UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Métodos de inferência para modelos de regressão aplicados a dados sorológicos de pacientes HIV+ com múltiplos níveis de censura à esquerda

Matheus Henrique Felix

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

ICMC USP SÃO CARLOS

Matheus Henrique Felix

Métodos de inferência para modelos de regressão aplicados a dados sorológicos de pacientes HIV+ com múltiplos níveis de censura à esquerda

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. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Vera Lucia Damasceno Tomazella
Coorientador: Prof. Dr. Afrânio Márcio Corrêa Vieira

USP – São Carlos
Abril de 2025

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

Fm

Felix, Matheus Henrique
Métodos de Inferência para Modelos de Regressão Aplicados a Dados Sorológicos de Pacientes HIV+ com Múltiplos Níveis de Censura à Esquerda / Matheus Henrique Felix; orientadora Vera Lucia Damasceno Tomazella; coorientador Afrânio Márcio Corrêa Vieira. -- São Carlos, 2025.
113 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Censura à esquerda. 2. Modelagem estatística. 3. HIV. 4. LMD. 5. Níveis de censura. I. Lucia Damasceno Tomazella, Vera, orient. II. Márcio Corrêa Vieira, Afrânio, coorient. III. 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

Matheus Henrique Felix

Inference methods for regression models applied to serological data from HIV+ patients with multiple levels of left-censoring

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Vera Lucia Damasceno Tomazella
Co-advisor: Prof. Dr. Afrânio Márcio Corrêa Vieira

USP – São Carlos
April 2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Matheus Henrique Felix, realizada em 14/03/2025.

Comissão Julgadora:

Profa. Dra. Vera Lucia Damasceno Tomazella (UFSCar)

Profa. Dra. Juliana Betini Fachini Gomes (UnB)

Prof. Dr. Fabio Pratavieira (ESALQ/USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.
Em especial, ao pesquisadores do Instituto de Ciências Matemáticas e de Computação (ICMC).

AGRADECIMENTOS

A realização deste trabalho foi possível graças ao apoio, incentivo e contribuição de muitas pessoas, às quais sou imensamente grato.

Aos meus orientadores, Vera Lucia Damasceno Tomazella e Afrânio Márcio Corrêa Vieira, pela paciência, dedicação e pelos ensinamentos transmitidos ao longo desta jornada. Esta orientação foi fundamental para o desenvolvimento deste trabalho e para o meu crescimento acadêmico e pessoal.

A minha mãe Maria, meu pai Adão, meus irmãos Lucas e Adan e minha irmã Sabrina, que sempre estiveram ao meu lado, oferecendo amor, apoio incondicional e confiança em todos os momentos. Vocês foram minha base e minha inspiração para persistir e alcançar este objetivo.

Ao meu namorado, Heitor, pelo carinho, compreensão e por ser meu porto seguro nos momentos de dificuldade. Sua parceria foi essencial para que eu pudesse superar os desafios e seguir em frente.

Aos meus amigos, Marília, Grazi, Retori, Cláudio, Giovanna e Shimada, pela amizade, motivação e pelas conversas que tornaram esta jornada mais leve e significativa. Vocês tiveram um papel importante no equilíbrio entre trabalho e vida pessoal.

Além disto, o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Por fim, a todos que, de alguma forma, contribuíram para este trabalho, direta ou indiretamente, deixo aqui minha eterna gratidão.

"Não importa o que aconteça,
continue a nadar."
(WALTERS, Graham; ***Procurando Nemo****, 2003.)*

RESUMO

FELIX, M. H. **Métodos de inferência para modelos de regressão aplicados a dados sorológicos de pacientes HIV+ com múltiplos níveis de censura à esquerda**. 2025. 113 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, 2025.

Neste trabalho, o interesse está em dados que apresentam a ocorrência de *censura à esquerda*. Este tipo de censura ocorre quando não conhecemos o momento da ocorrência do evento, mas sabemos que ele ocorreu antes do tempo registrado. Um exemplo que ocorre com frequência é na análise de dados laboratoriais sorológicos, os quais impactam diretamente na tomada de decisão de médicos, pesquisadores e entre outros especialistas correlatos. Na maioria dos trabalhos publicados, são explorados estudos transversais, onde apenas um nível de censura é apresentado. No entanto, aparecem ocorrências de múltiplas medidas censuradas à esquerda, em diferentes níveis de um único paciente, bastante frequente ao se acompanhar a saúde do paciente ao longo do tempo. O mesmo acontece para testes aplicados em laboratórios diferentes, visto que os equipamentos também se diferem, gerando assim níveis diferentes para a censura. Atualmente, pesquisadores da área clínica acabam excluindo esses dados da análise. Do ponto de vista da regressão, é fundamental modelar adequadamente a influência do tempo, dos fatores e/ou das covariáveis na carga viral, além da correlação entre medidas repetidas de um mesmo paciente. Esses aspectos são cruciais para garantir inferências estatísticas confiáveis, tanto em ensaios clínicos quanto em estudos observacionais de coorte ou caso-controle. Nesse contexto, o objetivo deste trabalho é apresentar uma metodologia apropriada para lidar com diferentes níveis de censura à esquerda, assumindo distribuições Weibull e Log-Normal e aplicando-a a dados reais de cargas virais de pacientes portadores de HIV.

Palavras-chave: Censura à esquerda, Modelagem estatística, HIV, LMD, Níveis de censura.

ABSTRACT

FELIX, M. H. **Inference methods for regression models applied to serological data from HIV+ patients with multiple levels of left-censoring**. 2025. 113 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, 2025.

In this study, the focus is on data that exhibit the occurrence of left-censoring. This type of censoring occurs when the exact time of the event is unknown, but it is known to have occurred before the recorded time. A common example is in the analysis of serological laboratory data, which directly impacts decision-making by physicians, researchers, and other related specialists. Most published studies explore cross-sectional designs, where only one level of censoring is considered. However, occurrences of multiple left-censored measurements at different levels for a single patient are quite common when monitoring a patient's health over time. The same applies to tests conducted in different laboratories, as equipment varies, resulting in different censoring levels. Currently, researchers in the clinical field often exclude such data from their analyses. From a regression perspective, it is essential to adequately model the influence of time, factors, and/or covariates on viral load, as well as the correlation between repeated measurements for the same patient. These aspects are crucial for ensuring reliable statistical inferences, both in clinical trials and in observational cohort or case-control studies. In this context, the objective of this work is to present an appropriate methodology for handling data with multiple levels of left-censoring, particularly assuming Weibull and Log-Normal distributions, applied to real viral load data from HIV-infected patients.

Keywords: Left censorship, Statistical modeling, HIV, LMD, Censorship levels.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Os acontecimentos na década de 1980 com a descoberta do Vírus da Imunodeficiência Humana (HIV) e, por consequência, a síndrome da imunodeficiência adquirida (aids), mesmo não sendo tão distantes, nos trouxeram um grande número de pessoas infectadas e doentes. Segundo Canini *et al.* (2004), HIV pode ser descrito como um retrovírus que promove a disfunção imunológica crônica e progressiva no organismo devido à redução dos níveis de linfócitos CD4. Quanto menores esses níveis, maior é o risco de desenvolver a AIDS.

Com ausência de cura para a doença, a criação dos anti-retrovirais têm causado bons resultados, assim como uma melhoria no tempo de vida dos indivíduos, tornando-se foco de estudos para averiguar os impactos desses tratamentos no bem estar dos pacientes (CANINI *et al.*, 2004). A redução da carga viral do HIV é constantemente associada ao primeiro sinal de avaliação de eficácia de tratamentos a base de retrovirais (ver Wu, Ding e Gruttola (1998); Jacqmin-Gadda *et al.* (2000)). Desse modo, a análise da carga viral se torna um dos pontos de estudo, sendo que, em muitos casos, estes valores são tão baixos, que os equipamentos utilizados para a medição não são capazes de retornar o valor exato para a carga viral, que segundo BRASIL (2013), pode-se definir como o menor valor que um método é capaz de detectar com confiança, tendo em vista determinado procedimento analítico, denominando este limiar de medição como o Limite Mínimo Detectável (LMD). A análise de dados sorológicos de pacientes HIV+ envolve a investigação de diversos aspectos, incluindo o tempo até a ocorrência de eventos importantes, como a progressão da infecção para AIDS ou o início da terapia antirretroviral.

Outra característica é que esses dados frequentemente apresentam múltiplos níveis de censura à esquerda, o que torna a análise ainda mais complexa. Essa característica ocorre quando os valores observados não são diretamente mensuráveis abaixo de determinados limites de detecção, que podem variar entre diferentes exames laboratoriais, equipamentos ou metodologias. Em tais casos, diferentes níveis de censura podem ser registrados para o mesmo paciente, refletindo a heterogeneidade das medições e exigindo abordagens estatísticas adequadas para

lidar com essas limitações.

Em outras palavras, este tipo de dado é conhecido como incompleto ou censurado, ou no contexto de exames sorológicos, vinculado como não detectável. Mais especificamente a censura à esquerda, como é conhecida este tipo de medida, é muito comum em estudos laboratoriais sorológicos. O trabalho original de um modelo para dados censurados à esquerda foi proposto por Tobin (1958) com base na ideia de se tratar as censuras como medidas latentes. Outros trabalhos surgiram na intenção de se estender este modelo de dados censurados. Bolfarine *et al.* (2013) possibilitam que a distribuição dos dados possam pertencer à classe das distribuições elípticas, que a resposta tenha excesso de zeros ou truncadas, além de demonstrar a estimação no paradigma Bayesiano.

Além disso Hughes (1999) apresenta uma estrutura bastante geral para modelagem de dados de carga viral censurados à esquerda contendo efeitos mistos no preditor linear do modelo. Samson, Lavielle e Mentré (2006) propuseram uma extensão para o algoritmo de expectativa-maximização de aproximação estocástica (SAEM) aplicados a dados da carga viral HIV. Lyles, Williams e Chuachoowong (2001) apresentam a proposta da estimação do coeficiente de correlação entre duas variáveis de carga viral HIV RNA, ambas com censura à esquerda com LMD conhecido. Mais recentemente, Solomon e Weissfeld (2017) desenvolveram uma abordagem de modelagem multivariada de dados censurados via pseudo-verossimilhança. Nos trabalhos revisados até o momento, nenhum trata de múltiplos níveis de censuras.

Outro ponto de impacto vem com o avanço da tecnologia, que contribui para que o LMD se modifique ao decorrer do tempo, ou mesmo que ele seja diferente para testes aplicados em laboratórios diferentes, visto que os equipamentos também se diferem, gerando assim níveis diferentes para a censura (HUGHES, 1999). Por outro lado, este tipo de característica nunca foi tratada desta maneira, Lyles, Williams e Chuachoowong (2001) por exemplo, tratou a presença de 2 níveis de censura de maneira intervalar, aplicando o *log* nos valores das cargas virais em cópias/ml para diminuição da escala. Wang, Lin e Lachos (2018) por outro lado, tratou as censuras fixando um valor para o limite mínimo de 50 cópias/ml, além da aplicação do logaritmo na base 10, com o mesmo intuito.

Neste contexto, considerando a complexidade dos dados, este trabalho propõe o uso de modelos de regressão que acomodam múltiplos níveis de censura à esquerda. Esses modelos possibilitam a consideração da incerteza associada ao tempo de infecção, uma vez que alguns indivíduos podem ter sido infectados antes do início do estudo. Além disso, permitem a inclusão de covariáveis, como idade, sexo, carga viral e contagem de células CD4, para avaliar o impacto desses fatores na progressão da doença. Outra vantagem desses modelos é a capacidade de estimar indicadores essenciais, como a taxa de progressão da infecção, o tempo até o início da terapia antirretroviral e a sobrevida média dos pacientes soropositivos. Dessa forma, eles proporcionam uma compreensão mais precisa da evolução da infecção e auxiliam os pesquisadores na tomada de decisões fundamentadas sobre estratégias de tratamento e prevenção.

Para garantir que os modelos ajustados ofereçam uma representação adequada dos dados, é fundamental avaliar sua qualidade de ajuste. Essa avaliação pode ser realizada por meio da análise de resíduos, especialmente adaptada para situações em que há censura. Embora existam resíduos tradicionais voltados para esse tipo de análise, como os de Cox-Snell, Martingale e Deviance Colosimo e Giolo (2006b), tais abordagens apresentam limitações em cenários mais complexos, como o que se propõe neste estudo. Como alternativa, Wu, Feng e Li (2019) propõem os Resíduos de Probabilidade de Sobrevivência Aleatorizada Transformados pela Normal (NRSP), os quais, sob modelo corretamente especificado, seguem distribuição normal padrão, facilitando a avaliação gráfica e inferencial do ajuste. Essa metodologia pode ser facilmente adaptada para situações com censura à esquerda, como é o caso dos dados analisados neste trabalho.

1.1 Objetivo

O objetivo geral desta dissertação é propor um modelo de regressão para dados de pacientes diagnosticados com HIV, considerando múltiplos níveis de censura à esquerda. Para isso, serão empregados métodos inferenciais frequentistas e bayesianos, utilizando as distribuições Weibull e Log-Normal para modelar a carga viral. Compreender a variação da carga viral entre diferentes pacientes é essencial para auxiliar na tomada de decisões clínicas personalizadas. Além disso, a análise da média da carga viral em distintos grupos de pacientes pode fornecer respostas relevantes para questões que, de outra forma, permaneceriam sem solução.

Os dados utilizados neste estudo foram disponibilizados por um ensaio clínico conduzido no Laboratório de Investigação em Dermatologia e Imunodeficiências da Faculdade de Medicina da Universidade de São Paulo, como parte de um projeto coordenado pelos Drs. Duarte e Silva.

1.2 Organização do Trabalho

Este trabalho está organizado como segue. O Capítulo 1 apresenta a introdução, com os objetivos e a estrutura do trabalho. O Capítulo 2 traz uma revisão metodológica sobre censura, modelos de probabilidade, além de conceitos relacionados a intervalos de confiança e testes de hipóteses assintóticos. O Capítulo 3 descreve a metodologia principal deste estudo, com foco no modelo de regressão com múltiplos níveis de censura, incluindo seu desenvolvimento, testes, simulações e aplicações. O Capítulo 4 aborda métodos bayesianos, detalhando *prioris*, técnicas de amostragem e aplicações práticas. Por fim, o Capítulo 5 apresenta as conclusões e propostas para trabalhos futuros.

CAPÍTULO

2

REFERENCIAL TEÓRICO

Este capítulo apresenta uma breve revisão dos conceitos fundamentais que servirão de base para a modelagem proposta, incluindo definições de diferentes tipos de censura, algumas distribuições de probabilidade e outras noções que possam, eventualmente, se mostrar relevantes para o desenvolvimento do trabalho.

2.1 Censuras

Os conceitos apresentados nesta subseção foram baseados em Colosimo e Giolo (2006a), salvo indicação em contrário. Um componente amplamente utilizado em diversas áreas de aplicação, como em modelos de regressão e, em especial, na análise de sobrevivência, é o conceito de observações incompletas ou parciais, conhecidas como censuras. Quando essas observações são incluídas, sua exclusão pode levar a resultados e interpretações enviesados. Por isso, sua incorporação torna-se não apenas interessante, mas também necessária.

2.1.1 Censura a Direita

Censura à direita é um termo utilizado em estatística para descrever uma situação em que o valor exato de uma observação é desconhecido, mas sabe-se que ele está acima de um determinado limite. Isso ocorre frequentemente em estudos de tempo até o evento, como na análise de sobrevivência, onde o evento de interesse (como a morte, falência de uma empresa, etc.) ainda não ocorreu até o fim do período de observação.

Dentro desta categoria, tipos mais conhecidos podem ser dados pela censura do Tipo I e II e aleatória, visto que o valor real da ocorrência do evento, mesmo que desconhecido, está a direita do valor observado. Este tipo de situação, é muito comum quando aplicado na área de sobrevivência, onde o interesse é estudar o tempo.

Censura do Tipo I

Comumente definidas pelo pré estabelecimento do término do estudo, sendo que alguns indivíduos neste intervalo de tempo não irão apresentar o evento de interesse, consequentemente, tendo suas observações censuradas. Este tipo de censura pode ser encontrado em várias áreas de estudo, incluindo a área médica, onde o tempo de acompanhamento dos indivíduos é observado por um período pré-estabelecido para verificar a recorrência de uma doença. Os indivíduos que não apresentaram este evento são considerados censurados, caracterizando, nesse caso, a censura do tipo I.

Censura do Tipo II

Outro caso, é quando o fim do estudo é definido após a ocorrência de um número pré-fixado de eventos de interesse, categorizando o restante dos indivíduos do estudo, como censuras do tipo II. A vantagem deste tipo de experimento, é que como certos indivíduos tendem a ter o seu tempo até a ocorrência do evento de interesse muito longo, faz com que, em muitos casos, o tempo até o fim do estudo seja minimizado, por consequência reduzindo tempo e dinheiro investidos.

Censura Aleatória

Em muitos casos, a censura foge do escopo do pesquisador, ou seja, mesmo tomando cuidados e precauções, existem eventos não esperados. Em muitos casos, este tipo é observada em experimentos como o abandono da unidade experimental da pesquisa, ou em casos médicos, a qual é muito observada, pode ocorrer a morte súbita do indivíduo por algum motivo, que não seja aquele esperado quando realizado o experimento.

A censura também pode ser definida utilizando objetos matemáticos, seja i o índice associado a i-ésima unidade experimental suscetível a ocorrência do evento, $i = 1, \ldots, n$.. Defina-se a censura por meio de uma função indicadora δ_i, com y_i sendo considerado o valor até a falha. Deste modo, tem-se que.

$$\delta_i = \begin{cases} 1 & , \text{ se } y_i \text{ for o tempo até a falha;} \\ 0 & , \text{ se } y_i \text{ for o tempo até a censura.} \end{cases}$$

Tendo conhecimento de como é realizado cada um dos três tipos de censuras definidos anteriormente, é possível então realizar a ilustração de como é definida suas realizações, dada pela Figura 1.

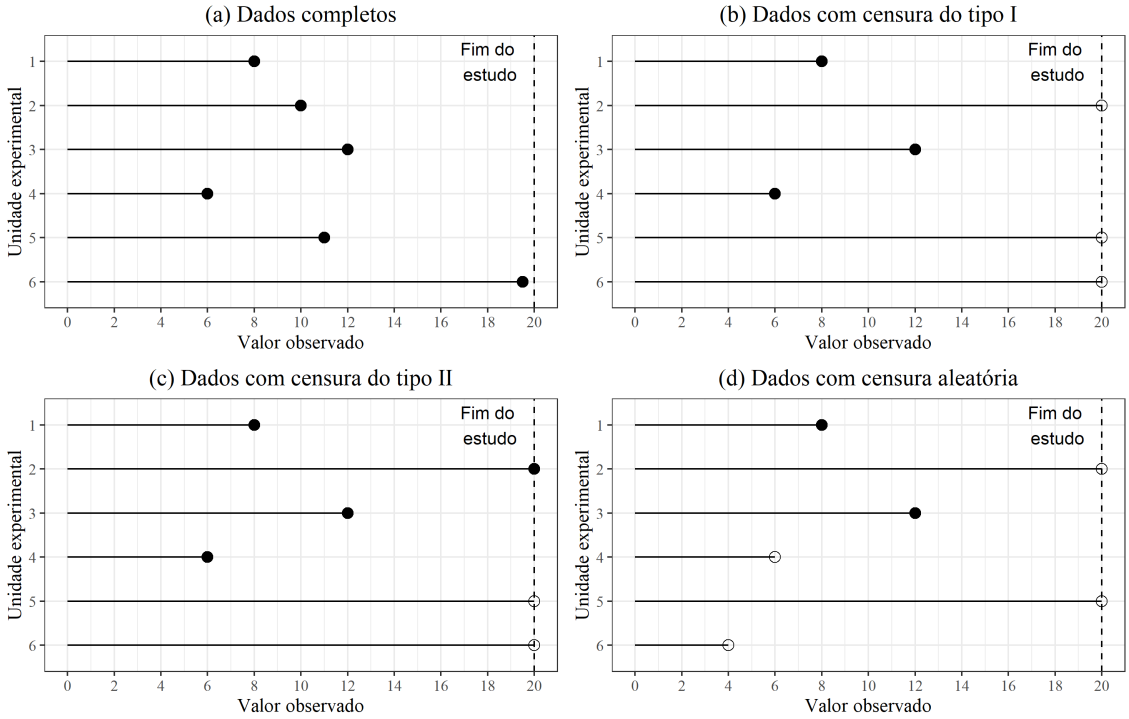

Figura 1 – Figura com algumas definições de censura em que • falha e ○ censura.

A Figura 1 apresenta algumas definições de censura. No painel (a), todas as unidades experimentais apresentaram o evento de interesse antes do final do estudo. Já no painel (b), algumas unidades não apresentaram o evento até o final. No painel (c), o estudo foi finalizado após um número pré-fixado de falhas. Por fim, em (d), algumas unidades experimentais tiveram seu acompanhamento interrompido devido a algum motivo, e outras não realizaram o evento até o final.

O mecanismo de censura apresentado até aqui, denominado como ’censura à direita’, é um dos mais comuns. No entanto, também existem outras formas de censura, como a ’censura intervalar’ e a ’censura à esquerda’.

2.1.2 Censura à Esquerda

Um caso em especial que será tratado neste trabalho é o caso da censura à esquerda, sendo possível ser observada na Figura 2.

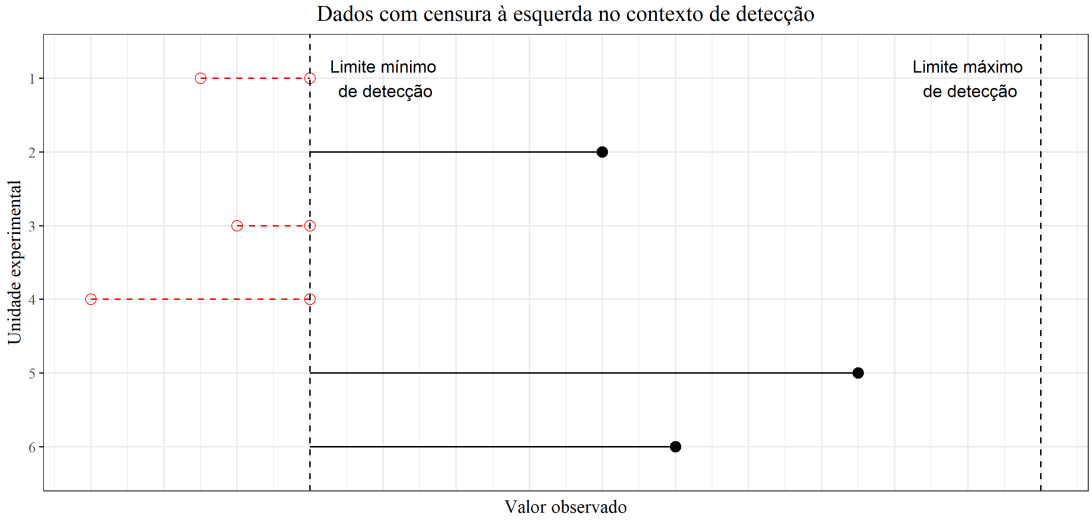

Figura 2 – Ilustração da censura à esquerda, sendo • a falha e ○ censura.

Na Figura 2, observa-se que o valor real da variável é inferior ao valor efetivamente registrado, determinado pelo limite mínimo de detecção. Nesses casos, os equipamentos não conseguem fornecer o valor exato devido à sua baixa magnitude. Em outros contextos, como em testes de carga viral, é possível que o valor ultrapasse o limite máximo de detecção, configurando uma censura à direita; entretanto, tais situações não serão abordadas neste trabalho. Em estudos sorológicos de pacientes HIV+, essa limitação pode ocorrer devido à sensibilidade restrita dos testes laboratoriais, que não detectam concentrações muito baixas de determinados biomarcadores.

