce[i])
    status[i] = ifelse(te[i] < ce[i], 1, 0)
  }
  X=model.matrix(~1+x1+x2)
  nc=ncol(X)
  vpar=c(lamb0,nu0,beta00,beta10,beta11)
  log_veroW(vpar,y,X,status)
  fit = try(optim(vpar,fn=log_veroW,y=y,X=X,status=status, control=list(fnscale=-1),
                  method="L-BFGS-B",hessian=T,lower = c(0.001,0.001,rep(-Inf,nc)),
                  upper = c(Inf,Inf, rep(Inf,nc))))
  if (class(fit) != "try-error" && fit$convergence == 0 &&fit$par[2]<3*nu0){
    summary(fit)
    cov1=-fit$hessian
    va=diag(solve(cov1))
    if(all(va>0)){
      EP=sqrt(va)
      I=fit$par-1.96*EP
      S=fit$par+1.96*EP
      MI=rbind(MI,I)
      MS=rbind(MS,S)
      M1=rbind(M1,fit$par)
      irep=irep+1
      print(irep)
    }
  }
}

options(show.error.messages = TRUE)
cat("\n Numero de amostras geradas = ", ntent,"\n")
cat("\n Numero de amostras geradas = ", n,"\n")
```

```
med=apply(M1,2,mean)
desv_pa=apply(M1,2,sd)
QM=EQM(M1,vpar)
VIE=VIES(M1,vpar)
Pc=PC(MI,MS,vpar)

saida=cbind(med,desv_pa,VIE,QM,Pc$Cf)
colnames(saida)=c("Mean","Standard Desv","VIES", "RMS","CP")
rownames(saida)=c("lambda","nu0","beta_0","beta_1","beta_2")
print(saida, digits = 3)
library(xtable)
xtable(t(saida), digits=3)

######## Histograma ##########
nu=M1[1:1000,2]
hist(nu,freq=F,main = NULL , xlab = expression(v), ylab = "Frequência")
d.rv=density(nu)
d.rv.max = d.rv$x[which.max(d.rv$y)]
lines(density(nu), col="black",lty=5)
abline(v=nu0,col="red", lwd=1)
```


ICMC
USP
SÃO CARLOS