Ao contrário do observado na censura à direita, este tipo de censura acontece quando o valor observado é maior do que o verdadeiro valor. Para casos envolvendo o tempo, significa que o evento de interesse já aconteceu antes de iniciar o estudo. Um exemplo clássico para este tipo é um estudo que tem como objetivo determinar a idade em que crianças são alfabetizadas em determinada comunidade, porém, ao iniciar verifica-se que algumas crianças já sabem ler e escrever, caracterizando assim a censura à esquerda. Do mesmo modo, pode ocorrer que até o final do experimento algumas crianças não saibam ler, caracterizando assim também a censura à direita. Estes casos especiais são denominados como duplamente censurados (TURNBULL, 1976).

Além das citadas acima, existem outras definições de censura (ver Colosimo e Giolo (2006a), Lawless (2011)). Outra característica comum em alguns experimentos e que diversas vezes é confundida com a censura é o truncamento. O truncamento pode ser definido pela exclusão de unidades experimentais devido a uma condição. Neste tipo de estudo, apenas os indivíduos que tiveram o evento de interesse são incluídos. Em estudos sobre a AIDS, o evento de interesse é frequentemente o desenvolvimento da doença. Indivíduos já infectados pelo HIV, mas que ainda não desenvolveram AIDS, podem não ser identificados, resultando em sua exclusão da amostra. Essa omissão pode introduzir vieses nos resultados, pois impede uma compreensão

completa da progressão da infecção pelo HIV até o desenvolvimento da AIDS.

2.2 Alguns modelos de probabilidades

Nesta seção apresenta-se alguns modelos probabilísticos que sejam adequados para aplicação aos dados reais, considerando as particularidades neles presentes. Com isso, serão discutidas as distribuições Exponencial, Weibull e Log-normal, que possuem como principal característica a modelagem de dados com valores estritamente positivos. Ressaltamos que a Log-normal será utilizada como uma alternativa à abordagem tradicional encontrada na literatura, que frequentemente aplica a transformação logarítmica à variável resposta para adequá-la à distribuição normal no contexto da modelagem da carga viral.

2.2.1 Distribuição Weibull

Proposta por Weibull (1939), a distribuição Weibull teve sua aplicação também estudada, posteriormente, pelo mesmo autor (WEIBULL, 1951). Dai em diante, foi possível verificar sua utilização nas mais diversas áreas, como modelos biométricos, áreas laboratoriais, estudos industriais e análise de sobrevivência e confiabilidade (MARTZ; WALLER, 1982). Sua grande vantagem para utilização, pode ser vista pela característica da sua grande flexibilidade, vista as diferentes formas que ela pode assumir para suas funções densidade, sobrevivência, acumuladas, e em especial a função taxa de falha, a qual pode apresentar comportamento monótono, sendo ele crescente, decrescente ou contante.

Seja Y uma variável aleatória definida em algum espaço de probabilidade $(\Omega, \mathscr{F}, P)$ com distribuição Weibull, com parâmetros dados por $\alpha > 0$ e $\gamma > 0$, sendo γ um parâmetro de forma e α um parâmetro de escala. Sua f.d.p. (função densidade de probabilidade) é dada por:

$$f_Y(y) = \frac{\gamma}{\alpha^\gamma} y^{\gamma-1} \exp\left\{ -\left(\frac{y}{\alpha}\right)^\gamma \right\} \mathbb{I}_{\{y \geq 0\}}. \tag{2.1}$$

A função de sobrevivência é dada por:

$$S_Y(y) = \int_y^\infty \frac{\gamma}{\alpha^\gamma} y^{\gamma-1} \exp\left\{ -\left(\frac{y}{\alpha}\right)^\gamma \right\} dy = \exp\left\{ -\left(\frac{y}{\alpha}\right)^\gamma \right\}, \tag{2.2}$$

A f.d.a. (função distribuição acumulada) da Weibull é dada por,

$$F_Y(y) = 1 - S_Y(y) = 1 - \exp\left\{ -\left(\frac{y}{\alpha}\right)^\gamma \right\}. \tag{2.3}$$

Uma peculiaridade desta distribuição é que, para específicos valores do parâmetro de forma γ, obtêm-se distribuições diferentes já relatadas na literatura. Podendo, seguir uma distri-

buição Exponencial, Rayleigh, além de duas situações que pode ser considerada aproximações das distribuições Log-Normal e Normal.

O interessante em se utilizar uma forma paramétrica a seus dados é que a forma das suas funções está suscetível à alteração do valor de seus parâmetros, assim como mostra a Figura 3.

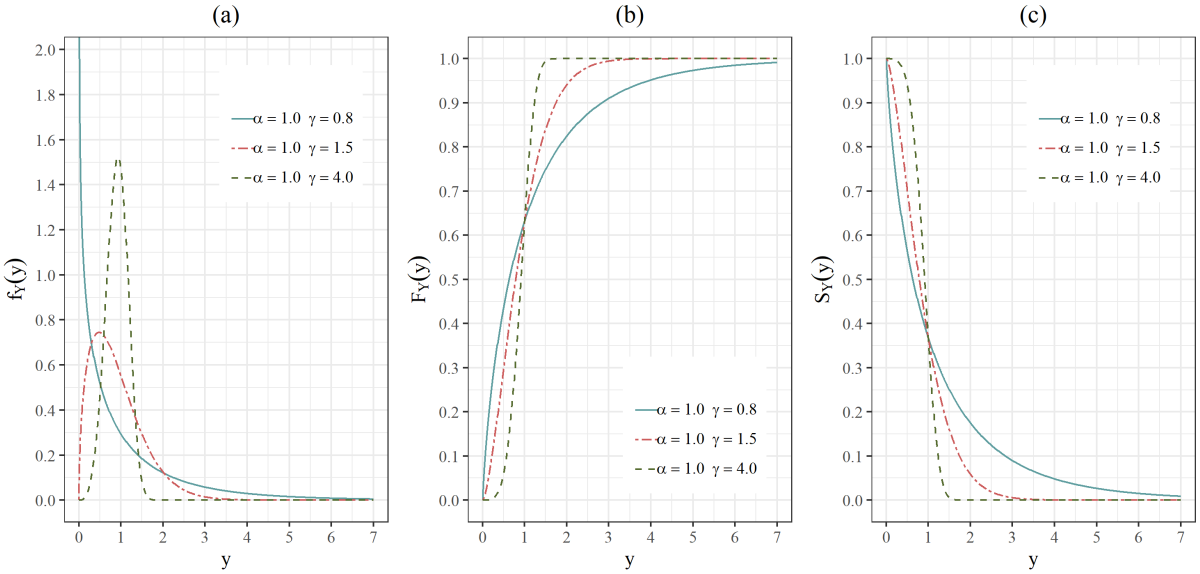

Figura 3 – Funções densidade de probabilidade (a), acumulada (b) sobrevivência (c) para a distribuição Weibull.

2.2.2 Distribuição Log-Normal

A distribuição Log-normal é uma excelente opção para diversas aplicações de modelagem. Esta distribuição assim como nome induz, tem uma relação direta com a distribuição normal, em que pode-se analisar dados advindos desta distribuição por meio da distribuição normal, desde que consideremos a aplicação do logaritmo nos dados, ao invés de sua utilização em sua escala original (COLOSIMO; GIOLO, 2006a).

Seja Y uma variável aleatória definida em algum espaço de probabilidade $(\Omega, \mathscr{F}, P)$ com distribuição log-normal, com parâmetros dados por $\mu \in \mathbb{R}$ e $\sigma > 0$, sendo μ um parâmetro de locação e σ um parâmetro de dispersão. Sua f.d.p. (função densidade de probabilidade) é dada por:

$$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma}} \exp\left\{\frac{-(ln(y)-\mu)^2}{2\sigma^2}\right\} \mathbb{I}_{\{y \geq 0\}}. \tag{2.4}$$

A função de sobrevivência é dada por:

$$S_Y(y) = \int_y^\infty \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} = 1 - \Phi\left(\frac{log(y)-\mu}{\sigma}\right). \tag{2.5}$$

A f.d.a. (função de distribuição acumulada) da log-normal é dada por:

$$F_Y(y) = 1 - S_Y(y) = \Phi\left(\frac{log(y) - \mu}{\sigma}\right). \tag{2.6}$$

A relevância de adotar uma abordagem paramétrica para a análise dos dados está no fato de que a forma das funções pode ser ajustada de acordo com a variação dos valores de seus parâmetros, conforme ilustrado na Figura 4.

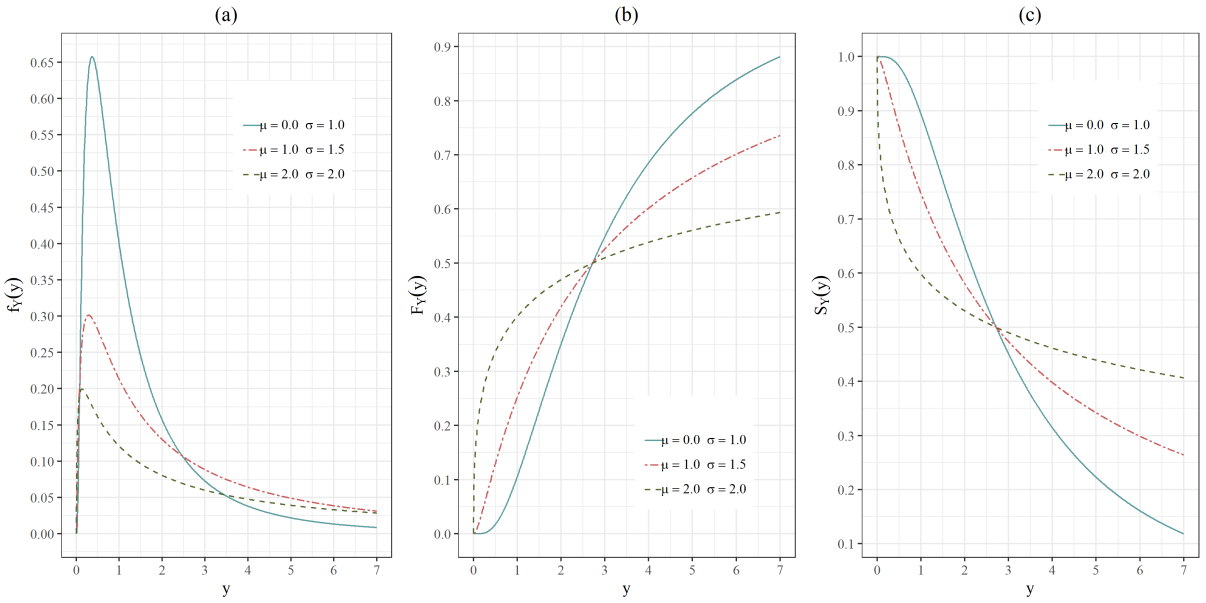

Figura 4 – Funções densidade de probabilidade (a), Função acumulada (b) Função de sobrevivência (c) para a distribuição Log-Normal.

2.3 Modelo de regressão paramétrico

Na literatura, diversos métodos são utilizados para estimar os parâmetros do modelo. No entanto, um dos mais conhecidos é o método de maximização da verossimilhança, especialmente por sua eficácia em cenários que envolvem dados censurados. Existem outras propriedades interessantes neste estimador, como possuir ótimas propriedades para quando temos um tamanho amostral grande, tendo como exemplo a normalidade assintótica dos estimadores Colosimo e Giolo (2006a).

Neste contexto pretende-se definir de forma ampla como é realizada a estimação via máxima verossimilhança para o caso em que existe a presença de censura à esquerda, tendo como caso especial a utilização a distribuição Weibull, sendo essa seção baseada em Colosimo e Giolo (2006a) e Estevam (2014), com algumas alterações para se adaptar ao contexto do tipo de censura.

Seja $(y_1, y_2, \ldots, y_n)$ uma amostra proveniente da distribuição de interesse. Defina-se $f_Y(y)$ como a função densidade de probabilidade associada a essa distribuição. Além disso, seja

$\mathbf{X} \in \mathbb{R}^{n \times p}$ a matriz de covariáveis associada às observações, em que cada linha $\mathbf{x}_i$ representa o vetor de covariáveis do i-ésimo indivíduo.

Suponha que não haja presença de dados censurados. Nesse caso, a função de verossimilhança para um vetor de parâmetros θ é dada por:

$$L(\theta \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} f(y_i \mid \theta, \mathbf{x}_i) \tag{2.7}$$

Neste método de estimação, busca-se encontrar os valores de θ que maximizam a função de verossimilhança $L(\theta \mid \mathbf{y})$. Pode-se observar que, na ausência de censura, a contribuição de cada observação é dada diretamente pela função densidade de probabilidade $f_Y(y)$ definida anteriormente.

Em situações em que os dados estão censurados à direita, como discutido em (COLOSIMO; GIOLO, 2006a), a contribuição das observações censuradas é representada pela função de sobrevivência. No entanto, no presente trabalho, consideramos o caso de censura à esquerda, em que a contribuição para a função de verossimilhança é dada pela função de distribuição acumulada $F_Y(y)$, a qual é complementar à função de sobrevivência, ou seja, $S_Y(y) = 1 - F_Y(y)$.

De forma geral, a função de verossimilhança, em um caso mais simples sem a presença de covariáveis e com censura à esquerda, pode ser expressa como o produto das contribuições das n observações, conforme a seguinte equação:

$$L(\theta \mid \mathbf{y}, \delta) = \prod_{i=1}^{n} [f(y_i \mid \theta)]^{\delta_i} \times [F_Y(y_i \mid \theta)]^{1-\delta_i} \tag{2.8}$$

em que $\delta_i = 1$ indica que a observação y_i não é censurada, e $\delta_i = 0$ indica que ela é censurada à esquerda.

Considerando a distribuição Weibull definida na Subseção 2.2.1 e substituindo as respectivas funções densidades(2.1) e acumuladas na equação 2.8, temos que a log-verossimilhança é dada por:

$$l(\alpha, \gamma | y, \delta) = \sum_{i=1}^{n} \{\delta_i [(\gamma - 1) y_i - \gamma ln(\alpha) + ln(\gamma) - (\frac{y_i}{\alpha})^{\gamma}] + (1 - \delta_i) ln(1 - exp\{-(\frac{y_i}{\alpha})^{\gamma}\})\}$$

com $l(\alpha, \gamma | y, \delta) = log[L(\alpha, \gamma | y, \delta)]$.

Considerando o conjunto de p variáveis independentes é possível expressar as covariáveis por meio da matriz $X_{pxn} = [x_1, \ldots, x_i, \ldots, x_n]$, sendo $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ para $i = 1, \ldots, n$. Podemos incorporar as covariáveis por meio da função de ligação logaritimica no parâmetro de escala, da seguinte maneira:

$$\alpha_i = \exp(x_i \beta) \Rightarrow \ln \alpha_i = x_i \beta;$$

em que β é o vetor de parâmetros de regressão associado às covariáveis, isto é, um vetor $p \times 1$ cujos elementos representam os efeitos (ou pesos) de cada variável explicativa sobre o parâmetro de escala α_i.

Deste modo, pode-se definir a log-verossimilhança da distribuição Weibull com a presença de covariáveis da seguinte maneira:

$$l(\alpha,\gamma,\beta|y,X,\delta) = \sum_{i=1}^{n}\{\delta_i[(\gamma-1)y_i - \gamma x_i\beta + ln(\gamma) - (\frac{y_i}{\exp(x_i\beta)})^{\gamma}] + (1-\delta_i)ln(1-exp\{-(\frac{y_i}{\exp(x_i\beta)})^{\gamma}\})\}$$

Para encontrar os estimadores de máxima verossimilhança, é necessário encontrar os valores de que maximizem $l(\theta)$. Logo, para encontrar basta resolver o seguinte sistema de equações abaixo:

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

Dado que na maioria dos casos a complexidades das equações inviabiliza soluções analíticas e forma fechada, é necessário a utilização de métodos numéricos para realizar a estimação, usando de aparato o *software R* para a utilização de métodos implementados. Na Seção 3.2 poderá ser encontrado o processo de estimação com maiores detalhes.

2.4 Estimação Intervalar

As inferências referentes aos parâmetros envolvidos, quando há a existência de uma amostra suficientemente grande, pode se apoiar em propriedades de normalidade assintótica (ver Migon, Gamerman e Louzada (2014) e Ospina e Ferrari (2012)). Seja, $\widehat{\theta_i}$ a estimativa de verossimilhança associada ao i-ésimo parâmetro associado ao modelo, de modo que $\widehat{\theta_i} - \theta_i$ apresenta distribuição assintótica normal p-variada, com vetor de médias zero, e matriz covariâncias $I^{-1}(\hat{\theta})$, conhecida como a inversa da matriz de informação de Fisher estimada. Consequentemente, empregando um nível de confiança de $100(1-\alpha)\%$, $\alpha \in [0,1]$, é possível realizar a obtenção dos intervalos de confiança assintóticos.

$$IC(\theta_i; 100(1-\alpha)\%) = \widehat{\theta_i} \pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{\theta_i})},$$

em que $z_{\alpha/2}$ é o percentil de $\alpha/2$ referente à distribuição normal padrão, e $\text{Var}(\widehat{\theta_i})$ representa os valores da diagonal da matriz $I^{-1}(\hat{\theta})$, associado ao parâmetro.

Quando o tamanho amostral é relativamente pequeno, pode acontecer de algumas propriedades do estimador de máxima verossimilhança sejam inconsistentes (CECCOTTI, 2015).Isso

pode resultar em alguns problemas, como a obtenção de valores-limite fora dos intervalos permitidos pelo espaço paramétrico, um fenômeno comumente conhecido como problema de fronteira. Outro ponto a considerar é que o intervalo obtido é simétrico devido à suposição de normalidade, o que pode impedir a captura da assimetria presente na distribuição de alguns estimadores de máxima verossimilhança (EMV).

Como solução, alguns autores sugerem a aplicação de uma transformação nos parâmetros. Dessa forma, o intervalo de confiança é construído para o parâmetro transformado, utilizando, por exemplo, o método delta, e posteriormente reajustado para a escala origina (CECCOTTI, 2015). No entanto, neste trabalho, esse problema não será abordado caso esteja presente.

2.5 Teste de Hipótese

Nesta subseção será apresentada uma explicação sobre o Teste de Wald e o Teste da Razão de Verossimilhanças (TRV). Serão abordados os conceitos fundamentais de cada teste, sua aplicabilidade e a maneira como são utilizados para avaliar hipóteses sobre parâmetros desconhecidos em modelos estatísticos.

2.5.1 Teste de Wald

Muitas vezes, nosso interesse é verificar, por meio de testes de hipóteses, restrições lineares nos parâmetros. Um dos testes mais utilizados é o teste de Wald, que também pode ser aplicado a estimadores não lineares. Como a obtenção da estatística teste em muitos casos é de difícil obtenção, um modo interessante é a utilização de resultados assintóticos, os quais levam em consideração as distribuições de qui-quadrado e a gaussiana ao invés de distribuições como a t-Student e F-Snedecor para amostras pequenas e com suposição de normalidade (CAMERON; TRIVEDI, 2005). Além disso a maneira de se estimar a matriz de covariância e variância para os estimadores, será realizada da mesma forma apresentada na seção 2.4.

Em Cameron e Trivedi (2005) o interesse é testar H restrições lineares para o vetor de parâmetros, porém é facilmente adaptado ao caso de teste de uma única restrição. Logo, considerando testar a restrição de um único parâmetro, a a hipótese nula sendo dada por H_0 contra a hipótese alternativa dada por H_1, onde:

$$H_0 = \theta = \theta_0$$
$$H_1 = \theta \neq \theta_0$$

em que sua estatística do teste é definida por:

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\text{var}(\hat{\theta})} \xrightarrow{\text{d}} \chi^2_1.$$

Para grande valores da estatística teste W, leva-se a rejeição da hipótese nula H_0, ou seja ao nível de significância α rejeita-se $W > \chi^2_1(1-\alpha)$, caso contrário, não há indícios de que a hipótese seja falha.

Frequentemente o interesse é centrado em testar a hipótese de nulidade de um único coeficiente. O qual pode ser testado utilizando a estatística teste sendo a raiz de W, resultando em:

$$t = \sqrt{W} = \frac{\hat{\theta}}{se(\hat{\theta})} \quad \xrightarrow{\text{d}} \quad N(0,1).$$

Também para este caso, rejeita-se a hipótese nula para grandes valores de t, e ao contrário de W, esta estatística pode ser utilizado para teste unilaterais.

De acordo com Cameron e Trivedi (2005), a estatística $\sqrt{W}$ de forma mais rigorosa é uma estatística assintótica z, porém a notação t é utilizado pelo fato de reproduzir a estatística t usual, da estimativa dividida pelo seu erro padrão. Muitos softwares para casos de amostras finitas utilizam a distribuição normal para a realização do teste para os cálculos de valores críticos e p-valores, porém, em contrapartida muitos também fazem o uso da distribuição t. O autor também afirma, que mesmo esta duas vertentes serem utilizadas para amostras finitas, nenhuma das duas está totalmente correta, exceto para o caso de regressão em que os erros são supostos como tendo distribuição normal, sendo neste, a distribuição t exata. Porém, em amostra grandes, os resultados são iguais, visto que a distribuição t converge para a distribuição normal.

2.5.2 Teste da razão de verossimilhança

Testes da razão de verossimilhança recebem esse nome por estarem diretamente ligados aos estimadores de máxima verossimilhança. Esse tipo de teste pode ser utilizado para verificar se um determinado parâmetro associado ao modelo satisfaz ou não uma certa restrição Casella e Berger (2021). Além disso, Bickel e Doksum (2015) propôs esse mesmo teste para avaliar se um modelo restrito pode ou não ser considerado apropriado, podendo esse segundo caso ser interpretado como um teste de comparação entre modelos.

Esse teste é particularmente interessante por diversos motivos, dentre os quais se destaca o fato de que toda a informação contida na amostra ou experimento é expressa pela função de verossimilhança (ver RESENDE (2007); PORTUGAL (1995)). Considerando, y_1, y_2, ..., y_n uma amostra das variáveis aleatórias i.i.d. Y_1, Y_2, ..., Y_n, a função de verossimilhança pode ser descrita como uma função dependente somente do vetor de parâmetros θ, pode ser descrita assim como definida pela Equação 2.7.

Ao restringir o vetor de parâmetros θ, dada pela hipótese nula de que ele pertença a algum subespaço. Este espaço, comumente refere-se à restrições determinadas no espaço paramétrico, e

neste contexto, a hipótese nula corresponde ao espaço paramétrico restrito, enquanto a alternativa, ao irrestrito (FERREIRA, 2008).

Sejam $y_1, y_2, \ldots, y_n$ uma realização de variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$, dependentes apenas do vetor de parâmetros θ. Além disso, seja $\mathbf{X} \in \mathbb{R}^{n \times p}$ a matriz de covariáveis associada às observações, onde cada linha $\mathbf{x}_i$ representa o vetor de covariáveis do i-ésimo indivíduo.

Definindo as hipóteses por $H_0 : \theta \in \Omega_0$ e $H_1 : \theta \in \Omega$, onde $\Omega_0 \subset \Omega$, ou seja, H_0 representa o espaço paramétrico restrito e H_1 o espaço irrestrito, a estatística da razão de verossimilhança é definida por (FERREIRA, 2008):

$$\Lambda = \frac{\sup_{\Omega_0} L(\theta \mid y, \mathbf{X})}{\sup_{\Omega} L(\theta \mid y, \mathbf{X})} \tag{2.9}$$

Para valores elevados da estatística Λ, não há indícios suficientes para rejeitar a hipótese nula H_0; por outro lado, valores baixos de Λ sugerem evidência em favor da hipótese alternativa H_1. Entretanto, como a distribuição exata de Λ não é trivial, utiliza-se a estatística transformada do teste, dada por $-2\ln(\Lambda)$, a qual, sob certas condições regulares, segue uma distribuição qui-quadrado com k graus de liberdade.

$$\text{Rejeita-se } H_0 \text{ se e somente se } -2ln(\Lambda) > \chi^2_{1-\alpha}(p),$$

em que p é definido como os graus de liberdade da distribuição de qui-quadrado, sendo que para modelos aninhados p é dado pela diferença de dimensionalidade de Ω_0 e Ω, ou seja, se os modelos estiverem aninhados p é a diferença no número de parâmetros livres nos dois modelos (WILKS, 1938). Este teste também pode ser realizado com base nas funções de log-verossimilhanças, sendo dadas por:

$$\Lambda = -2\left[\ell(\theta_0|y, \mathbf{X}) - \ell(\theta|y, \mathbf{X})\right]$$

em que $\ell(\theta \mid y, \mathbf{X}) = \ln\left[\sup_{\Omega} L(\theta \mid y, \mathbf{X})\right]$, sendo que o critério de rejeição baseia-se em uma estatística definida a partir das funções de verossimilhança.

2.6 Considerações Finais

Neste capítulo, foram apresentados os conceitos fundamentais sobre os diferentes tipos de censura, além de uma introdução às distribuições amplamente utilizadas na modelagem estatística. Esses conceitos são cruciais para a compreensão da metodologia proposta neste trabalho, proporcionando uma base sólida para a introdução e o entendimento dos novos conceitos que serão discutidos nos capítulos seguintes. Adicionalmente, foram abordados os testes de hipótese e os intervalos de confiança assintóticos, os quais desempenham um papel central na

avaliação inferencial dos parâmetros estimados, contribuindo para a fundamentação teórica dos procedimentos adotados nas análises.

CAPÍTULO

3

MODELO DE REGRESSÃO COM MÚLTIPLOS NÍVEIS DE CENSURA A ESQUERDA

Neste capítulo, serão apresentados os principais tópicos para o desenvolvimento deste trabalho, incluindo a definição de múltiplos níveis de censura à esquerda e sua incorporação na função de verossimilhança. Além disso, será explorada a aplicação dessa abordagem com as distribuições Weibull e Log-Normal. Por fim, serão discutidos testes estatísticos, dois estudos de simulação e a aplicação prática a um conjunto de dados reais

3.1 Múltiplos Níveis de Censura à Esquerda

Considere um vetor de dados com n observações, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, composto por observações independentes, amostradas de variáveis aleatórias Y_i, para $i = 1, \ldots, n$, onde Y_i representa a carga viral do i-ésimo indivíduo.

Sejam m os níveis distintos de censura considerados, representados por $\mathbf{k} = [k_1, k_2, \ldots, k_m]$, com $k_1 > k_2 > \cdots > k_m$, define-se a matriz de censura $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_n]$, onde cada vetor $\mathbf{c}_i = [c_{i1}, \ldots, c_{im}]$ indica a presença ou ausência de censura para a observação y_i, com $c_{ij} \in \{0, 1\}$ para $i = 1, \ldots, n$ e $j = 1, \ldots, m$, logo a matriz de censura $\mathbf{C}$ é definida como:

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix},$$

onde cada vetor linha $\mathbf{c}_i = [c_{i1}, c_{i2}, \ldots, c_{im}]$ indica a presença ou ausência de algum nível de censura para a observação y_i. Os elementos c_{ij} são definidos conforme os seguintes critérios:

- Se $y_i > k_m$, então $c_{i1} = c_{i2} = \cdots = c_{im} = 0$, indicando que a observação y_i não é censurada.

- Se $y_i \leq k_1$, então $c_{i1} = 1$ e $c_{i2} = c_{i3} = \cdots = c_{im} = 0$, indicando censura no nível mais baixo k_1.
- Se $k_j < y_i \leq k_{j+1}$, com $j \in \{1, \ldots, m-1\}$, então $c_{i,j+1} = 1$ e $c_{i,j+2} = \cdots = c_{im} = 0$, indicando censura no nível k_{j+1}.

A matriz de censura $\mathbf{C}$ é construída com base em limiares $k_1 > k_2 > \cdots > k_m$, que representam diferentes níveis de censura. De forma geral, se $y_i > k_1$, então a observação y_i não é censurada e todos os elementos da linha correspondente na matriz são iguais a zero. Se $y_i \leq k_m$, então a observação é censurada no nível mais baixo k_m, ou seja, $c_{im} = 1$ e os demais elementos da linha são zero.

Já para valores intermediários, ou seja, quando $k_{j+1} < y_i \leq k_j$ com $j \in \{1, \ldots, m-1\}$, a observação pode ser censurada no nível k_j, representado por $c_{ij} = 1$ e zeros nos níveis seguintes. No entanto, em aplicações reais, esse intervalo intermediário pode não implicar censura, dependendo do contexto; por exemplo, um exame realizado em um equipamento mais sensível pode detectar valores abaixo de k_j, mesmo que outro equipamento mais limitado os considerasse censurados. Assim, as regras para $y_i > k_1$ (sem censura) e $y_i \leq k_m$ (censura no nível mais baixo) são sempre válidas, enquanto os casos intermediários podem depender de fatores adicionais relacionados à sensibilidade da medição.

Por exemplo, considerando $m = 3$ níveis de censura com $k_1 = 400$, $k_2 = 50$ e $k_3 = 40$, a matriz $\mathbf{C}$ para cinco observações y_i poderia ser:

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

onde:

- A primeira e a quinta observações não são censuradas.
- A segunda observação é censurada no nível $k_1 = 400$.
- A terceira observação é censurada no nível $k_2 = 50$.
- A quarta observação é censurada no nível $k_3 = 40$.

Essa estrutura assegura que, para cada observação i, ou todos os elementos de $\mathbf{c}_i$ são zero (indicando ausência de censura), ou exatamente um dos elementos é 1 (indicando censura em um nível específico), garantindo que uma observação não esteja associada a múltiplos níveis de censura simultaneamente.

Adicionalmente, estabelece-se a restrição de que, para cada observação i, ou todos os elementos de $\mathbf{c}_i$ são iguais a zero (isto é, não há censura), ou exatamente um dos elementos c_{ij} é igual a 1 (indicando censura no nível k_j). Portanto, não é possível que uma mesma observação esteja associada simultaneamente a mais de um nível de censura.

É importante destacar que o objeto de censura indica apenas a presença (ou ausência) de censura em determinado nível, diferentemente do conceito apresentado na Seção 2.1, que define o tempo até a falha. Essa distinção exigirá adaptações na construção da função de verossimilhança apresentada na Seção 3.2.

Da mesma forma em que foi definida anteriormente o caso de censura à esquerda com um único nível de censura, também pode-se definir a mesma, com a utilização de múltiplos níveis.

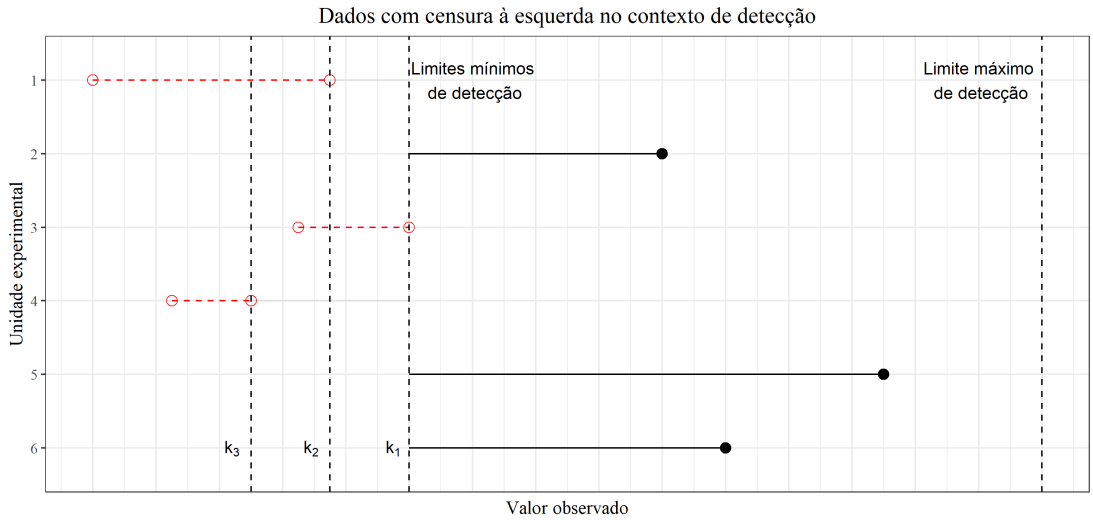

Figura 5 – Ilustração da censura à esquerda, sendo ● a falha e ○ censura.

Note na Figura 5 que o valor real da variável é menor do que o valor que realmente foi observado, dado pelo limites mínimos de detecção. Visto que neste caso, os equipamentos não são capazes de retornar o valor exato, pois o valor muito baixo para o equipamento conseguir mensurar, além disto como pode ser feito em equipamento diferentes, os LMD podem variar. Outro ponto é que em outros casos envolvendo testes de carga viral, pode acontecer do valor exceder ao limite máximo, caracterizando novamente por uma censura a direita, porém não serão tratados neste trabalho.

3.2 Modelo de Regressão Paramétrico com m Níveis de Censura

Uma vez definido o objeto principal deste trabalho - os múltiplos níveis de censura - nosso objetivo é construir uma função de verossimilhança que incorpore adequadamente essas

diferentes situações: tanto os casos sem censura quanto aqueles com diferentes níveis de censura nos dados.

- **Caso de um nível de censura** $m = 1$

Começamos pelo caso mais simples com um único nível de censura $k = [k_1]$. Neste cenário, cada observação pode ser classificada como:"

- Não censurada, com $c_i = 0$;
- Censurada à esquerda no ponto k_1, com $c_i = 1$.

A função de verossimilhança pode, então, ser definida como:

$$L(\theta \mid \mathbf{y}, k_1, \mathbf{C}) = \prod_{i=1}^{n} \left[f(y_i \mid \theta)^{1-c_i} \cdot P(y_i < k_1 \mid \theta)^{c_i} \right]$$

- **Caso de dois níveis de censura** $m = 2$

Considerenado dois níveis de censura $m = 2$, ou seja, dois pontos de censura, k_1 e k_2. Cada observação pode assumir uma das seguintes classificações:

- Não censurada: $c_{i1} = 0$, $c_{i2} = 0$;
- Censurada em k_1: $c_{i1} = 1$, $c_{i2} = 0$;
- Censurada em k_2: $c_{i1} = 0$, $c_{i2} = 1$.

A verossimilhança, então, pode ser escrita como:

$$L(\theta \mid \mathbf{y}, k_1, k_2, \mathbf{C}) = \prod_{i=1}^{n} \left[f(y_i \mid \theta)^{(1-c_{i1}-c_{i2})} \cdot P(y_i < k_1 \mid \theta)^{c_{i1}} \cdot P(y_i < k_2 \mid \theta)^{c_{i2}} \right]$$

- **Caso de** k **níveis de censura** $m = k$

De forma geral, a função de verossimilhança do modelo estudado pode ser descrita como o produto das contribuições das n observações da seguinte forma:

$$L(\theta \mid \mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C}) = \prod_{i=1}^{n} \left[f(y_i \mid \theta)^{\left(1-\sum_{j=1}^{m} c_{ij}\right)} \times \prod_{j=1}^{m} P(y_i < k_j \mid \theta)^{c_{ij}} \right]. \tag{3.1}$$

Neste contexto, de acordo com a Equação 3.1, é possível separar a contribuição de cada indivíduo, $i = 1, \ldots, n$, em dois casos distintos. O primeiro corresponde ao caso em que se tem uma informação completa, dado por:

$$f(y_i \mid \theta)^{\left(1-\Sigma_{j=1}^{m} c_{ij}\right)}, \tag{3.2}$$

quando $c_{ij} = 0 \quad \forall j$, com i fixado, a contribuição na verossimilhança é dada pela função densidade de probabilidade (f.d.p.). Em outras palavras, no caso em que não exista nenhum nível de censura para determinado indivíduo, sua contribuição é dada pela f.d.p.

Quando existe a presença de algum nível de censura, e a contribuição é dada por:

$$\prod_{j=1}^{m} P(y_i < k_j \mid \theta)^{c_{ij}}, \tag{3.3}$$

quando existe $c_{ij} \neq 0$, com i fixado, a contribuição na verossimilhança é dada pela função de distribuição acumulada (f.d.a.) no nível de censura k_j. Em palavras, quando há algum nível de censura, sua contribuição será dada pela f.d.a. correspondente ao ponto de censura.

3.2.1 Processo de Estimação do Vetor de Parâmetros θ

Dado a definição da função de verossimilhança para o modelo com múltiplos níveis de censura, $L(\theta \mid \mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C})$, definida pela Equação 3.1. De acordo com (CASELLA, 2002) o princípio da máxima verossimilhança dado um conjunto fixo de específico de variáveis explicativas, consiste em encontrar um valor para θ que maximize a função de verossimilhança.

Formalmente, o estimador de máxima verossimilhança (EMV) para $\hat{\theta}$ é definido como:

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C}).$$

No entanto, em práticas computacionais, maximiza-se a função log-verossimilhança $\ell(\theta|y, X, k, C) = \log L(\theta|y, X, k, C)$, que preserva o ponto de máximo e facilita os cálculos, pois transforma o produto das probabilidades em uma soma de logaritmos (CASELLA, 2002).

Primeiras Derivadas (Equações de Score)

As primeiras derivadas da função de log-verossimilhança em relação aos parâmetros θ são conhecidas como equações de score. Estas equações são fundamentais para a identificação os pontos de inclinação iguais a zero, indicando possíveis máximo, mínimos ou pontos de sela. Sendo estes, candidatos a estimadores de máxima verossimilhança. Matematicamente, as equações de score são dadas por:

$$U(\theta) = \frac{\partial \ell(\theta|\mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbb{C})}{\partial \theta} = 0.$$

Resolver este sistema de equações simultâneas para θ fornece os EMVs. A condição de primeira ordem, $U(\theta) = 0$, é necessária para maximizar $\ell(\theta)$, mas não suficiente, exigindo uma análise adicional das segundas derivadas (CASELLA, 2002).

Segundas Derivadas (Matriz Hessiana)

A matriz das segundas derivadas da função de log-verossimilhança em relação aos parâmetros θ é chamada de matriz Hessiana. A Hessiana avalia a curvatura da log-verossimilhança em torno dos EMVs e é utilizada para verificar se o ponto encontrado pelas equações de score corresponde a um máximo local (CASELLA, 2002). Se a Hessiana for negativa definida no ponto estimado, podemos concluir que os valores de θ encontrados são de fato máximos locais. Logo a matriz Hessiana é definida como:

$$H(\theta) = \frac{\partial^2 \ell(\theta|\mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C})}{\partial\theta\partial\theta'},$$

em que $H(\theta)$ é uma matriz simétrica de ordem $p \times p$ (onde p é o número de parâmetros em θ). Para um vetor de parâmetros θ, a Hessiana avaliada em $\hat{\theta}$ desempenha um papel importante na estimação da variância-covariância dos estimadores:

$$\text{Var}(\hat{\theta}) \approx -H(\hat{\theta})^{-1}.$$

Essa inversa da Hessiana fornece uma aproximação da matriz de covariância dos estimadores de θ, sendo essencial para a construção de intervalos de confiança e para a realização de testes de hipóteses.

Em resumo, as derivadas primeiras e segundas da log-verossimilhança são ferramentas cruciais na teoria da máxima verossimilhança, permitindo não apenas a estimação pontual dos parâmetros, mas também a avaliação de sua precisão e a realização de inferências robustas.

3.3 Modelo de regressão Weibull com *m* níveis de censura

Sejam n observações coletadas da variável dependente, representadas pelo vetor $\mathbf{y} = (y_1, \ldots, y_n)$, assumindo-se independência entre as respostas. Essas observações são amostras de uma variável aleatória Y, associada à quantidade de interesse em cada unidade amostral.

$$\begin{aligned} &Y_i \sim Weibull(\gamma, \alpha_i); \\ &\alpha_i = \exp(x_i\beta) \Rightarrow \ln \alpha_i = x_i\beta. \end{aligned}$$

Considerando um conjunto de p variáveis independentes, é possível representar as covariáveis por meio da matriz $\mathbf{X}_{nxp} \in \mathbb{R}^{n\times p}$, definida como

$$\mathbf{X}_{nxp} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

em que cada vetor linha $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}] \in \mathbb{R}^{1\times p}$ representa o vetor de covariáveis associado ao i-ésimo indivíduo, para $i = 1, \ldots, n$.

Utilizando a função de verossimilhança definida na Equação 3.1, juntamente com a função densidade de probabilidade (f.d.p.) e a função de distribuição acumulada (f.d.a.) apresentadas da distribuição Weibull como apresentadas na Subseção 2.2.1, tem-se a seguinte função de verossimilhança do modelo Weibull com múltiplos níveis de censura

$$L_i(\beta, \gamma | y_i, x_i, k, c_{ij}) = \left\{ \gamma y_i^{\gamma-1} \left(e^{x_i\beta}\right)^{-\gamma} \exp\left[-\left(y_i e^{-x_i\beta} \right)^{\gamma} \right] \right\}^{1-\sum_{j=1}^{m} c_{ij}} \times \prod_{j=1}^{m} \left\{ 1 - \exp\left[-\left(k_j e^{-x_i\beta} \right)^{\gamma} \right] \right\}^{c_{ij}}. \tag{3.4}$$

assim, a função de verossimilhança pode ser descrita como um produto das contribuições das n observações da seguinte forma:

$$L(\beta, \gamma | \mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C}) = \prod_{i=1}^{n} \left[\left\{ \gamma y_i^{\gamma-1} \left(e^{x_i'\beta}\right)^{-\gamma} \exp\left[-\left(y_i e^{-x_i'\beta} \right)^{\gamma} \right] \right\}^{1-\sum_{j=1}^{m} c_{ij}} \times \prod_{j=1}^{m} \left\{ 1 - \exp\left[-\left(k_j e^{-x_i'\beta} \right)^{\gamma} \right] \right\}^{c_{ij}} \right].$$

O função da log-verossimilhança é dada por

$$\ell(\beta, \gamma | \mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C}) = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij} \right) \left[\log(\gamma) + (\gamma - 1)\log(y_i) - \gamma x_i'\beta - \left(y_i e^{-x_i'\beta} \right)^{\gamma} \right] + \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \log\left(1 - \exp\left[-\left(k_j e^{-x_i'\beta} \right)^{\gamma} \right] \right).$$

Para realizar a estimação dos parâmetros do modelo β e γ através do método de máxima verossimilhança, assim como definido em 3.2 é necessário calcular as primeiras e segundas derivadas da função de log-verossimilhança.

Derivada Primeira

As derivadas primeiras, ou funções score, são calculadas separadamente em relação aos parâmetros β e γ.

Derivada em Relação a β

$$\frac{\partial \ell(\beta,\gamma)}{\partial \beta} = \sum_{i=1}^{n}\left(1-\sum_{j=1}^{m} c_{ij}\right)\left[-\gamma x_i + \gamma y_i^{\gamma} e^{-\gamma x_i'\beta} x_i\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\left[\frac{-\gamma k_j^{\gamma} e^{-\gamma x_i'\beta} x_i}{1-\exp\left[-\left(k_j e^{-x_i'\beta}\right)^{\gamma}\right]}\right].$$

Derivada em Relação a γ

$$\frac{\partial \ell(\beta,\gamma)}{\partial \gamma} = \sum_{i=1}^{n}\left(1-\sum_{j=1}^{m} c_{ij}\right)\left[\frac{1}{\gamma} + \log(y_i) - x_i'\beta - \log\left(y_i e^{-x_i'\beta}\right)\left(y_i e^{-x_i'\beta}\right)^{\gamma}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\frac{\log\left(k_j e^{-x_i'\beta}\right)\left(k_j e^{-x_i'\beta}\right)^{\gamma}}{1-\exp\left[-\left(k_j e^{-x_i'\beta}\right)^{\gamma}\right]}.$$

Hessiana (Matriz de Segundas Derivadas)

Assim como comentado na 3.2, a matriz Hessiana, composta pelas segundas derivadas parciais, fornece informações sobre a curvatura da log-verossimilhança e é fundamental para verificar se as soluções encontradas pelas derivadas primeiras são máximos locais. A Hessiana é dada por:

$$H(\beta,\gamma) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta^2} & \frac{\partial^2 \ell}{\partial \beta \partial \gamma} \\ \frac{\partial^2 \ell}{\partial \gamma \partial \beta} & \frac{\partial^2 \ell}{\partial \gamma^2}. \end{pmatrix}$$

Segunda Derivada em Relação a β

$$\frac{\partial^2 \ell}{\partial \beta^2} = \sum_{i=1}^{n}\left(1-\sum_{j=1}^{m} c_{ij}\right)\gamma y_i^{\gamma} e^{-\gamma x_i'\beta} x_i x_i' + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\frac{\gamma^2 k_j^{\gamma} e^{-\gamma x_i'\beta} x_i x_i'}{\left(1-\exp\left[-\left(k_j e^{-x_i'\beta}\right)^{\gamma}\right]\right)^2}.$$

Segunda Derivada em Relação a γ

$$\frac{\partial^2 \ell}{\partial \gamma^2} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[-\frac{1}{\gamma^2} - \left(\log(y_i e^{-x_i'\beta})\right)^2 \left(y_i e^{-x_i'\beta}\right)^{\gamma}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \frac{\left(\log(k_j e^{-x_i'\beta})\right)^2 \left(k_j e^{-x_i'\beta}\right)^{\gamma}}{1 - \exp\left[-\left(k_j e^{-x_i'\beta}\right)^{\gamma}\right]}.$$

Derivada Cruzada $\frac{\partial^2 \ell}{\partial \beta \partial \gamma}$

$$\frac{\partial^2 \ell}{\partial \beta \partial \gamma} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[x_i \left(y_i^{\gamma} e^{-\gamma x_i'\beta} \log\left(y_i e^{-x_i'\beta}\right)\right)\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \frac{\gamma k_j^{\gamma} \log(k_j e^{-x_i'\beta}) e^{-\gamma x_i'\beta} x_i}{1 - \exp\left[-\left(k_j e^{-x_i'\beta}\right)^{\gamma}\right]}.$$

Assim, as derivadas primeiras nos dão a direção do gradiente para a maximização da log-verossimilhança, e as segundas derivadas (Hessiana) nos fornecem uma visão mais profunda da curvatura da função, essencial para métodos de otimização e estimação de parâmetros.

3.4 Modelo de regressão Log-Normal com *m* níveis de censura

Sejam n observações coletadas da variável dependente, representadas pelo vetor $\mathbf{y} = (y_1, \ldots, y_n)$, assumindo-se independência entre as respostas. Essas observações são amostras de uma variável aleatória Y, associada à quantidade de interesse em cada unidade amostral.

$$\begin{aligned} &Y_i \sim LogNormal(\mu_i, \sigma_i); \\ &\mu_i = x_i \beta_\mu; \\ &\sigma_i = exp(x_i \beta_\sigma). \end{aligned}$$

Diferentemente da modelagem realizada com a distribuição Weibull, onde as covariáveis foram incorporadas apenas ao parâmetro de escala, optou-se por incluir variáveis explicativas em ambos os parâmetros da distribuição LogNormal, ou seja, tanto na média μ_i quanto na escala σ_i. Essa decisão foi motivada pela observação de uma variabilidade expressiva entre os diferentes grupos nos dados.

Ao permitir que o parâmetro σ_i varie de acordo com covariáveis, busca-se investigar se essa variabilidade refletida nas diferentes dispersões observadas entre os grupos contribui para melhorar o ajuste do modelo. Em outras palavras, deseja-se avaliar se há evidências de que a heterogeneidade dos dados não se restringe apenas à média, mas também afeta a dispersão das respostas, justificando, assim, uma modelagem mais flexível e informativa.

Considerando um conjunto de p variáveis independentes, é possível representar as covariáveis por meio da matriz $\mathbf{X}_{nxp} \in \mathbb{R}^{n\times p}$, definida como

$$\mathbf{X}_{nxp} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

em que cada vetor linha $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}] \in \mathbb{R}^{1\times p}$ representa o vetor de covariáveis associado ao i-ésimo indivíduo, para $i = 1, \ldots, n$.

Considerando a função de verossimilhança definida na Equação 3.1, juntamente com a função densidade de probabilidade (f.d.p.) e a função de distribuição acumulada (f.d.a.) apresentadas da distribuição Weibull como apresentadas na Subseção 2.2.2, tem-se a seguinte função de verossimilhança do modelo Weibull com múltiplos níveis de censura

$$L_i(\beta_\mu, \beta_\sigma | y_i, x_i, k, c_{ij}) = \left\{ \frac{1}{y_i \exp(x_i\beta_\sigma)\sqrt{2\pi}} \exp\left(-\frac{(\log y_i - x_i\beta_\mu)^2}{2\exp(2x_i\beta_\sigma)} \right) \right\}^{1-\sum_{j=1}^{m} c_{ij}} \times \prod_{j=1}^{m} \left\{ \Phi\left(\frac{\log y_i - x_i\beta_\mu}{\exp(x_i\beta_\sigma)} \right) \right\}^{c_{ij}},$$

em que $\Phi(\dot{})$ é dada pela função de distribuição acumulada da normal padrão, assim, a função de verossimilhança pode ser descrita como um produto das contribuições das n observações da seguinte forma:

$$L(\beta_\mu, \beta_\sigma | \mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C}) = \prod_{i=1}^{n} \left\{ \left[\frac{1}{y_i \exp(x_i'\beta_\sigma)\sqrt{2\pi}} \exp\left(-\frac{(\log y_i - x_i'\beta_\mu)^2}{2\exp(2x_i'\beta_\sigma)} \right) \right]^{1-\sum_{j=1}^{m} c_{ij}} \times \prod_{j=1}^{m} \left[\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)} \right) \right]^{c_{ij}} \right\}. \quad (3.5)$$

O logaritmo da função de verossimilhança é

$$l(\beta_\mu, \beta_\sigma | \mathbf{y}, \mathbf{X}, \mathbf{k}, \mathbf{C}) = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[-\log y_i - x_i'\beta_\sigma - \frac{1}{2}\log(2\pi) - \frac{(\log y_i - x_i'\beta_\mu)^2}{2\exp(2x_i'\beta_\sigma)}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \log\left[\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)\right].$$

Para estimar os parâmetros β_μ e β_σ através do método de máxima verossimilhança, assim como definido na 3.2 é necessário calcular as primeiras e segundas derivadas da função de log-verossimilhança.

Derivada Primeira

A motivação para calcular a primeira derivada da função de verossimilhança em relação aos vetores β_μ e β_σ é encontrar os estimadores de máxima verossimilhança (EMV) para esses parâmetros. Os EMV são os valores que maximizam a função de verossimilhança, correspondendo aos parâmetros que tornam os dados observados mais prováveis, dados o modelo.

Derivada em relação a β_μ

$$\frac{\partial \ell}{\partial \beta_\mu} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[\frac{(\log y_i - x_i'\beta_\mu)x_i}{\exp(2x_i'\beta_\sigma)}\right] - \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \left[\frac{\phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right) x_i}{\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)\exp(x_i'\beta_\sigma)}\right].$$

em que $\phi(\dot{})$ é dada pela função de probabilidade da normal padrão.

Derivada em relação a β_σ

$$\frac{\partial \ell}{\partial \beta_\sigma} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[-x_i - \frac{(\log y_i - x_i'\beta_\mu)^2 x_i \exp(-2x_i'\beta_\sigma)}{\exp(2x_i'\beta_\sigma)}\right] - \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \left[\frac{\phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)(\log y_i - x_i'\beta_\mu)x_i \exp(-x_i'\beta_\sigma)}{\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)\exp(2x_i'\beta_\sigma)}\right].$$

Matriz hessiana

Assim como comentado na 3.2, a matriz Hessiana, composta pelas segundas derivadas parciais, fornece informações sobre a curvatura da log-verossimilhança e é fundamental para verificar se as soluções encontradas pelas derivadas primeiras são máximos locais. A Hessiana é dada por:

Derivada Segunda em relação a β_μ

$$\frac{\partial^2 \ell}{\partial \beta_\mu^2} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[-\frac{x_i x_i'}{\exp(2x_i'\beta_\sigma)}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \left[\frac{\partial}{\partial \beta_\mu}\left(\frac{\phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right) x_i}{\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)\exp(x_i'\beta_\sigma)}\right)\right].$$

Derivada Segunda em relação a β_σ

$$\frac{\partial^2 \ell}{\partial \beta_\sigma^2} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[-\frac{x_i x_i'(\log y_i - x_i'\beta_\mu)^2 \exp(-4x_i'\beta_\sigma)}{\exp(2x_i'\beta_\sigma)}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \left[\frac{\partial}{\partial \beta_\sigma}\left(\frac{\phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)(\log y_i - x_i'\beta_\mu) x_i \exp(-x_i'\beta_\sigma)}{\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right)\exp(2x_i'\beta_\sigma)}\right)\right].$$

Derivada Cruzada entre β_μ *e* β_σ

$$\frac{\partial^2 \ell}{\partial \beta_\mu \partial \beta_\sigma} = \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{m} c_{ij}\right) \left[-\frac{(\log y_i - x_i'\beta_\mu) x_i x_i' \exp(-2x_i'\beta_\sigma)}{\exp(2x_i'\beta_\sigma)}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \left[\frac{\partial}{\partial \beta_\sigma}\left(\frac{\phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i'\beta_\sigma)}\right) x_i}{\Phi\left(\frac{\log y_i - x_i'\beta_\mu}{\exp(x_i' \; beta_\sigma)}\right)\exp(x_i'\beta_\sigma)}\right)\right].$$

3.5 Adequação do modelo ajustado

Avaliar se o modelo proposto é adequado é uma das parte cruciais ao se realizar uma modelagem de dados, existem várias técnicas para se realizar esta validação, assim como técnicas gráficas que fazem usos de diferentes resíduos para examinar aspectos algumas suposições do modelo Colosimo e Giolo (2006b).

Para modelos com a incorporação de censura, existem várias resíduos que buscam realizar um estudo da adequabilidade das suposições, em Colosimo e Giolo (2006b) são apresentados alguns como Cox-Snell, Martingale e Deviance. Podemos então, como é o caso deste trabalho, verificar se é adequado considerar se as distribuição Weibull ou Log-normal são adequadas para a modelagem através da análise residual.

Ademais, os resíduos citados acima apresentam certas limitações, o que dificulta a verificação da adequação do modelo com base em resíduos tradicionais. Diante disso, a utilização dos Resíduos de Probabilidade de Sobrevivência Aleatorizada Transformados pela Normal

mostra-se uma alternativa viável, dada sua fácil implementação, uma vez que requer apenas o cálculo da função de sobrevivência proposta para a variável resposta (WU; FENG; LI, 2019).

3.5.1 Resíduo de Probabilidade de Sobrevivência Aleatório Transformado pela Normal (NRSP)

Em Wu, Feng e Li (2019), os autores definem os resíduos NRSP aleatorizando as probabilidades de sobrevivência para as observações censuradas com base em uma distribuição uniforme no intervalo $[0, S_i(Y_i)]$. Em suma, a sua aplicação é realizada aplicando um fator de redução nas probabilidades de sobrevivência com base em uma variável aleatória uniforme padrão no intervalo (0, 1). Deste modo, pode-se definir a probabilidade de sobrevivência aleatoriezada (RSP) para Y_i pela seguinte equação:

$$S_i^R(Y_i, d_i, u_i) = \begin{cases} S_i(Y_i), & \text{se } Y_i \text{ não é censurado,} \\ U_i S_i(Y_i), & \text{se } Y_i \text{ é censurado,} \end{cases} \tag{3.6}$$

em que u_i é uma variável aleatória uniforme no intervalo (0, 1) e $S_i(Y_i)$ é a função de sobrevivência da distribuição que temos interesse. Wu, Feng e Li (2019) demonstra que os RSP, $S_i^R(T_i, d_i, U_i)$, são uniformemente distribuídos em (0, 1) quando a suposição de adequação existe. Consequentemente, pelo método da inversa podemos obter outra distribuição. Para isso então, é aplicado a função inversa da distribuição acumulada da normal:

$$r_i^{RSP}(Y_i, d_i, u_i) = \Phi^{-1}(S_i^R(Y_i, d_i, U_i)), \tag{3.7}$$

em que $\Phi^{-1}(x)$ é a função quantílica da distribuição normal padrão.

Chamando então os resíduos 3.7 de resíduos de probabilidade de sobrevivência aleatorizado transformados pela normal, (ou da sigla em inglês (*Normal-transformed Randomized Survival Probability* (NRSP)). Uma das vantagens na transformação dos resíduos RSP para NRSP, é que sob hipótese de que o modelo seja adequado, estes serão normalmente distribuídos, tendo a vantagem que existe uma ampla gama de testes na literatura Wu, Feng e Li (2019).

Contudo para a aplicação neste relatório, como o intuito é a aplicação para dados com censura à esquerda, não é muito difícil adaptar esta metodologia para se utilizar a função acumulada no lugar da função de sobrevivência, ainda teremos que sob condições de bom ajuste. Primeiramente, reescrevemos $F_i^R(T_i^*, C_i, U_i)$ como uma função de (T_i^*, C_i, U_i) da seguinte forma:

$$F_i^R(T_i^*, C_i, U_i) = \begin{cases} U_i F_i(C_i), & \text{se } T_i^* \leq C_i, \\ F_i(T_i^*), & \text{se } T_i^* > C_i. \end{cases}$$

Aqui:

- Quando $T_i^* \leq C_i$, o valor censurado é representado por $U_i F_i(C_i)$, onde U_i é uniforme em $[0,1]$, escalado para o intervalo $[0, F_i(C_i)]$.

- Quando $T_i^* > C_i$, o pseudovalor é dado por $F_i(T_i^*)$, que segue a distribuição acumulada padrão para valores não censurados.

Assume-se que T_i^* e C_i são independentes, ou seja, os tempos de censura não são informativos em relação aos tempos de falha originais. Logo, o objetivo então é provar que $F_i^R(T_i^*, C_i, U_i)$ dado que $C_i = c$ é uniforme no intervalo $(0,1)$. Assim para demonstrar é necessário definir alguns casos, o primeiro sendo quando temos $T_i^* \leq C_i$, o pseudovalor é dado por $U_i F_i(C_i)$, onde U_i é uniforme em $[0,1]$. Isso implica que $U_i F_i(C_i)$ será uniformemente distribuído no intervalo $[0, F_i(C_i)]$. E dado $T_i^* > C_i$ o pseudovalor é $F_i(T_i^*)$, e sabe-se que $T_i^* > C_i$ implica $F_i(T_i^*) \in [F_i(C_i), 1]$. Assim, $F_i(T_i^*)$ é uniformemente distribuído no intervalo $[F_i(C_i), 1]$. Além disso, denota-se $\lambda(B)$ como o comprimento de um intervalo B em $(0,1)$. Com base nos casos acima, a probabilidade condicional para $F_i^R(T_i^*, C_i, U_i) \in B$, dado $C_i = c$, é:

$$P(F_i^R(T_i^*, C_i, U_i) \in B \mid C_i = c) = P(U_i F_i(C_i) \in B \mid C_i = c, T_i^* \leq c) \cdot P(T_i^* \leq c)$$

$$+P(F_i(T_i^*) \in B \mid C_i = c, T_i^* > c) \cdot P(T_i^* > c).$$

Substituindo os termos:

- Para $T_i^* \leq c$, $U_i F_i(C_i)$ está em $[0, F_i(c)]$, e a probabilidade é proporcional a

$$\lambda(B \cap [0, F_i(c)]) \cdot F_i(c).$$

- Para $T_i^* > c$, $F_i(T_i^*)$ está em $[F_i(c), 1]$, e a probabilidade é proporcional a

$$\lambda(B \cap [F_i(c), 1]) \cdot (1 - F_i(c)).$$

Logo, tem-se:

$$P(F_i^R(T_i^*, C_i, U_i) \in B \mid C_i = c) = \lambda(B \cap [0, F_i(c)]) \cdot F_i(c) + \lambda(B \cap [F_i(c), 1]) \cdot (1 - F_i(c)).$$

Como $F_i(c) + (1 - F_i(c)) = 1$, a probabilidade total é:

$$P(F_i^R(T_i^*, C_i, U_i) \in B \mid C_i = c) = \lambda(B).$$

Concluímos que com a censura à esquerda, a distribuição condicional de $F_i^R(T_i^*, C_i, U_i)$, dado $C_i = c$, é uniforme no intervalo $(0,1)$. Isso demonstra que o pseudovalor F_i^R, baseado na função acumulada $F_i(\cdot)$, satisfaz a uniformidade, mesmo sob censura à esquerda. Logo se F_i^R sob hipótese de que o modelo seja adequado é uniformemente distruído no intervalo $(0,1)$, também é válido aplicar a função quantílica da normal, igual feito para quando há o uso da função sobrevivência para os casos de censura a direita.

3.6 Estudo de Simulação

Nesta capítulo conduzimos um estudo de simulação para investigar a consistência e eficiência dos EMVs que podem ser obtidos através das equações com base em diferentes tamanhos amostrais. Para tanto utizamos 3 cenários e três critérios: o viés, raiz quadrada do erro quadrático médio (EQM) e a probabilidade de cobertura (PC), os quais são dados, respectivamente, por:

$$\text{Vies}\left(\widehat{\theta}_w\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_w^{(m)} - \theta_w\right);$$

$$\text{REQM}\left(\widehat{\theta}_w\right) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_w^{(m)} - \theta_w\right)^2}$$

e

$$\text{PC}\left(\widehat{\theta}_w\right) = \frac{1}{M}\sum_{m=1}^{M} 1\left(\theta_w \in \widehat{\theta}_w \pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{\theta}_w)}\right)$$

para $w = 1,\ldots,\kappa$, em que M é o numero de replicações Monte Carlo e $\theta = (\theta_1,\ldots,\theta_\kappa)$ representa o vetor de parâmetros. Entretanto, $\widehat{\theta}_w^{(m)}$ denota o EMV de θ_w obtida da amostra m, para $m = 1,\ldots,M$. Por meio de 1000 amostras *bootstrap*.

Por esta abordagem, espera-se que bons estimadores tenham viés e REQM próximos de zero. Por sua vez, os intervalos de confiança razoáveis, que são obtidos aqui usando a normalidade assintótica dos EMVs, devem ter amplitude pequena e com probabilidade de cobertura próxima ao valor nominal de 95%. Neste trabalho, todos os cálculos e simulações foram realizados no software R (*R Core Team*, 2025).

De forma geral, pode-se definir o processo de simulação da seguinte maneira para uma distribuição com dois parâmetros. Seja n observações coletadas da variável dependente $y_1 = (y_1,\ldots,y_i,\ldots,y_n)$, assumindo independência entre as repostas, sendo elas amostradas de uma variável aleatória qualquer $Y_i \sim Dist(\alpha_{1,i}, \alpha_{2,i})$, considera-se o conjunto p de variáveis independentes, deste modo é possível expressar as covariáveis por meio da matrix $X_p = [x_1,\ldots,x_i,\ldots,x_n]$, sendo $x_i = [x_{i1}, x_{i2},\ldots,x_{ip}]$ para $i = 1,\ldots,n$, por fim, assumi-se $g(.)$ e $h(.)$ duas funções de ligações que satisfaça os espaços amostrais respectivamente de $\alpha_{1,i}$ e $\alpha_{2,i}$, temos então:

$$
\begin{aligned}
Y_i &\sim Dist(\alpha_{1,i}, \alpha_{2,i}) \\
\alpha_{1,i} &= g(x_i\beta_{\alpha_1}); \\
\alpha_{2,i} &= h(x_i\beta_{\alpha_2});.
\end{aligned}
$$

sendo β_{α_1} e β_{α_2} 2 vetores de parâmetros:

O seguinte algoritmo será utilizado para a geração de observações para o modelo:

1. Estabeleça valores para os vetores de parâmetros do modelo β_{α_1} e β_{α_2};
2. Fixe m distintos níveis de censura $k_1 = [k_1, k_2, \ldots, k_m]$, considerando que $k_1 > k_2 > \cdots > k_m$;
3. Simule o conjunto de variáveis $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$;
4. Gere y_i distribuído da sua função de interesse com parâmetros $g(x_i\beta_{\alpha_1})$ e $h(x_i\beta_{\alpha_2})$;
5. Gere u_i distribuido uniformemente no intervalo (0,1);
6. Caso $y_i > k_1$, assuma $y_i = y_i$;
7. Caso $y_i \leq k_m$, assuma $y_i = k_m$;
8. Para $2 < j < m$ e $k_{j-1} \leq y_i < k_j$
   8.1. Assuma $y_i = k_{j-1}$, caso $u_i < 0.5$;
   8.2. Assuma $y_i = y_i$. caso contrário.

Do modo como foi criado, a nossa amostra garante que todo valor que for menor que o menor nível de censura k_m, seja dado como censurado, já para valores entre níveis de censura distintos, ele somente sera censurado, caso a probabilidade gerada pela Bernoulli for maior que 0.5. Isto garante que a amostra gerada seja fidedigna aos dados de estudo, dado que a existência de equipamentos diferentes, faz com que por exemplo o LMD para um equipamento, não seja para outro, porém o valor mínimo de detecção de todos é um valor comum fixado até que haja uma evolução com relação as tecnologias utilizadas.

3.6.1 Estudo de simulação do modelo de regressão Weibull

Para avaliar a performance do estimador sob os níveis de censura para o Modelo de regressão Weibull com m níveis de censura, foi realizado um estudo de simulação utilizando diferentes tamanhos amostrais de $n = 50$, $n = 100$, $n = 250$ e $n = 500$, com isto a variável neste processo é dada por:

$$Y_i \sim Weibull(\gamma, \alpha_i)$$
$$\alpha_i = \exp(x_i\beta) \Rightarrow \ln \alpha_i = x_i\beta;$$

logo, como só é possível a incorporação de co-variáveis no α_i e neste primeiro estudo de simulação foi escolhido pela não utilização de co-variáveis, o vetor β é dado apenar pelo intercepto.

Com isso, consideraram-se os seguintes cenários para verificar se as propriedades frequentistas eram satisfeita

- **Cenário 1:** 1 nível de censura $m = 1$, considerando o quantil de 5% da distribuição Weibull;

- **Cenário 2:** 2 níveis de censura $m = 2$, considerando os quantis de 5% e 2.5% da distribuição Weibull;

- **Cenário 3:** 3 níveis de censura $m = 3$, considerando os quantis de 5%, 2.5% e 1% da distribuição Weibull.

Para todos os cenários, foram simuladas 1000 amostras, sendo os dados gerados de uma Weibull($\alpha = 15$, $\gamma = \exp(\beta_0 = 2)$), sendo assim, é possível obter as estimativas dos parâmetros, e por fim os critérios.

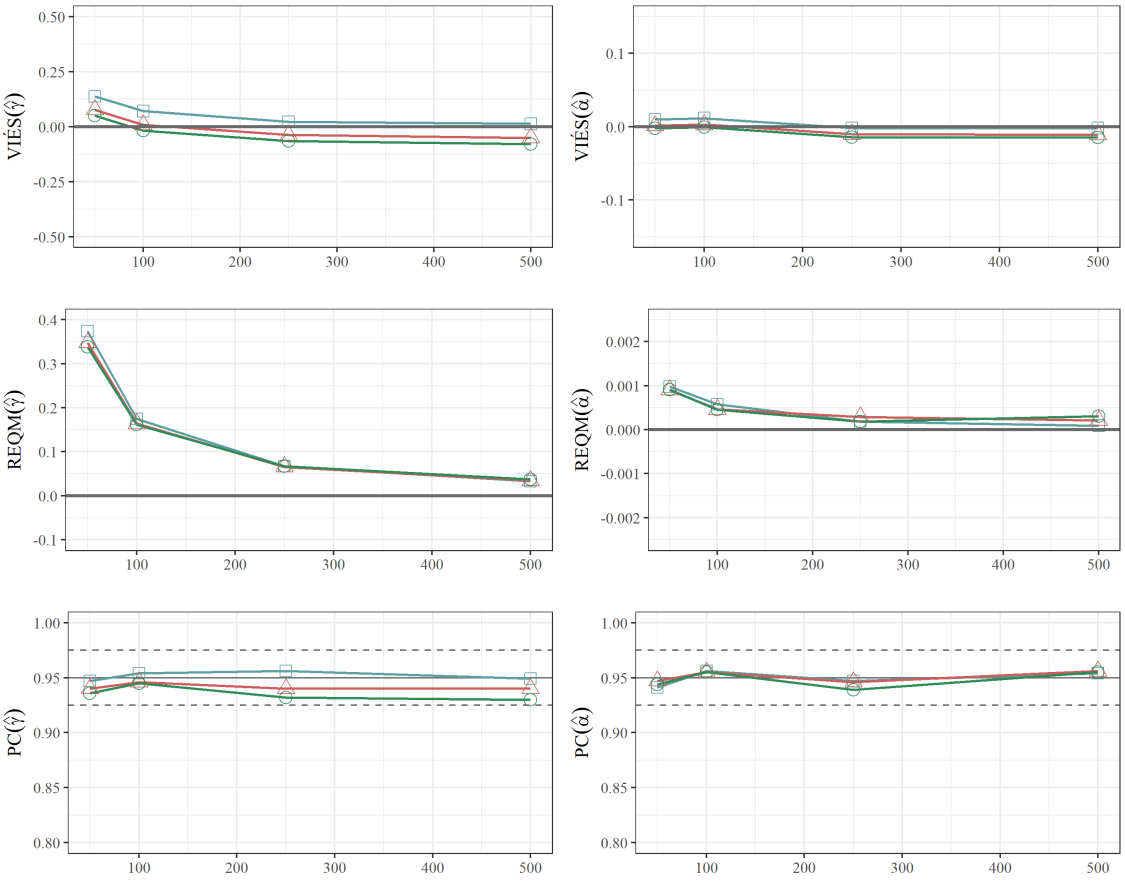

Figura 6 – Viés, raiz quadrada do erro quadrático médio e probabilidade de cobertura (CP) do estimador de máxima verossimilhança de $(\hat{\alpha}, \hat{\gamma})$ do modelo de múltiplos níveis de censura Weibull utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

A Figura 6 mostra o resultado do estudo de simulação considerando, em que cenário 1 (■), cenário 2 (▲) e cenário 3 (●). Para os resultados, serão mostrados as métricas em cima de *gamma*, e como foi escolhido o valor para o intercepto no processo de simulação, foi realizado o cálculo da variância deste parâmetro via método delta .Com isto, nota-se que o viés e o REQM diminuem conforme o tamanho da amostra aumenta, indicando que os estimadores se tornam mais precisos com mais dados, assim como a probabilidade de cobertura dos parâmetros chegaram próximas de 95%, já que este foi o nível de significância escolhido para simularmos. Para acessar mais informações sobre os valores exatos, acessar Apêndice A.

3.6.2 Estudo de simulação do modelo de regressão Log-Normal

Para avaliar a performance do estimador sob os níveis de censura, foi realizado um estudo de simulação utilizando diferentes tamanhos amostrais de $n = 50$, $n = 100$, $n = 250$ e $n = 500$, com a incorporação de duas variáveis, supondo distribuição normal padrão para ambas além da utilização do intercepto. Com isto a variável neste processo é dada por:

$$
\begin{aligned}
Y_i &\sim LogNormal(\mu_i, \sigma_i) \\
\mu_i &= x_i \beta_\mu; \\
\sigma_i &= \exp(x_i \beta_\sigma).
\end{aligned}
$$

sendo $x_i = (1, x_{1,i}, x_{2,i})$ e β_μ e β_σ vetores de dimensões (1×3).

Com isto, foi-se considerado os seguintes cenários, para verificar se as propriedades frequentistas eram atendidas:

- **Cenário 1:** 1 nível de censura, considerando o quantil de 5% da distribuição Log-Normal;
- **Cenário 2:** 2 níveis de censura, considerando os quantis de 5% e 2.5% da distribuição Log-Normal;
- **Cenário 3:** 3 níveis de censura, considerando os quantis de 5%, 2.5% e 1% da distribuição Log-Normal.

Para todos os cenários descritos anteriormente, foram simulados valores aleatórios para as duas covariáveis, considerando a distribuição normal padrão. A escolha dos parâmetros foi a mesma em todos os casos, sendo definidos pelos seguintes valores.

Tabela 1 – Coeficientes para o estudo de simulação do modelo com múltiplos níveis de censura à esquerda baseado na distribuição log-normal.

i	$\beta_{\mu i}$	$\beta_{\sigma i}$
1	1.00	0.15
2	0.75	0.40
3	0.30	0.80

Para todos os cenários, foram simuladas 1000 amostras, sendo os dados gerados de uma Log-Normal com os parâmetros definidos pela Tabela 1, sendo assim, é possível obter as estimativas dos parâmetros, e por fim os critérios. Para esta seção apenas será apresentado os valores em forma de gráficos, para acessar as tabelas com os valores exatos, acessar Apêndice A.

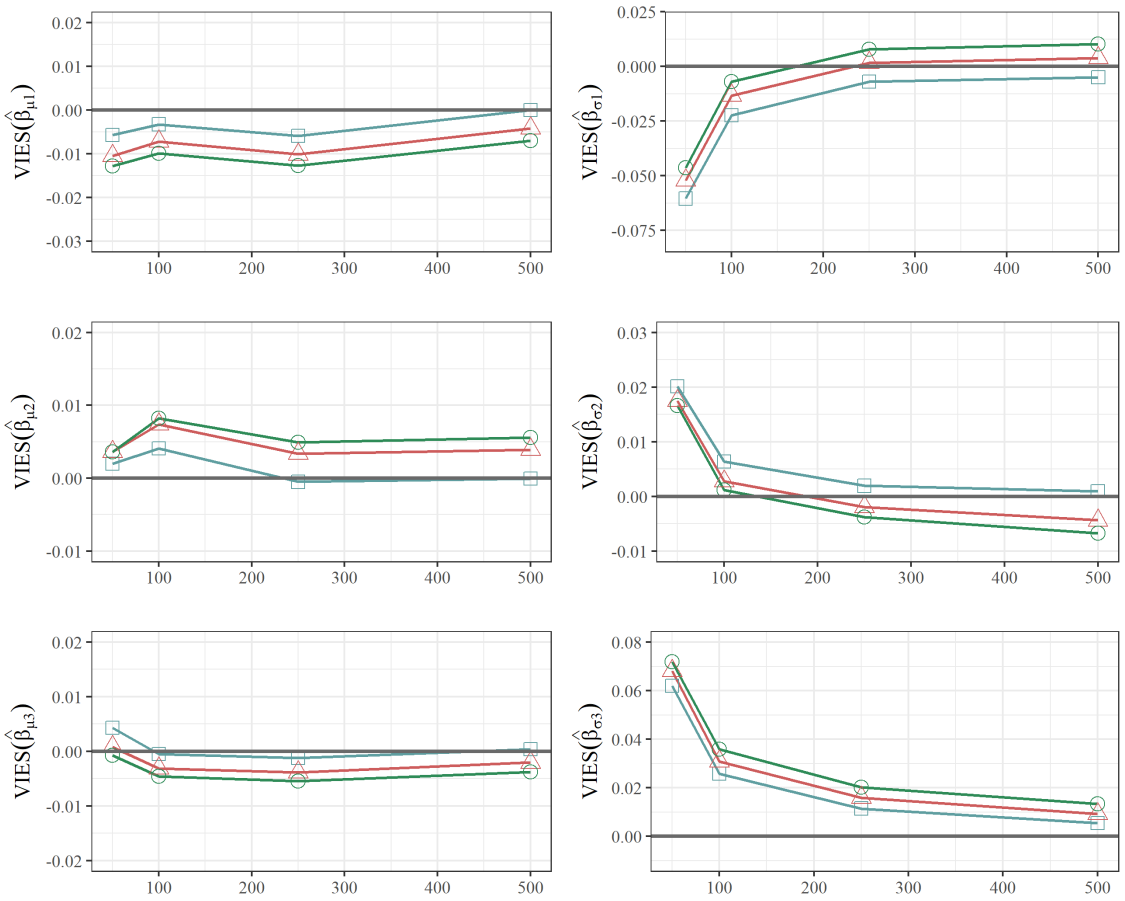

Figura 7 – Variação do viés (VIES($\hat{\beta}_\mu$) e VIES($\hat{\beta}_\sigma$)) em um estudo de simulação com intercepto e duas covariáveis, para diferentes tamanhos de amostra ($n = 50, 100, 250, 500$) para o modelo de regressão Log-Normal com m níveis de censura.

A Figura 7 mostra o viés (VIES) dos estimadores $\hat{\beta}_\mu$ e $\hat{\beta}_\sigma$ em um estudo de simulação para diferentes tamanhos de amostra. Na inferência frequentista, espera-se que os estimadores de máxima verossimilhança (EMV) sejam assintoticamente imparciais e consistentes. Assim, à medida que o tamanho da amostra aumenta, o viés dos estimadores deve diminuir, aproximando-se de zero.

Os gráficos indicam que, para amostras menores ($n = 50$), o viés é mais significativo, mas tende a diminuir com o aumento do tamanho da amostra, conforme esperado. Esse comportamento reflete a maior precisão dos estimadores com amostras maiores, evidenciando a importância de utilizar grandes amostras para estimativas mais confiáveis dos parâmetros β_μ e β_σ.

A Figura 8 mostra o Erro Quadrático Médio (EQM) dos estimadores $\hat{\beta}_\mu$ e $\hat{\beta}_\sigma$ em um estudo de simulação para diferentes tamanhos de amostra. Na inferência frequentista, espera-se que o EQM dos estimadores diminua à medida que o tamanho da amostra aumenta, refletindo maior precisão. Os gráficos indicam que, para amostras menores ($n = 50$), o EQM é mais elevado, indicando maior erro. Conforme o tamanho da amostra cresce, o EQM diminui significativamente,

estabilizando-se em $n = 500$. Isso confirma que amostras maiores proporcionam estimativas mais precisas e confiáveis dos parâmetros β_μ e β_σ.

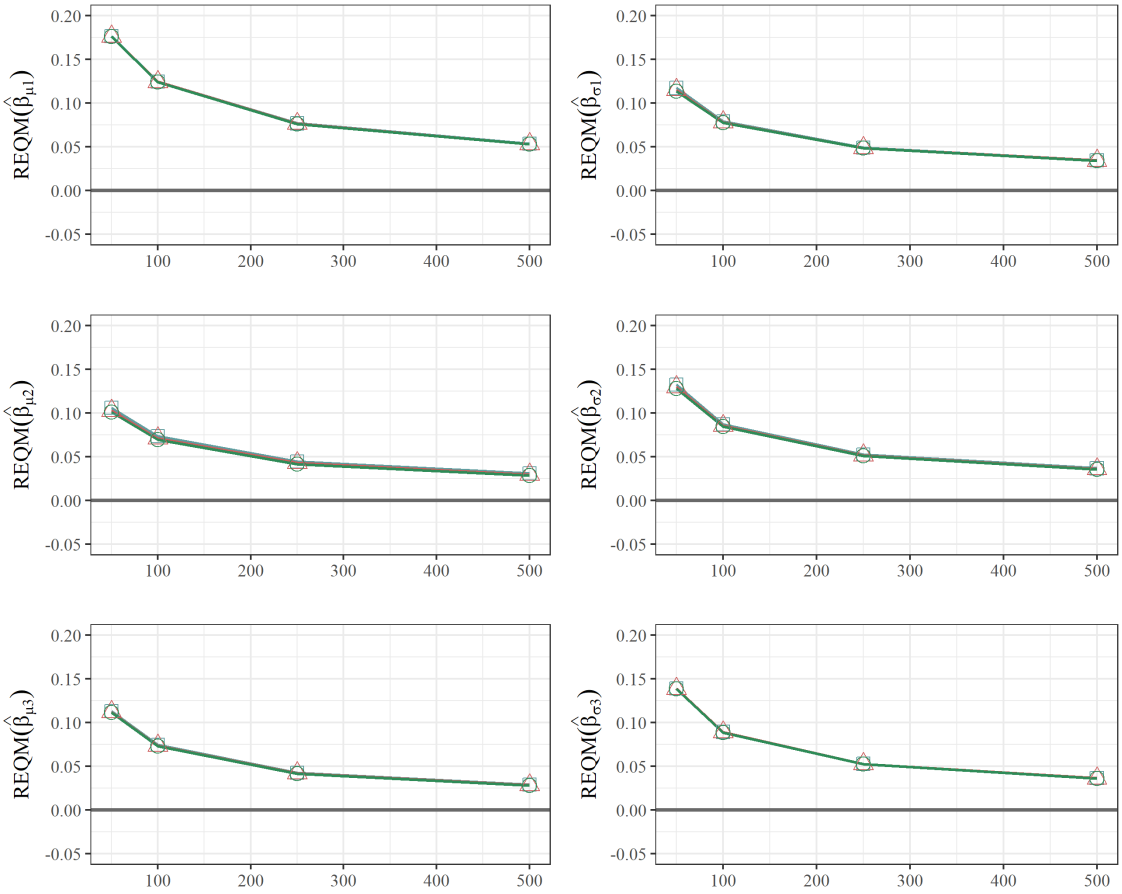

Figura 8 – Variação da raiz do erro quadrático médio (REQM($\hat{\beta}_\mu$) e REQM($\hat{\beta}_\sigma$)) em um estudo de simulação com intercepto e duas covariáveis, para diferentes tamanhos de amostra ($n = 50, 100, 250, 500$) para o modelo de regressão Log-Normal com m níveis de censura.

A Figura 9 mostra a proporção de cobertura (PC) dos intervalos de confiança dos estimadores $\hat{\beta}_\mu$ e $\hat{\beta}_\sigma$ em um estudo de simulação para diferentes tamanhos de amostra, utilizando um nível de confiança de 95%.

Na inferência frequentista, espera-se que a proporção de cobertura se aproxime do nível de confiança especificado, que neste caso é 95% (indicado pela linha pontilhada no gráfico). Isso significa que, idealmente, 95% dos intervalos de confiança devem conter o verdadeiro valor do parâmetro.

Os gráficos mostram que, para tamanhos de amostra menores ($n = 50$), a PC pode ser ligeiramente inferior ou superior ao esperado, mas tende a se estabilizar em torno de 95% conforme o tamanho da amostra aumenta. Isso indica que os intervalos de confiança se tornam mais confiáveis com amostras maiores, alinhando-se com o nível de confiança desejado e garantindo uma inferência estatística robusta.

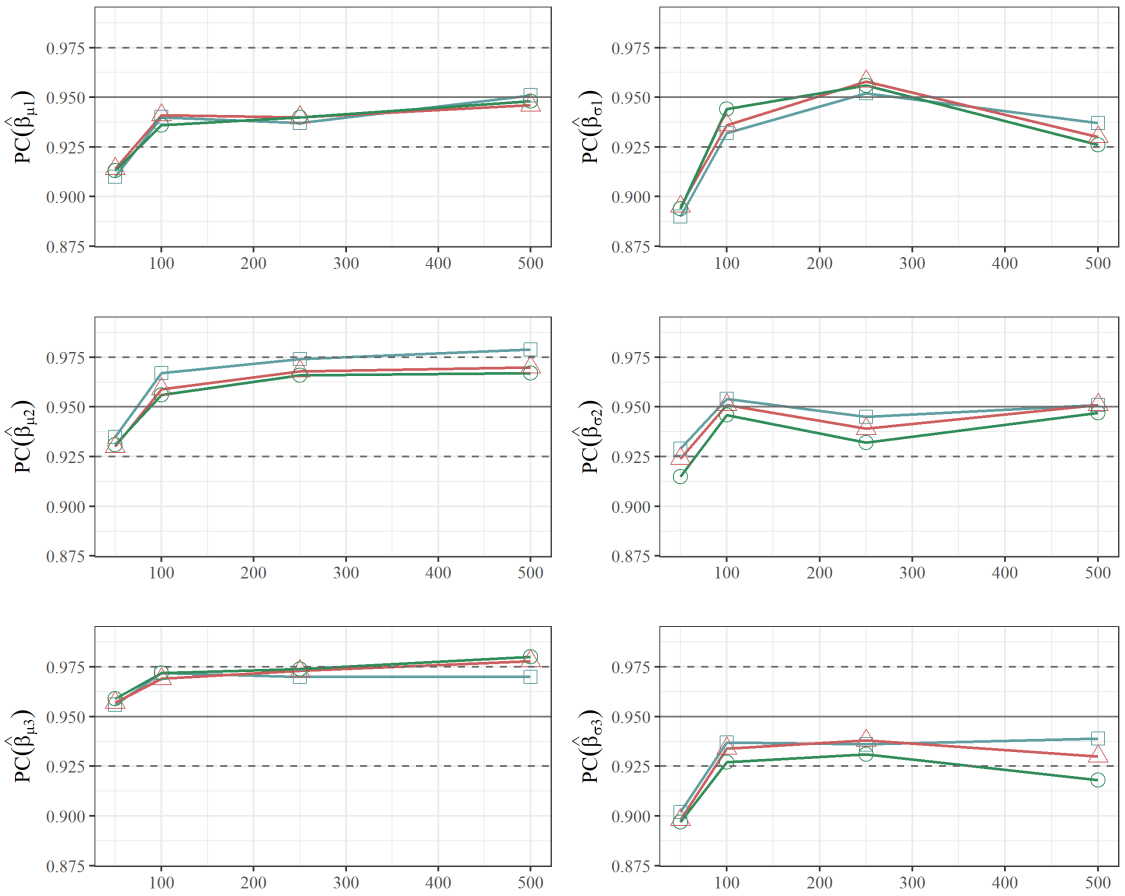

Figura 9 – Variação da probabilidade de cobertura (PC($\hat{\beta}_\mu$) e PC($\hat{\beta}_\sigma$)) em um estudo de simulação com intercepto e duas covariáveis, para diferentes tamanhos de amostra ($n = 50, 100, 250, 500$) para o modelo de regressão Log-Normal com m níveis de censura.

3.7 Aplicação

Os dados utilizados neste trabalho foram disponibilizados por um estudo clínico desenvolvido no Laboratório de Investigação em Dermatologia e Imunodeficiências da Faculdade de Medicina da Universidade de São Paulo. O conjunto de dados refere-se a pacientes diagnosticados com o vírus HIV, divididos em dois grupos principais. O primeiro grupo, denominado CTRL (grupo controle), é composto por indivíduos soropositivos que, caso não recebam tratamento com antirretrovirais, podem apresentar cargas virais elevadas e progressão da doença. O segundo grupo, denominado LNTP (*Long Term Non-Progressors*), inclui pacientes soropositivos que, mesmo sem tratamento antirretroviral, não desenvolvem sintomas e conseguem manter a carga viral sob controle por longos períodos. Segundo Shah e Nadiger (2013), essa característica está presente em cerca de 5 a 15% da população pediátrica infectada pelo HIV.

Além desses dois grupos, existe um terceiro, conhecido como controladores de elite. Esses indivíduos, embora apresentem resultado reagente nos exames sorológicos, não necessitam de terapia antirretroviral para manter a infecção controlada. Eles são capazes de manter, de forma

constante, a carga viral abaixo do limite mínimo de detecção dos exames laboratoriais. Estima-se que menos de 1% da população vivendo com HIV pertença a esse grupo, conforme apontado por BRASIL (2013). No entanto, é importante destacar que esse grupo de pacientes não está presente nos dados analisados neste estudo.

Para melhor compreender a estrutura dos dados que motivaram o desenvolvimento desta abordagem metodológica, é importante observar as principais variáveis envolvidas. O conjunto de dados contempla informações sobre o número de células CD4, carga viral, indicador de censura e o grupo ao qual o paciente pertence. Embora os dados tenham uma natureza longitudinal com múltiplas observações por paciente ao longo do tempo este aspecto não será tratado neste trabalho. Vale ressaltar que os níveis de censura estão associados aos limites mínimos de detecção dos equipamentos utilizados, conforme ilustrado na Tabela 2.

Tabela 2 – Amostra dos dados disponibilizados por um estudo clínico desenvolvido no Laboratório de Investigação em Dermatologia e Imunodeficiências, da Faculdade de Medicina, da Universidade de São Paulo.

Paciente	Data	CD4	Carga Viral	Cens	Tratamento
1	01/01/2000	800	400	1	LNTP
1	02/01/2001	849	400	1	LNTP
1	05/01/2001	768	400	1	LNTP
4	12/02/2008	788	400	1	CTRL
4	02/06/2009	898	400	1	CTRL
4	29/05/2009	662	59	0	CTRL

Como pode ser observado na Tabela 2, os dados apresentam uma estrutura longitudinal, com registros distribuídos entre os anos de 1992 e 2016. No entanto, para a aplicação prática proposta neste trabalho, adotou-se um ponto de corte baseado na última observação disponível para cada paciente, considerando apenas os dados registrados entre 2012 e 2016. Essa seleção resultou na composição final da amostra, que incorpora três níveis distintos de censura, distribuídos entre os dois grupos analisados.

Para o grupo CTRL, foram identificadas 21 observações, das quais 14 são censuradas. Os níveis de censura correspondentes a esse grupo são representados pelo vetor $\mathbf{k}_{\text{CTRL}} = [400, 40]$. Por outro lado, o grupo LNTP possui 19 observações, com 4 censuras associadas aos valores $\mathbf{k}_{\text{LNTP}} = [400, 50]$. Com isso, o conjunto total de dados abrange três níveis distintos de censura, dados por $\mathbf{k} = [400, 50, 40]$. A visualização gráfica desses níveis pode ser consultada na Figura 10.

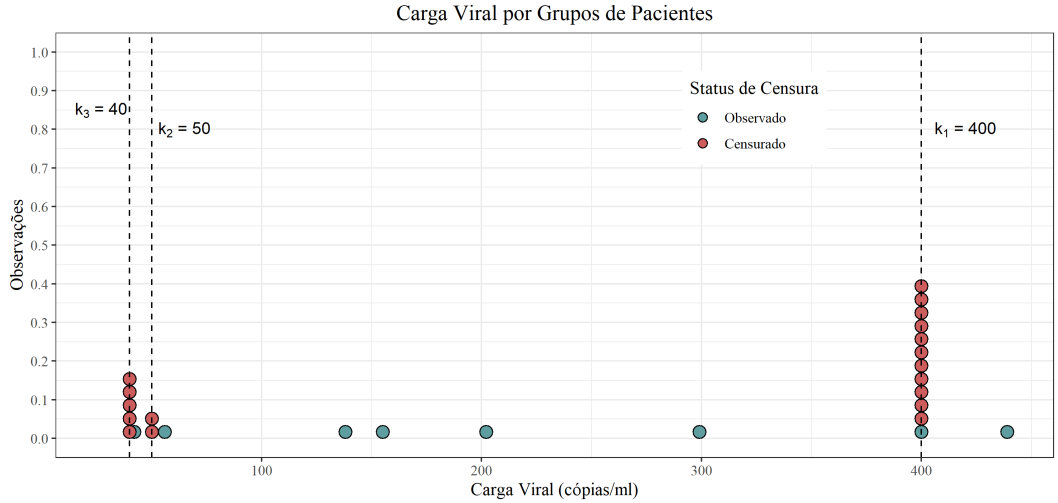

Figura 10 – Níveis de censura - Dados reais.

A Figura 10 apresenta a distribuição das observações de carga viral em relação aos três níveis de censura considerados: $k_3 = 40$, $k_2 = 50$ e $k_1 = 400$, representados pelas linhas verticais tracejadas. Os pontos azuis indicam observações não censuradas, enquanto os vermelhos correspondem a observações censuradas à esquerda. Nota-se a presença de censura concentrada principalmente nos extremos inferior e superior do eixo, refletindo a heterogeneidade entre os grupos analisados. Essa variação nos limites de detecção evidencia a necessidade de se considerar múltiplos níveis de censura na modelagem estatística dos dados.

Inicialmente é possível verificar outras medidas descritivas pela Tabela 3.

Tabela 3 – Medidas sumárias para a carga viral para diferentes grupos.

	Média da Carga Viral	**Variância da Carga Viral**	**Desvio Padrão da Carga Viral**
CTRL	397,00	287.888,00	537,00
LNTP	4.535,00	49.299.304.00	7021,00

Analisando as estatísticas descritivas apresentadas na Tabela 3, observa-se que a média e a variância das cargas virais dos pacientes em tratamento (grupo LNTP) são significativamente maiores do que as observadas no grupo controle (CTRL). Ou seja, mesmo não necessitando do uso de terapia antirretroviral, os indivíduos do grupo CTRL apresentam cargas virais mais baixas e menos dispersas, como evidenciado pelas medidas de dispersão e pela menor amplitude da distribuição.

Embora esse resultado possa parecer contraditório à primeira vista, ele é consistente com o contexto clínico. Em um cenário hipotético sem terapia antirretroviral, esperar-se-ia que os indivíduos do grupo LNTP apresentassem naturalmente carga viral mais baixa, pois sua inclusão no estudo decorre justamente do fato de estarem em tratamento. Contudo, como os pacientes do grupo LNTP estão sob medicação enquanto os do grupo CTRL não, as cargas virais observadas

refletem não apenas o estado imunológico basal dos indivíduos, mas também o efeito da terapia medicamentosa.

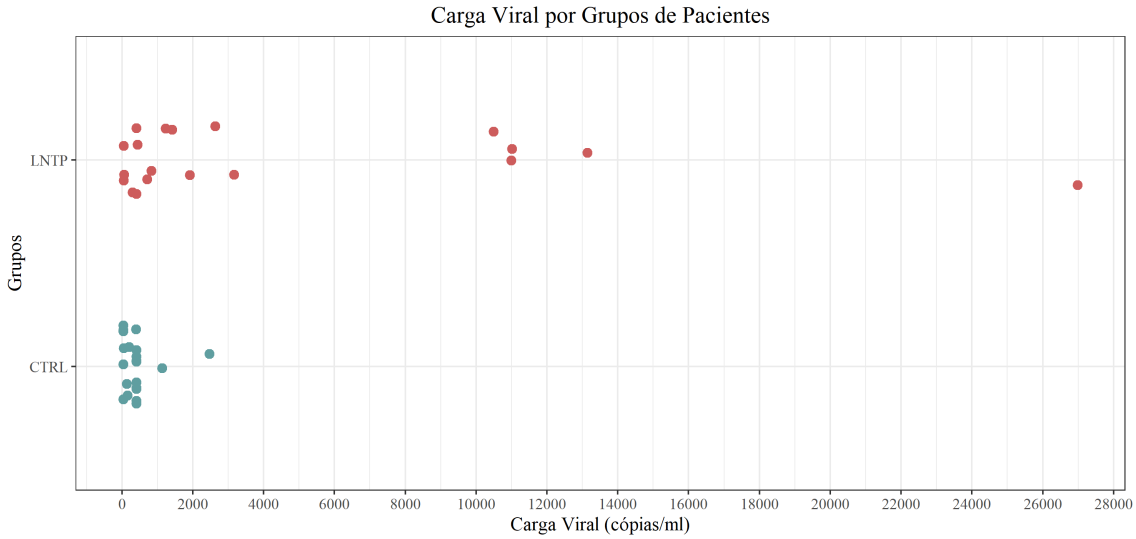

Figura 11 – Stripchart da carga viral em cópias/ml por grupo.

Analisando a Figura 11, percebe-se que os dados do grupo CTRL estão mais concentrados, bem próximos de zero, tendo assim uma amplitude relativamente pequena, sendo que seu maior valor se encontra à esquerda de 5000 e seu menor valor se encontra próximo de 0, com a presença de dois pontos discrepantes. Já no grupo LNTP, os dados possuem uma concentração menor, com uma maior amplitude, sendo que seu maior valor se encontra à direita de 25000 e seu menor valor se encontra próximo de 0, e contém 4 pontos mais distantes dos demais, ou seja, 4 pacientes que tiveram uma carga viral mais elevada que o normal.

Assim, pode-se dizer que o grupo CTRL apresenta um maior número de pacientes com uma menor carga viral, quando comparado ao grupo LNTP. É notória a forte assimetria à direita da distribuição da carga viral em ambos os grupos, porém mais evidente no grupo LNTP, razão pela qual a distribuição Weibull foi sugerida para a aplicação.

Também pode ser analisado se há diferença no número de células CD4 entre os diferentes grupos, considerando que elas são muito importantes, pois são tipos de leucócitos que têm funções essenciais no sistema imunológico.

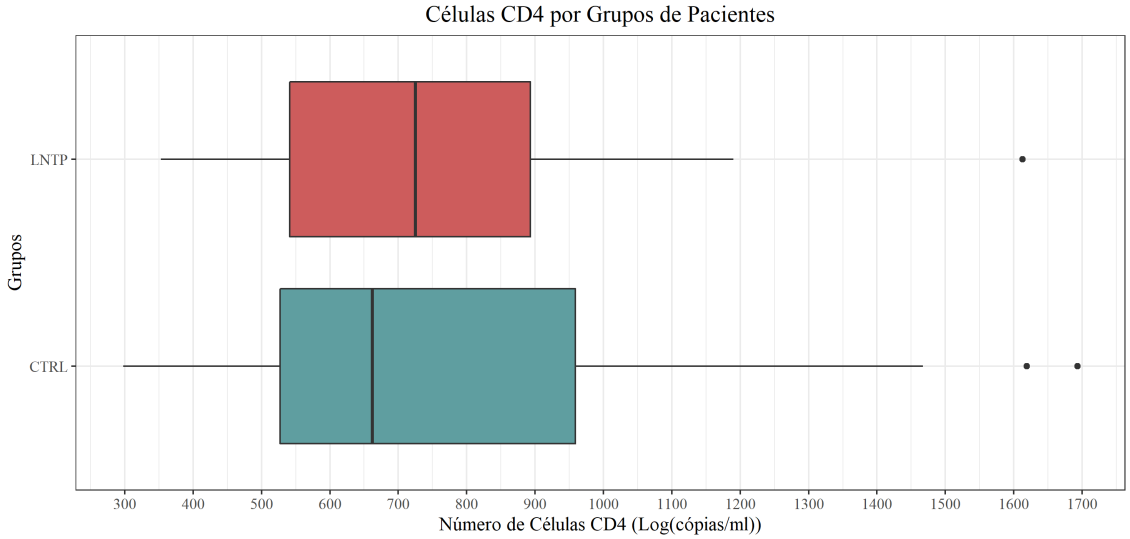

Figura 12 – Box-plot do número de célular cd4 em cópias/ml por grupo.

Nota-se pela Figura 12 que não há muitas diferenças entre o número de células CD4 para os dois grupos, mesmo com o grupo LNTP tendo valores medianos maiores para a carga viral, não se notou um menor número de células CD4 para eles, o que pode ser explicado também pelo motivo dele ser um grupo que consegue controlar bem a sua carga viral, sem a utilização dos coquetéis antirretrovirais, logo o seu sistema imunológico, não sofre tanto com isso.

Para a aplicação serão considerados três casos distintos, inicialmente será aplicado o modelo Weibull, com a acoplação das covariáveis apenas no parâmetro α, o segundo caso, considerando a distribuição Log-Normal com covariáveis acopladas no parâmetro μ, e por fim, novamente a distribuição Log-Normal, mas agora com as variáveis ligadas aos dois parâmetros associados a ela. O uso destas duas distribuições foi feito devido as particularidades delas, dado que como os dados possuem uma grande variabilidade e com valores assumindo apenas no intervalo positivo dos números reais.

3.7.1 Modelo de regressão Weibull

Inicialmente, o modelo Weibull com três níveis de censura foi aplicado considerando a variável resposta em sua escala original, e adotando o grupo e o número de células CD4 como covariáveis. A fim de verificar a significância dessas covariáveis no modelo, foi realizado o Teste da Razão de Verossimilhança (TRV).

- H_0: Nenhuma dos coeficientes possui efeito significativo sobre os parâmetros do modelo, ou seja, todos os coeficientes associados aos parâmetros α e γ são iguais a zero;

- H_1: Pelo menos uma dos coeficientes possui efeito significativo sobre algum dos parâmetros do modelo (α ou γ), ou seja, ao menos um dos coeficientes é diferente de zero.

Tabela 4 – Teste da Razão de Verossimilhança (TRV) para Avaliação de Covariáveis (Modelo Weibull com covariáveis no α)

	Estatística	**p-Valor**
Covariáveis	19.627	<0.0001

A Tabela 4, referente ao teste TRV, exibe uma estatística de 19.627 com um p-valor próximo de zero. Este p-valor extremamente baixo indica que rejeitamos a hipótese nula de que o modelo sem as covariáveis se ajusta tão bem quanto o modelo com as covariáveis. Em outras palavras, espera-se que pelo menos um das covariáveis incluídas no modelo sejam estatisticamente significativas e melhore significativamente o ajuste do modelo. Agora para visualizar qual das covariáveis são significativas ao modelo, foi realizado o teste de wald para cada uma delas.

Tabela 5 – Resultados do Teste de Wald para Parâmetros do Modelo Weibull (com covariáveis no α).

	Estimativa	**Err. Pad.**	**Lim. Inf.**	**Lim. Sup.**	**t Valor**	**p-Valor**
γ	0.4948	0.0844	0.3293	0.6602	5.8610	0.0000
Intercepto	5.6937	0.8676	3.9933	7.3941	6.5628	0.0000
CD4	-0.0016	0.0009	-0.0033	0.0002	-1.7488	0.0803
Tratamento	3.1669	0.6703	1.8532	4.4807	4.7249	0.0000

A Tabela 8 mostra que os p-valores confirmam a significância ao se considerar um nível de 5% para os coeficientes relacionados ao parâmetro γ, além do intercepto e a variável ligada ao tratamento. O número de células CD4 por outro lado ao nível de significância de 5% não se mostrou significativa.

Com isto, para verificar se os resultados são confiáveis vamos calcular os resíduos assim como especificados na Subseção 3.5.1 para verificar se o modelo está bem ajustado. Isso nos permitirá avaliar a adequação dos resíduos à distribuição Normal.

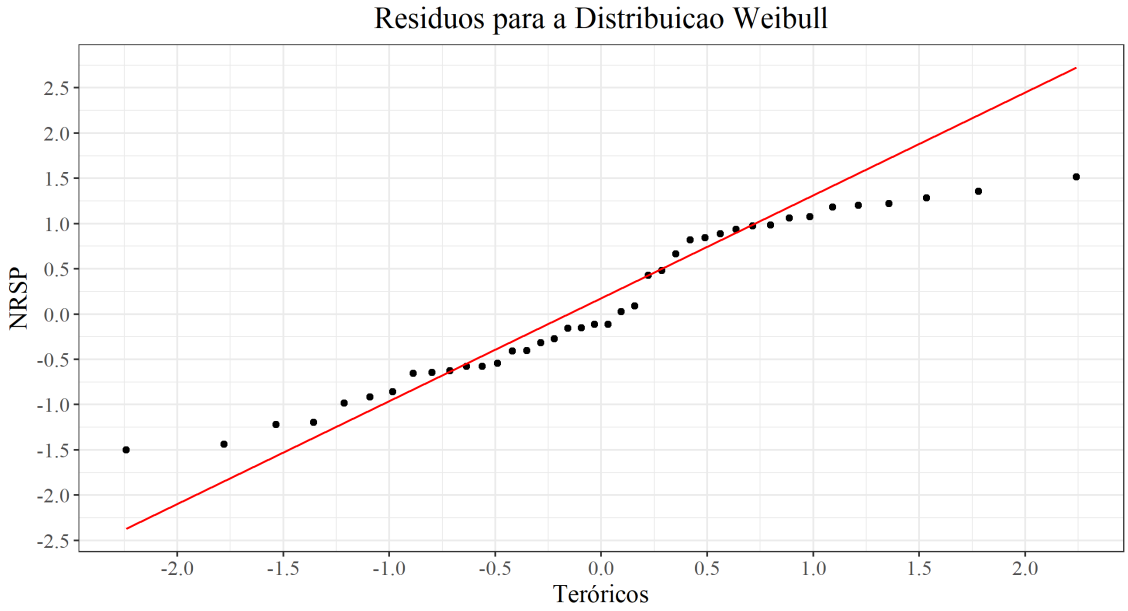

Figura 13 – Gráfico quantil x quantil para o modelo weibull com variáveis acopladas no α.

De acordo com a Figura 14, o gráfico QQ mostra que os resíduos do modelo, quando comparados com uma distribuição Log Normal, não seguem perfeitamente essa distribuição, especialmente nas extremidades onde há desvios significativos em relação à linha de referência. Isso sugere que os resíduos não seguem uma distribuição Normal.

Tabela 6 – Testes de normalidade para os resíduos NRSP para o modelo Weibull com variáveis acopladas no α.

	Estatística	**p-Valor**
Shapiro-Wilk	0.9234	0.0099
Shapiro-Francia	0.9223	0.0112
Kolmogorov-Smirvov	0.2304	0.0286
Lilliefors	0.1838	0.0016
Cramer-von Mises	0.3638	0.0898

Os resultados dos testes de normalidade reforçam essa conclusão. Os p-valores dos testes são todos menores que 0.05, indicando que os resíduos não seguem uma distribuição normal. Apenas o teste Cramer-von Mises apresenta um p-valor maior (0.0898), mas isso não é suficiente para contradizer os resultados dos outros testes. Portanto, concluímos que os resíduos não são normalmente distribuídos.

3.7.2 Modelo Log-Normal com covariáveis no parâmetro μ

Também foi aplicado para o modelo Log-Normal com 3 níveis de censura considerado a variável resposta na sua escala original e considerando o grupo e o número de células CD4 como

uma covariável acopladas no parâmetro de locação (μ), tendo inicialmente testado se algumas das covariáveis são ou não significativas por meio do TRV.

- H_0: Nenhuma dos coeficientes possui efeito significativo sobre os parâmetros do modelo, ou seja, todos os coeficientes associados aos parâmetros μ e são iguais a zero;

- H_1: Pelo menos uma dos coeficientes possui efeito significativo sobre algum dos parâmetros do modelo (μ ou σ), ou seja, ao menos um dos coeficientes é diferente de zero.

Tabela 7 – Teste da Razão de Verossimilhança (TRV) para Avaliação de Covariáveis (Modelo log-normal com covariáveis no μ)

	Estatística	**p-Valor**
Covariáveis	21.0446	<0.0001

A Tabela 7, referente ao teste TRV, exibe uma estatística de 21.0446 com um p-valor próximo de 0. Este p-valor extremamente baixo indica que rejeitamos a hipótese nula de que o modelo sem as covariáveis se ajusta tão bem quanto o modelo com as covariáveis. Em outras palavras, espera-se que pelo menos um das covariáveis incluídas no modelo sejam estatisticamente significativas e melhore significativamente o ajuste do modelo. Agora para visualizar qual das covariáveis são significativas ao modelo, foi realizado o teste de wald para cada uma delas.

Tabela 8 – Resultados do Teste de Wald para Parâmetros do Modelo (Modelo com covariáveis no μ).

	Estimativa	**Err. Pad.**	**Lim. Inf.**	**Lim. Sup.**	**t Valor**	**p-Valor**
σ	0.6872	0.1602	0.3733	1.0011	4.2907	0.0000
Intercepto	5.7123	1.0124	3.7280	7.6966	5.6423	0.0000
CD4	-0.0026	0.0011	-0.0047	-0.0005	-2.4416	0.0146
Tratamento	3.1211	0.7635	1.6246	4.6175	4.0880	0.0000

A Tabela 8 Esses p-valores confirmam a significância dos coeficientes de todos os coeficientes ao se considerar um nível de confiança de 5%. Contudo para ter confiança nos resultados obtidos, é preciso verificar os resíduos.

Com isto, para verificar se os resultados são confiáveis vamos calcular os resíduos quantílicos para verificar se o modelo está bem ajustado. Isso nos permitirá avaliar a adequação dos resíduos à distribuição Normal.

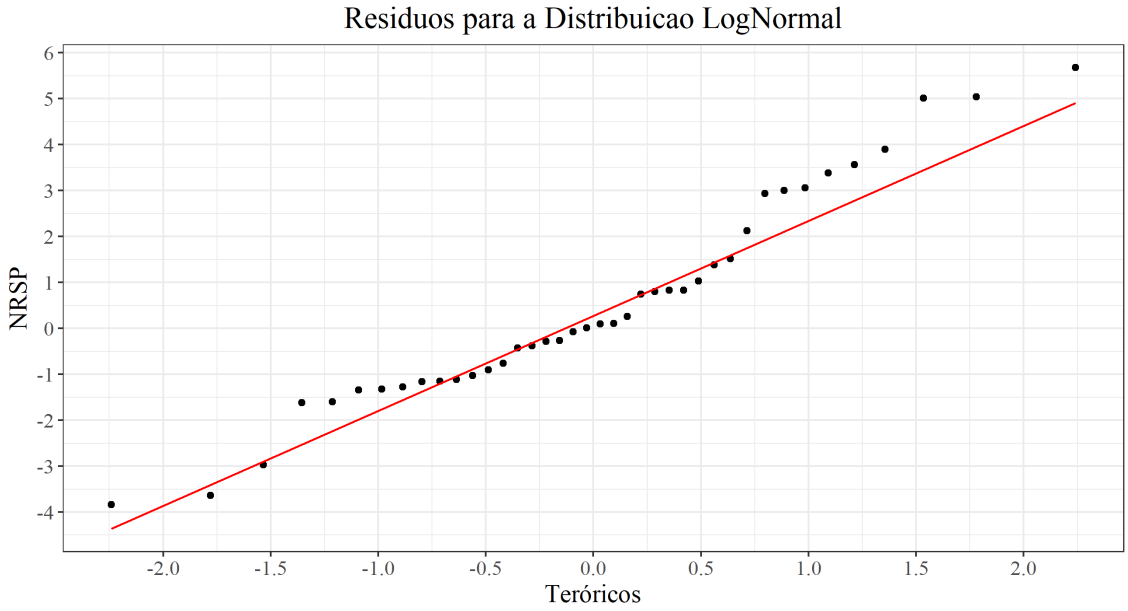

Figura 14 – Gráfico quantil x quantil para o modelo lognormal com variáveis acopladas no μ.

De acordo com a Figura 14, o gráfico QQ mostra que os resíduos do modelo, quando comparados com uma distribuição Log Normal, não seguem perfeitamente essa distribuição, especialmente na cauda superior onde há desvios significativos em relação à linha de referência. Isso sugere que os dados originais não são log-normalmente distribuídos.

Tabela 9 – Testes de normalidade para os resíduos quantílicos para o modelo lognormal com variáveis acopladas no μ.

	Estatística	**p-Valor**
Shapiro-Wilk	0.9412	0.0379
Shapiro-Francia	0.9422	0.0416
Kolmogorov-Smirvov	0.2332	0.0213
Lilliefors	0.1497	0.0242
Cramer-von Mises	0.3982	0.0752

Os resultados dos testes de normalidade reforçam essa conclusão. Os p-valores dos testes são todos menores que 0.05, indicando que os resíduos não seguem uma distribuição normal. Apenas o teste Cramer-von Mises apresenta um p-valor maior (0.0752), mas isso não é suficiente para contradizer os resultados dos outros testes. Portanto, concluímos que os resíduos não são normalmente distribuídos.

3.7.3 Modelo Log-Normal com covariáveis nos parâmetros μ e σ

Também foi aplicado para o modelo Log-Normal com 3 níveis de censura considerado a variável resposta na sua escala original e considerando o grupo e o número de células CD4 como

uma covariável acopladas no parâmetro de locação (μ) e escala (σ), tendo inicialmente testado se algumas das covariáveis são ou não significativas por meio do TRV.

- H_0: Nenhuma dos coeficientes possui efeito significativo sobre os parâmetros do modelo, ou seja, todos os coeficientes associados aos parâmetros μ e σ são iguais a zero;
- H_1: Pelo menos uma dos coeficientes possui efeito significativo sobre algum dos parâmetros do modelo (μ ou σ), ou seja, ao menos um dos coeficientes é diferente de zero.

Tabela 10 – Teste da Razão de Verossimilhança (TRV) para Avaliação de Covariáveis (Modelo log-normal com covariáveis no μ e σ.)

	Estatística	**p-Valor**
Covariáveis	20,3666	0.0004

A Tabela 10, referente ao teste TRV, exibe uma estatística de 20.3666 com um p-valor de 0.0004. Este p-valor extremamente baixo indica que rejeitamos a hipótese nula de que o modelo sem as covariáveis se ajusta tão bem quanto o modelo com as covariáveis. Em outras palavras, espera-se que pelo menos um das covariáveis incluídas no modelo sejam estatisticamente significativas e melhore significativamente o ajuste do modelo. Agora para visualizar qual das covariáveis são significativas ao modelo, foi realizado o teste de wald para cada uma delas.

Tabela 11 – Resultados do Teste de Wald para Parâmetros do Modelo (Modelo com covariáveis no μ e σ).

	Estimativa	**Err. Pad.**	**Lim. Inf.**	**Lim. Sup.**	**t Valor**	**p-Valor**
Intercepto (μ)	6.8909	1.0520	4.8290	8.9527	6.5505	0.0000
CD4 (μ)	-0.0040	0.0019	-0.0076	-0.0003	-2.1389	0.0324
Trat (μ)	2.5287	0.6871	1.1820	3.8753	3.6804	0.0002
Intercepto (σ)	-0.4800	0.2677	-1.0047	0.0446	-1.7933	0.0729
CD4 (σ)	0.0011	0.0001	0.0008	0.0014	7.6274	0.0000
Trat (σ)	0.5086	0.3072	-0.0935	1.1106	1.6557	0.0978

A Tabela 11 mostra que que ao se considerar um nível de significância de 5% todas as variáveis ligadas ao parâmetro de locação são dadas como significativas. Já para as variáveis ligadas ao parâmetro de escala, σ, apenas a variável ligada ao número de célular CD4 é significativa a um nível de significância de 5%.

Com isto, para verificar se os resultados são confiáveis vamos calcular os resíduos quantílicos para verificar se o modelo está bem ajustado. Isso nos permitirá avaliar a adequação dos resíduos à distribuição Normal.

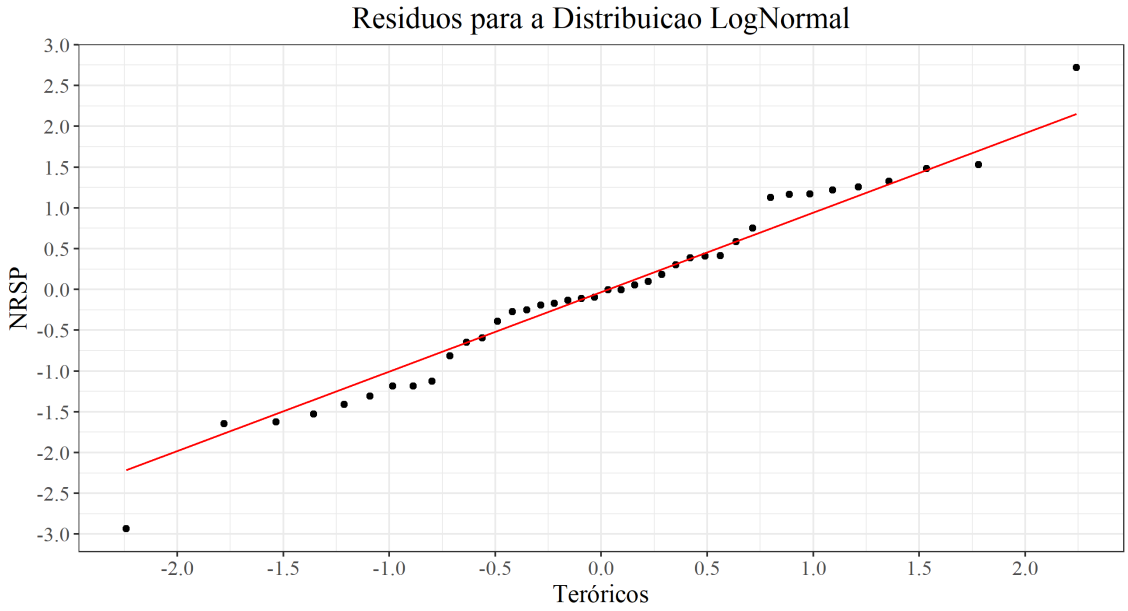

Figura 15 – Gráfico quantil x quantil para o modelo lognormal com variáveis acopladas no μ e σ.

De acordo com a Figura 15, o gráfico QQ apresentado compara os resíduos do modelo com uma distribuição Log Normal, considerando variáveis acopladas no μ e σ. Neste gráfico, os pontos seguem mais de perto a linha de referência vermelha, sugerindo que os resíduos estão mais alinhados com a distribuição Log Normal, tendo principalmente 2 pontos mais distantes nas extremidades, porém não indicando que a suposição não seja atendida

Tabela 12 – Testes de normalidade para os resíduos quantílicos para o modelo lognormal com variáveis acopladas no μ e σ.

	Estatística	**p-Valor**
Shapiro-Wilk	0.9814	0.7434
Shapiro-Francia	0.9757	0.4551
Kolmogorov-Smirvov	0.0949	0.8306
Lilliefors	0.0911	0.5510
Cramer-von Mises	0.0552	0.8469

Os resultados dos testes de normalidade mostram que os p-valores para os testes são maiores que 0.05, indicando que não há evidências suficientes para rejeitar a hipótese de normalidade dos resíduos. Por isso, para este caso pode-se dizer que a suposição de que os dados seguem uma distribuição Log-Normal é adequada.

3.8 Escolha do Modelo

Nesta seção, realizamos uma análise comparativa de diferentes modelos para ajustar os dados, utilizando critérios de seleção de modelos amplamente reconhecidos: *AIC* (*Akaike*

Information Criterion), *BIC* (*Bayesian Information Criterion*) e *AICc* (*Akaike Information Criterion* Corrigido). Esses critérios são usados para avaliar o desempenho dos modelos levando em consideração tanto o ajuste aos dados quanto a complexidade do modelo, penalizando um número maior de parâmetros. A comparação foi feita entre três modelos: o modelo *Weibull*, o *Log-Normal* considerando apenas a acoplação no parâmetro μ, e o *Log-Normal* considerando acoplação em ambos os parâmetros μ e σ. O objetivo é selecionar o modelo que melhor explica os dados com base nesses critérios, buscando o menor valor de *AIC*, *BIC* e *AICc*.

Tabela 13 – AIC, BIC e AICC para as estimativas frequentistas realizadas.

	Weibull	**Log Normal** (μ)	**Log Normal** (μ, σ)
AIC	-396.09	-396.01	-396.69
BIC	-389.33	-389.26	-389.93
AICC	-394.95	-394.87	-395.55

A Tabela 13 apresenta os valores de *AIC*, *BIC* e *AICc* para três modelos: Weibull, Log-Normal (μ) e Log-Normal (μ, σ). O modelo *Log-Normal* (μ, σ) apresentou o menor valor de *AIC* (-396.69), indicando que este modelo oferece o melhor ajuste aos dados, levando em consideração a penalização pela complexidade. O *BIC* segue a mesma tendência, mostrando que o modelo *Log-Normal* (μ, σ) também é preferível (-389.93), pois apresentou o menor valor dentre os três modelos. O *AICc*, que corrige o *AIC* para amostras pequenas, reforça a preferência pelo modelo *Log-Normal* (μ, σ), com um valor de -395.55.

Portanto, com base nos três critérios (*AIC*, *BIC* e *AICc*), o modelo *Log-Normal* (μ, σ) se destaca como o modelo mais adequado para explicar os dados, pois ele apresenta os menores valores em todos os critérios, sugerindo o melhor equilíbrio entre ajuste aos dados e complexidade do modelo.

3.9 Considerações finais

Este capítulo apresentou a metodologia principal deste trabalho, abordando a inserção de múltiplos níveis de censura e mostrando como cada observação contribui para o modelo, integrando informações de indivíduos com cargas virais mensuráveis e daqueles com valores censurados pelos limites inferiores. Foram estimados os parâmetros associados, seus intervalos de credibilidade e realizados testes de significância, além da análise de resíduos. Nesta etapa, também foi conduzido um estudo de simulação para investigar a consistência e eficiência dos estimadores de máxima verossimilhança (EMVs), considerando diferentes tamanhos amostrais. Para isso, foram analisados três cenários e avaliados três critérios: viés, raiz quadrada do erro quadrático médio (EQM) e a probabilidade de cobertura (PC). Os resultados das simulações frequentistas corroboraram a eficiência da metodologia, destacando sua robustez. Além disso, três modelos foram comparados, e os resultados indicaram que o modelo *Log-Normal* (μ, σ)

é o mais adequado, apresentando os menores valores de *AIC*, *BIC* e *AICc*, bem como melhor adequação dos resíduos. Isso demonstra o equilíbrio ideal entre ajuste e complexidade, reforçando a aplicabilidade do modelo proposto.

CAPÍTULO

4

MÉTODOS BAYESIANOS

De acordo com Kinas e Andrade (2017), a estatística bayesiana difere das metodologias convencionais por não estar apenas atrelada a probabilidades em um contexto restrito, já que, no cenário bayesiano, a probabilidade é encarada como uma "medida racional" e "condicional de incerteza". Um ponto interessante é que o avanço tecnológico fez com que o enfoque bayesiano rompesse suas barreiras de utilização, pois era considerado, em décadas passadas, computacionalmente pesado.

Em um modelo bayesiano, considera-se a informação a respeito dos parâmetros do modelo (θ), além daquela referente às observações de uma variável aleatória Y (dados coletados).

A informação sobre o parâmetro usada para modelar os parâmetros é denominada função ou distribuição a priori, denotada por $\pi(\theta)$. Já a informação sobre os dados é representada pela verossimilhança, denotada por $\pi(y|\theta)$. A informação especificada pela priori é incorporada ao modelo, acrescentando uma camada extra de conhecimento por meio do teorema de Bayes.

Na perspectiva do teorema de Bayes, após observar $Y = y$, a informação disponível sobre θ é incorporada utilizando-se:

$$\pi(\theta|y) = \frac{\pi(\theta,y)}{\pi(y)} = \frac{\pi(\theta)\pi(y|\theta)}{\int \pi(\theta,y)\,d\theta}, \tag{4.1}$$

sendo essa denominada distribuição a posteriori.

Como o denominador da Equação (4.1) é uma constante (isto é, não depende de θ), a equação pode ser expressa como:

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta). \tag{4.2}$$

Em alguns casos, pode haver pouco conhecimento prévio sobre o parâmetro θ, levando à necessidade de utilizar uma distribuição a priori que não introduza informação adicional ao

modelo. Nesses casos, adota-se a chamada *priori não informativa*, cuja escolha deve ser feita com cautela. Uma consideração importante nessa escolha é a possibilidade de a priori ser *imprópria*, o que pode impactar a inferência estatística e exigir uma análise criteriosa.

4.1 Informação a priori

De acordo com DeGroot (2002) escolha da distribuição a priori é um dos passos fundamentais para a construção de um modelo baseado em inferência bayesiana. Caso a determinação da escolha da priori não seja realizada, não será possível o calculo da posteriori, logo, comprometendo a análise bayesiana. Em suma, diferentes prioris levam a diferentes resultados, especialmente quando temos prioris muito diferentes sendo comparadas. Diferenciando-se de quando se utiliza a inferência frequentista, pois se uma variável aleatória X possui distribuição binomial, dado determinado critério, X/n será o seu melhor estimador O'Hagan e Forster (2004). Sendo válido para todo caso em que o interesse é modelar uma variável aleatória assumindo distribuição binomial. Diferenciando da estatística bayesiana, em que a informação do especialista é levada em consideração através da construção da priori, podendo assim diferenciar-se para cada problema. Pois, mesmo para casos em que a verossimilhança seja a mesma, diferentes usos de prioris, resultaram em diferentes posterioris para a condução das análises. Para se conduzir a escolha das prioris, deve-se levar em consideração as seguintes características, sendo elas:

- Pertencer ao espaço paramétrico;
- Resultar em uma posteriori integrável;
- Reproduzir adequadamente o conhecimento do especialista sobre o parâmetro.

4.1.1 Prioris subjetivas

Em determinadas ocasiões é possível obter informações sobre o parâmetro, podendo assim explicitar alguma forma para a priori. Contudo, alguns atores como Paulino, Turkman e Murteira (2018) abordam um problema comum no ambiente científico, que é a dificuldade em quantificar estas informações para que possa ser utilizada na inferência bayesiana. Medidas sumárias é um dos principais objetos que auxiliam a determinação da quantificação, a escolha da média ou desvio padrão que o pesquisador acredita sobre a distribuição a priori é fundamental para a escolha de uma distribuição explicitamente. Para casos em que há de antemão a informação sobre a média e a variância, é habitual a utilização da distribuição normal, ou para casos em que é condicionado no espaço paramétrico do parâmetro assumir apenas valores positivos, a utilização da distribuição Gama pode ser interessante O'Hagan e Forster (2004). Além disso existem alguns métodos utilizados para a escolha de prioris subjetivas, como o método do histograma, método preditivo de eliciação e um dos mais conhecidos a utilização de famílias conjugadas.

4.1.2 Prioris objetivas (não informativas)

Um cenário diferente do caso anterior, é quando a informação sobre o parâmetro é pouca ou inexistente, para isso vários autores apresentam teorias para suprir esta carência, sendo distribuições compostas pela mínima informação acima do parâmetro, o que é conhecido na literatura como priori não informativa. A priori não informativa pode ser definida como a distribuição que não privilegia nenhum valor específico de θ em relação aos outros Berger (2013). Esta classe de priori pode ser útil para servir como referência, sendo úteis para casos em que o conhecimento disponível não é suficiente para a utilização de uma distribuição subjetiva. Elas também são interessantes quando há o interesse de comparação com a teoria clássica ou para fim de análise de impacto de prioris subjetivas Paulino, Turkman e Murteira (2018). Algumas das prioris objetivas conhecidas na literatura, são a priori de *Jeffreys*, priori de máxima entropia e priori de Berger e Bernado.

4.1.3 Misturas de Prioris

Em algumas situações os dados se comportam de formas específicas para diferentes grupos. Não sendo adequado o uso dos grupos de prioris especificadas anteriormente, dado que não é levado em consideração a informação de disparidade. Um método interessante para a aplicação, é a mistura de distribuições para a composição da priori. Deste modo, dado que os dados são definidos por k grupos, pode-se definir a priori da seguinte maneira:

$$\pi(\theta) = \sum_{i=1}^{k} w_i \pi_i(\theta),$$

com $k \geq 1$, $w_i > 0$ com $w_1 + w_2 + \cdots + w_k = 1$ para $j = 1, 2, \ldots, k$ e cada $\pi_j(\theta) > 0$ e $\int \pi_j(\theta) d\theta = 1$ para $j = 1, 2, \ldots, k$. Além disso, como os pesos w_i são desconhecidos, há a necessidade da atribuição de uma distribuição a priori para o vetor de pesos $w = (w_1, w_2, \ldots, w_k)$, sendo intuitivo o uso $w \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

4.2 Monte Carlo via Cadeia de Markov (MCMC)

O método de Monte Carlo via Cadeias de Markov, conhecido como MCMC, é realizado de maneira iterativa, baseando-se em cadeias de Markov que, para o propósito deste trabalho, é utilizado para a obtenção da distribuição a posteriori. Este tipo de método é diverso: amostrador de gibbs, algoritmos de metropolis-hastings, hamiltoniano monte carlo, entre outros, em que, os dois últimos serão abordados ainda nesta seção.

4.2.1 Metropolis-Hastings

Inicialmente introduzido por Metropolis *et al.* (1953), o algoritmo de Metropolis-Hastings é o método mais básico e simples dentro da classe de métodos baseados em MCMC. Em Hanson

e Cunningham (1998) é visto que sua eficiência, em problemas de baixa dimensionalidade ou complexidade, é realmente notada, porém, à medida que sua distribuição começa a ganhar um grau maior de complexidade e dimensão, o algoritmo acaba por ser menos eficiente. Outro ponto, é que pode ocorrer presença de altas correlações para estados sucessivos, como consequência da natureza aleatória dentro do espaço de parâmetros, resultando assim em um processo mais lento para a obtenção de sua distribuição (TORRES; STUTZ; NETO, 2019).

Hastings (1970) generalizou o algoritmo para que pudesse também ser utilizado para distribuições assimétricas. O conceito deste algoritmo é simular valores de uma distribuição auxiliar e, posteriormente, aceitá-lo ou não com certa probabilidade. Considerando θ, e $\theta^{'}$ como sendo o estado atual da cadeira e o valor obtido pela distribuição proposta $q(\cdot|\theta)$ respectivamente, aceita-se $\theta^{'}$ com probabilidade igual a

$$\alpha(\theta,\theta^{'}) = min\left\{1, \quad \frac{\pi(\theta^{'})q(\theta|\theta^{'})}{\pi(\theta)q(\theta^{'}|\theta)}\right\},$$

em que, $\pi(\cdot)$ é a distribuição a qual se está interessado.

Algo interessante associado a este método é que não é necessário se ter o conhecimento total da distribuição a qual está interessado, isto é, também contempla distribuição as quais são parcialmente conhecidas a menos de uma constante, o que é interessante para o ponto de vista bayesiano, em que muitas vezes não se conhece a distribuição exata de suas posterioris. Deste modo, o algoritmo de Metropolis-Hasting é visto como a sequência das seguintes etapas.

1. Ativar o contador t = 0 e atribuir um valor inicial para o parâmetro $\theta^{(0)}$;
2. Gerar um novo valor $\theta^{'}$ da distribuição proposta $q(\cdot|\cdot)$;
3. Calcular a probabilidade de aceitação $\alpha(\theta,\theta^{'})$;
4. Gerar um valor $u \sim U[0,1]$;
5. Se $u \leq \alpha(\theta,\theta^{'})$, então aceita-se o novo valor, isto é, $\theta^{(t+1)} = \theta^{(')}$. Caso contrário, rejeita-se o novo valor e $\theta^{(t+1)} = \theta^{(t)}$;
6. Repetir até $t = t+1$.

Um caso particular, em que apenas é considerado distribuições simétricas, denominado como algoritmo de metropolis, foi apresentando inicialmente por Metropolis e Ulam (1949), isto é isto é, $q(\theta|\theta^{'}) = q(\theta^{'}|\theta)$ para todos os valores de θ e θ^{i}. Especificamente, para estes casos, é possível simplificar a probabilidade de aceitação, a qual é dada por,

$$\alpha(\theta,\theta^{'}) = min\left\{1, \quad \frac{\pi(\theta^{'})}{\pi(\theta)}\right\}.$$

Outro ponto, é que pelo algoritmo sofre muita influência da escolha de uma boa distribuição auxiliar, o que por muitas vezes é algo fácil de se realizar. Tendo esse tipo de dificuldade, surgem outros algoritmos na literatura, sendo um destes o Monte Carlo Hamiltoniano.

4.2.2 Monte Carlo Hamiltoniano (HMC)

Visto a dificuldade de algoritmos como o de Metropolis-Hasting em situações que a complexidade da distribuição aumenta, Duane *et al.* (1987) propuseram o método Monte Carlo Hamiltoniano (HMC). No artigo inicial, em que o método HMC foi apresentado, tal método foi chamado de método de Monte Carlo Híbrido. O interessante do HMC é que ele se baseia na dinâmica Hamiltoniana para gerar amostras para a sua distribuição de interesse (NEAL *et al.*, 2011).

O método do HMC pode ser entendido por duas vertentes, uma do ponto de vista de dinâmica moleculares, para suprir a necessidade de correção dos erros ao discretizar equações de Hamilton. Em contrapartida, pelo panorama estatístico, a utilização dele é visto como um aparato eficiente para a utilização em métodos baseados em MCMC. Nesta metodologia, o interessante é que os estados são aceitos com altas probabilidades, visto que os estados atuais estarão longes dos novos estados (CALDERHEAD, 2011).

Em poucas palavras, pode-se dizer que este método baseia-se na resolução de problemas baseados em simulação dinâmica, transformando as variáveis originais em variáveis de posição, e incluindo variáveis artificiais ao sistema, as quais normalmente são adotadas como sendo normalmente distribuidas (NEAL *et al.*, 2011). Conforme Dias (2018), o HMC “simula-se o movimento do deslocamewnto de uma partícula sob uma energia potencial igual ao logaritmo negativo da densidade de probabilidade de interesse", em que cada iteração é aleatorizada a velocidade da partícula representando, assim, o movimento em um tempo específico. Consequentemente, obtendo a nova posição, um novo valor para a distribuição alvo, sendo aceito ou não como a regra do Metropolis.

Dinâmica Hamiltoniana

Inicialmente, para se introduzir o que seria a dinâmica hamiltoniana, deve-se entender, primeiramente, os conceitos de energia cinética e energia potencial, o que será feito com a utilização da Figura 16.

Assim como mostra a Figura 16, considere um cenário em que uma bola é posta sobre uma rampa, sendo desconsiderado o atrito. Pela Mecânica Clássica, à medida que a bola, após ser solta sobre a rampa, há uma conversão de energia potencial gravitacional em energia cinética até que a bola chegue ao ponto mais baixo da superfície. Já quando a bola começa a se deslocar para o ponto B da rampa, a energia cinética é novamente convertida em energia potencial gravitacional.

Como o atrito está sendo desconsiderado, a soma das duas energias é constante em qualquer ponto da rampa Danilevicz e Filho (). Desta forma, há a conservação de energia no

Figura 16 – Representação da energia cinética e potencial.

sistema.

Considere θ como sendo a localização da bola sobre a rampa, e o movimento da bola (massa $\times$ velocidade), dado por um vetor p. Adicionalmente, seja $U(\theta)$ a energia potencial, o qual terá sua energia proporcional à altura que a bola está posicionada naquele instante de tempo, e $K(p)$ a energia cinética, a qual será dada por $p^{'}p/2m$, sendo m a massa relativa a bola (XAVIER, 2019).

Em Xavier (2019) é ser visto que o deslocamento da bola em qualquer posição da rampa é dado de maneira constante. Porém, à medida que a inclinação da rampa começa a ascender, a energia cinética começa a diminuir ao mesmo tempo que a energia potencial começa a aumentar, até que energia cinética chegue a zero. A partir deste ponto em questão, ocorre o oposto com o aumento da energia cinética e o decréscimo da potencial.

Deste modo, a representação Hamiltoniana da energia total $H(\theta, p)$ é

$$H(\theta, p) = U(\theta) + K(p) = U(\theta) + p^{'}M^{-1}p,$$

em que a energia potencial, $U(\theta)$, é dada pelo o negativo do logaritmo da densidade de probabilidade da distribuição de θ a qual tem-se o interesse de obter valores dela. E sua energia cinética, $K(p)$, envolve a quantidade M, que é uma matriz definida positiva positiva, usualmente simétrica, e em muitos casos sendo o produto entre uma matriz identidade e um escalar (NEAL *et al.*, 2011).

Com a caracterização da função hamiltoniana, agora, tem-se o interesse em, que dado um instante de tempo t, identificar como θ e p se movem durante este tempo. Assim, determina-se seus movimentos através das derivadas parciais da função Hamiltoniana (equações de movimento):

$$\begin{aligned}\frac{\partial\theta}{\partial t} &= \frac{\partial H(\theta,p)}{\partial p} = \nabla_p K(p),\\ \frac{\partial p}{\partial t} &= -\frac{\partial H(\theta,p)}{\partial \theta} = \nabla_\theta U(\theta),\end{aligned}$$

sendo ∇_x o gradiente da função em relação à variável x. Por meio deste sistema de equações, é possível identificar e determinar a posição e velocidade de uma partícula durante um intervalo de tempo, ou seja, considerando s o tamanho deste intervalo, é determinado assim a velocidade e a posição para qualquer tempo, t, até o tempo atual, $t+s$ (NEAL *et al.*, 2011).

4.2.3 Algoritmo HMC

Pode ser visto em Burda e Maheu (2013) uma representação do método monte carlo hamiltoniano, panorama mais voltado ao lado probabilístico, com ênfase Bayesiana, o que diferencia de outros autores, os quais costumam manter o foco em terminologias das leis da física.

Considere o vetor de parâmetros $\theta \in \mathbb{R}^d$ e distribuição a posteriori $\pi(\theta)$. Admita também um vetor auxiliar de parâmetros, $p \in \mathbb{R}^d$, sendo normalmente distribuído com vetor de médias $\mathbf{0}$ e matriz de covariâncias $\mathbf{M}$, $p \sim \mathrm{N}_d(\mathbf{0},\mathbf{M})$, sendo independente de θ. Consequentemente, é visto que o negativo do logarítimo da distribuição conjunta entre (θ,p) é a própria equação Hamiltoniana.

$$H(\theta,p) = -\ln\big(\pi(\theta)\big) + \frac{1}{2}\ln\big((2\pi)^d|M|\big) + \frac{1}{2}p^{'}Mp.$$

E como usualmente acorre em muitas aplicações reais, a complexidade das equações de movimento (derivadas parciais da função hamiltoniana) é bem alta, impossibilitando assim a resolução deste sistema de forma analítica, fazendo-se necessário a utilização de métodos numéricos. O método de Störmer-Verlet é um dos métodos mais conhecidos na literatura, caracterizado por

$$\begin{aligned}p^{(t+\varepsilon/2)} &= p^{(t)} + (\varepsilon/2)\nabla_\theta \ln\big(\pi(\theta^{(t)})\big),\\ \theta^{(t+\varepsilon)} &= \theta^{(t)} + \varepsilon M^{-1} p^{(t+\varepsilon/2)},\\ p^{(t+\varepsilon)} &= p^{(t+\varepsilon/2)} + (\varepsilon/2)\nabla_\theta \ln\big(\pi(\theta^{(t+\varepsilon)})\big).\end{aligned}$$

O método utiliza a *log-posteriori*, o que faz com que aponte para áreas com maior probabilidade, resultando na distribuição de equilibro da cadeia de Markov de maneira mais acelerada.

Após a aplicação do método de Störmer-Verlet por L vezes, é esperado que a diferença dos Hamiltonianos seja aproximadamente perto de zero. Assim, também, pode-se definir o estado

final da trajetória como sendo $(\theta^{(*)}, p^{(*)})$. Após este processo para a resolução do sistema de forma numérica, é utilizado a regra de aceitação de Metropolis, para assim poder corrigir os possíveis erros inseridos no sistema. Logo, a probabilidade de aceitação proposta é dada por

$$\alpha\left[\left(\theta^{(0)}, p^{(0)}\right), \left(\theta^{(*)}, p^{(*)}\right)\right] = \min\left\{1, \quad \exp\left[H\left(\theta^{(0)}, p^{(0)}\right) - H\left(\theta^{(*)}, p^{(*)}\right)\right]\right\}. \tag{4.3}$$

Assim, como visto em Dias (2018), ε e M são de livre escolha. Porém, uma boa escolha dele pode agilizar o processo até o alcance para a distribuição estacionária. Comumente, a matriz M é tida como sendo uma matriz identidade ou determinística definida positiva, pois quanto mais aumenta a generalização dos casos, mais difícil fica sua especificação.

Deste modo, o algoritmo Monte Carlo Hamiltoniano é dado pela sequência das seguintes etapas:

1. Comece o processo com uma posição inicial, $\theta^{(0)}$, e tamanho dos passos ε;
2. Ative o contador com $i = 1$;
3. Simule $p \sim \mathrm{N}_d(\mathbf{0}, \mathbf{M})$ e $u \sim \mathrm{U}(0,1)$;
4. Faça $(\theta^I, p^I) = (\theta^{(i-1}, p^{(*)})$, $H^I = H(\theta^I, p^I)$;
5. Repita o método de Störmer-Verlet um número suficiente de vezes.
6. Para o estado final, faça $(\theta^F, p^F) = (\theta^{(*)}, p^{(*)})$ e $H(\theta^F, p^F)$;
7. Calcule $\alpha\left[\left(\theta^{(F)}, p^{(F)}\right), \left(\theta^{(I)}, p^{(I)}\right)\right] = \min\left\{1, \quad \exp\left[H^I - H^F\right]\right\}$;
8. Efetue $\theta^{(i)} = \theta^F$ com probabilida $u < \alpha(\cdot)$ e $\theta^{(i)} = \theta^I$, caso contrário.

Algo interessante é que, para o algoritmo de Metropolis Hasting, a taxa de aceitação do método fica em torno de 70%, ao mesmo tempo, que para boas escolhas de ε e M esta taxa de aceitação para o Monte Carlo Hamiltoniano oscila entre 80% e 99%. Porém, é importante lembrar que dentro do processo de simulação, o método HMC utiliza mais passos dentro de cada processo, o que nos leva a um processo mais demorado até a obtenção de uma amostra significativa, ainda mais quando a complexidade do modelo começa a ficar mais complexa (DIAS, 2018).

Ao implementar o algoritmo HMC basicamente é preciso escolher dois valores, o valor de ε de cada passo e o número de passos *leapfrog*. Enontra-se na literatura diversas abordagens para a seleção desses parâmetros. Além disso, em vez de fixar um número pré-determinado de passos *leapfrog*, pode-se recorrer a uma versão adaptativa do HMC chamada NUTS (No-U-Turn Sampler), conforme detalhado no trabalho de (HOFFMAN; GELMAN *et al.*, 2014). O NUTS ajusta automaticamente o número de passos em cada iteração.

4.2.4 Plataforma para modelagem estatística Stan

Comumente utilizado para se obter amostras MCMC ou para realização de otimizações, o *software* estatístico Stan Team (2014) tem características interessantes, que é a utilização de estratégias para de adaptar para a escolha de parâmetros ideais. A implementação dos modelos foi realizada via o pacote *rstan* (interface *R*) para o *Stan*, o qual realiza a comunicação entre as ferramentas de forma satisfatória.

Quando comparado a outros métodos como o Metropolis-Hastings ou Gibbs, mesmo que o método de Monte Carlo Hamiltoniano leve mais tempo por iteração, por motivos de como o algoritmo funciona, ele por si só constrói iterações com menor autocorrelação. Assim, enquanto o Stan funciona com 10.000 iterações dando uma séria pouco correlacionada, o BUGS exigiria 100.000 iterações para uma boa mistura. Além disto, alguns dos aspectos principais sobre o Stan pode ser vistos a seguir:

1. Linguagem e Ferramenta de Modelagem
    - Stan oferece uma linguagem de modelagem que permite aos usuários especificar distribuições probabilísticas para seus dados e parâmetros de interesse.
    - A linguagem é projetada para ser simples e expressiva, permitindo que os usuários definam seus modelos de forma declarativa e se concentrem nos aspectos estatísticos, sem se preocupar tanto com a implementação computacional.
2. Inferência Bayesiana
    - O Stan é voltado principalmente para a inferência bayesiana, onde a estimativa dos parâmetros é feita por meio da distribuição posterior, que combina a informação a priori com os dados observados.
    - Ele usa métodos como o Hamiltonian Monte Carlo (HMC) e sua variante No-U-Turn Sampler (NUTS), que são mais eficientes do que os métodos tradicionais de Monte Carlo via Cadeias de Markov (MCMC), especialmente para modelos de alta dimensão e com correlações entre parâmetros.
3. Flexibilidade de Modelagem
    - O Stan pode ajustar uma ampla gama de modelos, incluindo modelos lineares, modelos hierárquicos (ou modelos de efeitos mistos), modelos de regressão, séries temporais, análise de sobrevivência, entre outros.
    - A flexibilidade vem da sua capacidade de permitir que os usuários definam suas próprias distribuições de probabilidade e funções de verossimilhança, além de incorporar covariáveis, efeitos não-lineares e distribuições complexas.
4. Interpretabilidade

- O Stan é utilizado por meio de interfaces em várias linguagens de programação, como: RStan (para R), PyStan (para Python), CmdStan (interface de linha de comando), CmdStanR e CmdStanPy (interfaces mais recentes para R e Python)
- Isso facilita a integração do Stan em fluxos de trabalho de análise de dados nas linguagens mais populares.

5. Desempenho

- O Stan é otimizado para oferecer um bom desempenho computacional. Seu uso de HMC e NUTS permite uma exploração mais eficiente do espaço de parâmetros, reduzindo a autocorrelação entre amostras e acelerando a convergência em comparação com outros métodos de MCMC.
- No entanto, como o Stan realiza cálculos intensivos, ele pode ser computacionalmente caro para modelos muito grandes ou complexos. Para esses casos, é possível utilizar aproximações ou otimizações personalizadas.

endo apresentado alguns dos principais aspectos da modelagem com o Stan, pode-se agora explorar um exemplo simples de sua implementação. Em Team (2018), é apresentado um caso prático de aplicação envolvendo um modelo linear simples com um único preditor e intercepto, assumindo que os resíduos seguem uma distribuição normal. Esse modelo pode ser descrito da seguinte forma:

$$y_n = \alpha + \beta x_n + \varepsilon_n \quad \text{onde} \quad \varepsilon_n \sim \text{normal}(0, \sigma).$$

Isso é equivalente à seguinte amostragem envolvendo o resíduo,

$$y_n - (\alpha + \beta x_n) \sim \text{normal}(0, \sigma),$$

e, reduzindo ainda mais, para

$$y_n \sim \text{normal}(\alpha + \beta x_n, \sigma).$$

Esta última forma do modelo é codificada em Stan da seguinte maneira.

```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
}
parameters {
```

```
  real alpha;
  real beta;
  real<lower=0> sigma;
}
model {
  y ~ normal(alpha + beta * x, sigma);
}
```

Considere N observações, cada uma associada a uma covariável $x[n]$ e uma variável resposta $y[n]$. Os parâmetros correspondentes ao intercepto e à inclinação são representados por α e β, respectivamente. O modelo pressupõe a presença de um termo de ruído normalmente distribuído com escala σ. Esse modelo adota *priors* impróprios para os dois coeficientes de regressão, embora seja igualmente possível inferir distribuições para os coeficientes regressores.

Assim, o Stan se destaca como uma ferramenta robusta e flexível para a implementação de modelos bayesianos, oferecendo grande facilidade de uso para os usuários. Sua capacidade de flexibilização na modelagem faz com que seja amplamente utilizado por pesquisadores e profissionais que lidam com dados complexos em suas análises diárias.

4.3 Aplicação

Para a aplicação do conjunto de dados, definido na Seção 3.7, para isso utilizou-se o *RStan*, dentro do *Software R* para a implementação, o qual é baseado na metodologia de estimação do Hamiltoniano Monte Carlo. Para a aplicação serão considerados três casos distintos, assim como a aplicação frequentista, inicialmente será aplicado o modelo Weibull, com a acoplação das covariáveis apenas no parâmetro α, o segundo caso, considerando a distribuição Log-Normal com covariáveis acopladas no parâmetro μ, e por fim, novamente a distribuição Log-Normal, mas agora com as variáveis ligadas aos dois parâmetros associados a ela.

4.3.1 Modelo Weibull

Dado o modelo *Weibull* proposto anteriormente (Seção 3.3) e sua respectiva verossimilhança, foram definidas *prioris* gaussianas para os parâmetros regressores do modelo, sendo o intercepto $\beta_0 \sim N(0, 10^2)$ e $\beta_i \sim N(0, 10^2)$, $i = 1, 2$, sendo estes últimos ligadas as variáveis grupo e número de células CD4. Para o outro parâmetro da Weibull, γ, foi utilizada uma *priori* $\gamma \sim Gamma(50, 1)$, considerando a alta variância dessas distribuições, dado que não temos informações substanciais sobre esses parâmetros.

Além disso, foi escolhido um *burn-in* de 2 mil iterações, seguido pela amostragem de mais 2 mil valores. Na aplicação bayesiana, será inicialmente verificada a convergência do modelo para, em seguida, apresentar as inferências, como as estimativas dos parâmetros e os

intervalos de credibilidade. Para fins de facilidade na interpretação a Tabela 14 mostra um depara entre as variáveis utilizadas no modelo e como serão apresentadas nas figuras, as quais não foi-se possível alterar os valores dos eixos.

Tabela 14 – Correspondência entre os Parâmetros do Modelo Weibull e sua Representação nas Figuras

Parâmetro	**Figuras**
Intercepto	mu
CD4	beta[1]
Tratamento	beta[2]
γ	gamma

Com isso, o gráfico de trajetória das cadeias simuladas em um modelo bayesiano é crucial para verificar a convergência das cadeias MCMC. Esses gráficos mostram se as cadeias estão explorando o espaço de parâmetros de maneira adequada e convergindo para a mesma distribuição. Caso as cadeias se misturem bem e não apresentem comportamentos anômalos, podemos confiar nas estimativas dos parâmetros do modelo. Assim, esses gráficos desempenham um papel importante na garantia de que as inferências bayesianas sejam precisas e confiáveis.

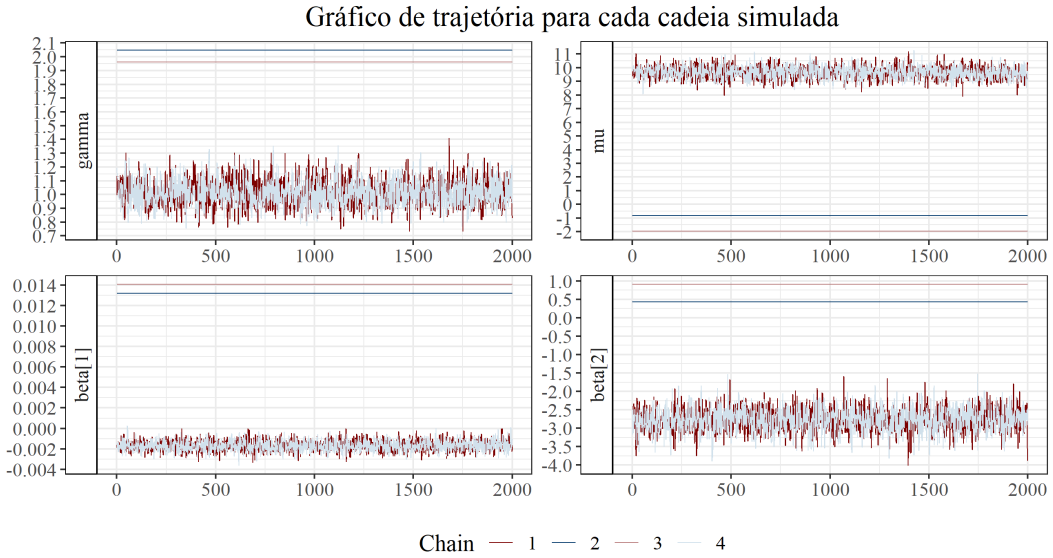

Figura 17 – Gráfico de trajetória de todas as cadeias simuladas para os parâmetros do modelo Weibull.

A Figura 17, mostra a trajetória das cadeias simuladas para os parâmetros do modelo Weibull sendo possível verificar que algumas cadeias caíram em estados absorventes, permanecendo praticamente estáticas e sem explorar adequadamente o espaço paramétrico. Esse comportamento compromete a convergência e a confiabilidade das inferências, indicando a necessidade de ajustes no modelo ou nos parâmetros do MCMC, como valores iniciais ou configuração do burn-in. Sem corrigir esses problemas, as estimativas e os intervalos de credibilidade não podem ser considerados confiáveis.

Calcular o $\hat{R}$ é essencial para verificar a convergência das cadeias MCMC. Esse índice mede a razão entre a variação entre as cadeias e a variação dentro de cada cadeia, permitindo avaliar se todas as cadeias convergiram para a mesma distribuição estacionária. Um valor de $\hat{R}$ próximo de 1 indica boa convergência, enquanto valores significativamente maiores sugerem que algumas cadeias ainda não exploraram adequadamente o espaço de parâmetros. Assim, o cálculo do $\hat{R}$ é uma etapa fundamental para garantir que as inferências obtidas sejam confiáveis e baseadas em cadeias bem misturadas e estabilizadas.

Tabela 15 – $\hat{R}$ estimado para cada parâmetro (Modelo Weibull).

Intercepto	CD4	Tratamento	γ
18.1389	22.6663	8.2289	7.8795

Complementando o que foi ditado acima, os valores de $\hat{R}$ apresentados indicam que as cadeias não convergiram, estando muito distantes do ideal de 1. Isso sugere problemas graves de exploração do espaço paramétrico, discrepâncias entre cadeias ou configuração inadequada do MCMC (Tabela 15).

Calcular o tamanho efetivo da amostra é importante para avaliar a qualidade das estimativas das cadeias MCMC. Ele mede a quantidade de amostras independentes equivalentes, ajudando a determinar se as cadeias exploraram bem o espaço de parâmetros.

Tabela 16 – Tamanho efetivo da amostra para cada parâmetro (Modelo Weibull).

Intercepto	CD4	Tratamento	γ
0.00025	0.00025	0.00025	0.00025

Os valores de tamanho efetivo da amostra apresentados na Tabela 16 são extremamente baixos (0.00025 para todos os parâmetros), indicando que as cadeias MCMC possuem alta autocorrelação e não estão fornecendo amostras independentes suficientes. Isso compromete seriamente a qualidade das estimativas e reforça a necessidade de ajustes no modelo ou na configuração do MCMC antes de proceder com inferências.

Por fim, deve-se calcular as medidas de autocorrelação da série de valores estimados para cada um dos parâmetros, afim de visualizar de existe ou não independência entre os valores estimados, ou seja, se existe ou não correlação serial entre os valores estimados.

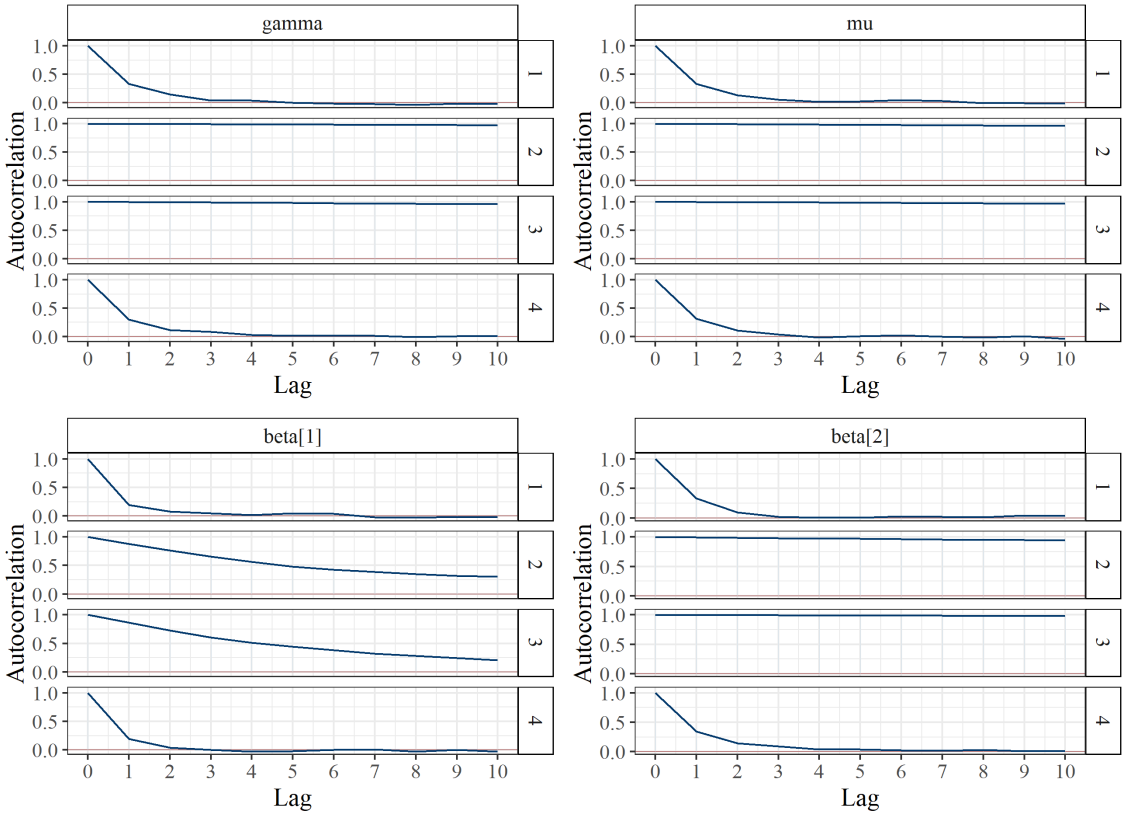

Figura 18 – Autocorrelação dos valores amostrados para cada parâmetro do modelo Weibull

A Figura 20 apresenta de autocorrelação apresentados mostram que os valores amostrados para todos os parâmetros do modelo Weibull apresentam alta autocorrelação, mesmo com pequenos lags. Essa alta dependência serial entre as amostras compromete a eficiência da amostragem e explica o baixo tamanho efetivo da amostra (Tabela 16). Com isto, mesmo tendo resultados não confiáveis serão apresentadas as estimativas do modelo estimado, assim como seu erro padrão e intervalo de credibilidade afim de comparação com os modelos seguintes.

Tabela 17 – Estimativas obtidas para o Modelo Weibull.

Parâmetro	**Média**	**Desv. Pad.**	**Lim. Inf.**	**Lim. Sup**
Intercepto	4.1299	3.9197	-1.9684	10.4025
CD4	0.006	0.0054	-0.0025	0.0141
Tratamento	-1.0523	1.2267	-3.307	0.9102
γ	1.5088	0.3514	0.8632	2.047

A Tabela 17 apresenta as estimativas obtidas para os parâmetros do modelo Weibull. Contudo, devido à falta de convergência das cadeias ($\hat{R}$ elevado e tamanho efetivo próximo a zero), os valores estimados, incluindo médias, desvios padrão e intervalos de credibilidade, não são confiáveis. A alta incerteza observada, especialmente nos parâmetros Intercepto e Tratamento, reforça a necessidade de ajustes no modelo e na configuração do MCMC para obter inferências

robustas. Para os modelos subsequentes também serão apresentados graficamente os intervalos de credibilidade e a matriz dos valores simulados para os parâmetros, porém como para este os resultados não foram satisfatórios, seja resguardada ao tempo de amostragem e taxa de aceitação.

Outra informação interessante que podemos retirar é o tempo levado para gerar o *burn-in* e amostragem para cada uma das 4 cadeias geradas, o que pode ser visto pela Tabela 18.

Tabela 18 – Tempo de amostragem e *burn-in* para cada uma das cadeias para o modelo Weibull.

	Burn-in	**Amostragem**
Cadeia 1	2.6170	1.1060
Cadeia 2	0.3540	0.8990
Cadeia 3	0.3240	0.3230
Cadeia 4	2.3880	1.1600

É possível verificar pela Tabela 18 que o tempo de aquecimento da primeira e quarta cadeia são maiores que a amostragem, já inverso para a segunda e terceira. O que não seria esperado, pois caso existisse uma boa convergência o tempo de *burn-in* seria superior, mostrando que os passos da amostragem são mais rápidos, pois já obteve a convergência da cadeia.

Além disso, como se trata de uma estimação via métodos MCMC, é interessante verificar a taxa de aceitação da cadeia.

Tabela 19 – Taxa de aceitação para cada uma das cadeias simuladas após o *burn-in*.

Cadeia 1	**Cadeia 2**	**Cadeia 3**	**Cadeia 4**
0,9331	0,9020	0,7912	0,9569

Nota-se pelos valores expostos na Tabela 19 que houve uma taxa de aceitação bem alta para quase todas as cadeias com valores acima de 90%, com exceção da cadeira 3.

4.3.2 Modelo Log-Normal com covariáveis no parâmetro de locação μ

Dado o modelo *Log-Normal* proposto anteriormente (Seção 3.4) e sua respectiva verossimilhança, foram definidas *prioris* gaussianas para os parâmetros regressores do modelo ligadas ao parâmetro μ, sendo o intercepto $\beta_0 \sim N(0, 10^2)$ e $\beta_i \sim N(0, 10^2)$, $i = 1, 2$, sendo estes últimos ligadas as variáveis grupo e número de células CD4. Para o outro parâmetro da Log-Normal, não foi feita a inclusão de covariáveis, σ, e foi utilizada uma *priori* $\gamma \sim Gamma(50, 1)$, considerando a alta variância dessas distribuições, dado que não temos informações substanciais sobre esses parâmetros.

Além disso, também foi escolhido um *burn-in* de 2 mil iterações, seguido pela amostragem de mais 2 mil valores. Na aplicação bayesiana, será inicialmente verificada a convergência do modelo para, em seguida, apresentar as inferências, como as estimativas dos parâmetros e os intervalos de credibilidade. Para fins de facilidade na interpretação a Tabela 20 mostra um depara entre as variáveis utilizadas no modelo e como serão apresentadas nas figuras, as quais não foi-se possível alterar os valores dos eixos.

Tabela 20 – Correspondência entre os Parâmetros do Modelo Log-Normal com acoplação no μ e sua Representação nas Figuras

Parâmetro	**Figuras**
Intercepto	intercept_loc
CD4	loc[1]
Tratamento	loc[2]
σ	sigma

Com isso, o gráfico de trajetória das cadeias simuladas em um modelo bayesiano é crucial para verificar a convergência das cadeias MCMC. Esses gráficos mostram se as cadeias estão explorando o espaço de parâmetros de maneira adequada e convergindo para a mesma distribuição. Caso as cadeias se misturem bem e não apresentem comportamentos anômalos, podemos confiar nas estimativas dos parâmetros do modelo. Assim, esses gráficos desempenham um papel importante na garantia de que as inferências bayesianas sejam precisas e confiáveis.

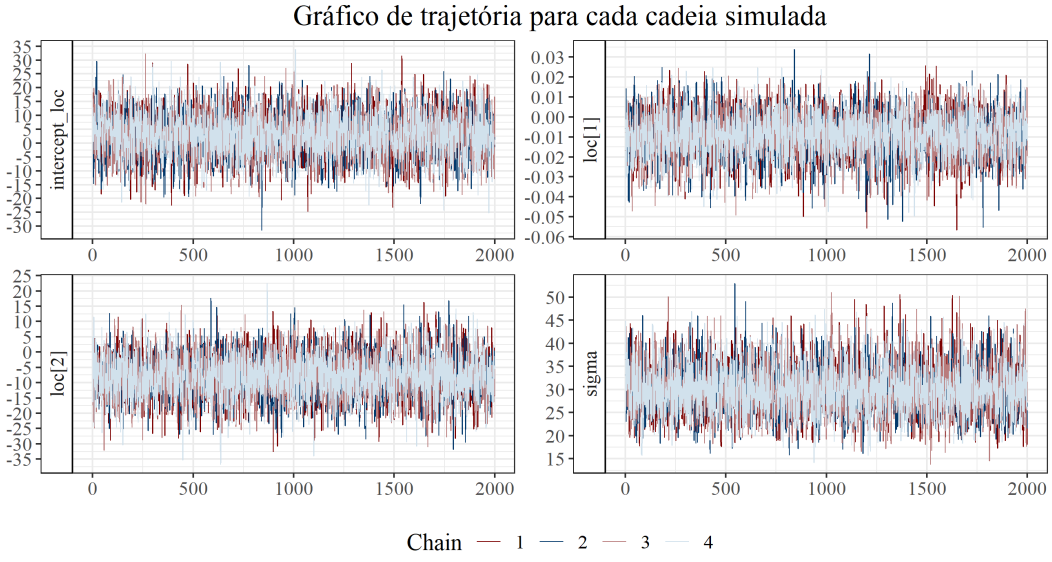

Figura 19 – Gráfico de trajetória de todas as cadeias simuladas para os parâmetros do modelo.

A Figura 19, mostra a trajetória das cadeias simuladas para os parâmetros do modelo bayesiano mostra um comportamento referente a uma boa convergência. As cadeias para cada parâmetro (intercepto, loc[1], loc[2], e σ) se misturam bem e exploram o espaço de parâmetros

de maneira semelhante, sem grandes desvios ou padrões anômalos. Isto é um indicativo que o modelo está bem ajustado e as inferências são confiáveis, mas para confirmar ainda realizaremos outras análises.

Calcular o $\hat{R}$ é essencial para verificar a convergência das cadeias MCMC. Esse índice mede a razão entre a variação entre as cadeias e a variação dentro de cada cadeia, permitindo avaliar se todas as cadeias convergiram para a mesma distribuição estacionária. Um valor de $\hat{R}$ próximo de 1 indica boa convergência, enquanto valores significativamente maiores sugerem que algumas cadeias ainda não exploraram adequadamente o espaço de parâmetros. Assim, o cálculo do $\hat{R}$ é uma etapa fundamental para garantir que as inferências obtidas sejam confiáveis e baseadas em cadeias bem misturadas e estabilizadas.

Tabela 21 – $\hat{R}$ estimado para cada parâmetro (Modelo acoplado no μ).

Intercepto	CD4	Tratamento	σ
1.000	1.000	1.000	1.000

O gráfico mostra que os valores de $\hat{R}$ para todos os parâmetros estão próximos de 1, indicando boa convergência das cadeias MCMC. Valores próximos de 1 sugerem que as cadeias se misturaram bem e que as estimativas dos parâmetros são confiáveis (Tabela 21).

Calcular o tamanho efetivo da amostra é importante para avaliar a qualidade das estimativas das cadeias MCMC. Ele mede a quantidade de amostras independentes equivalentes, ajudando a determinar se as cadeias exploraram bem o espaço de parâmetros.

Tabela 22 – Tamanho efetivo da amostra para cada parâmetro (Modelo acoplado no μ).

Intercepto	CD4	Tratamento	σ
0.609	0.572	0.700	0.645

A Tabela 22 mostra que os tamanhos efetivos da amostra (N_{eff}/N) para os parâmetros variam, com alguns valores próximos a 0.5, indicando uma boa eficiência na amostragem. Contudo, alguns parâmetros têm valores um pouco maiores, sugerindo que as cadeias podem não estar explorando tão bem o espaço de parâmetros para esses casos específicos.

Por fim, queremos calcular as medidas de autocorrelação da série de valores estimados para cada um dos parâmetros, afim de visualizar de existe ou não independência entre os valores estimados, ou seja, se existe ou não correlação serial entre os valores estimados.

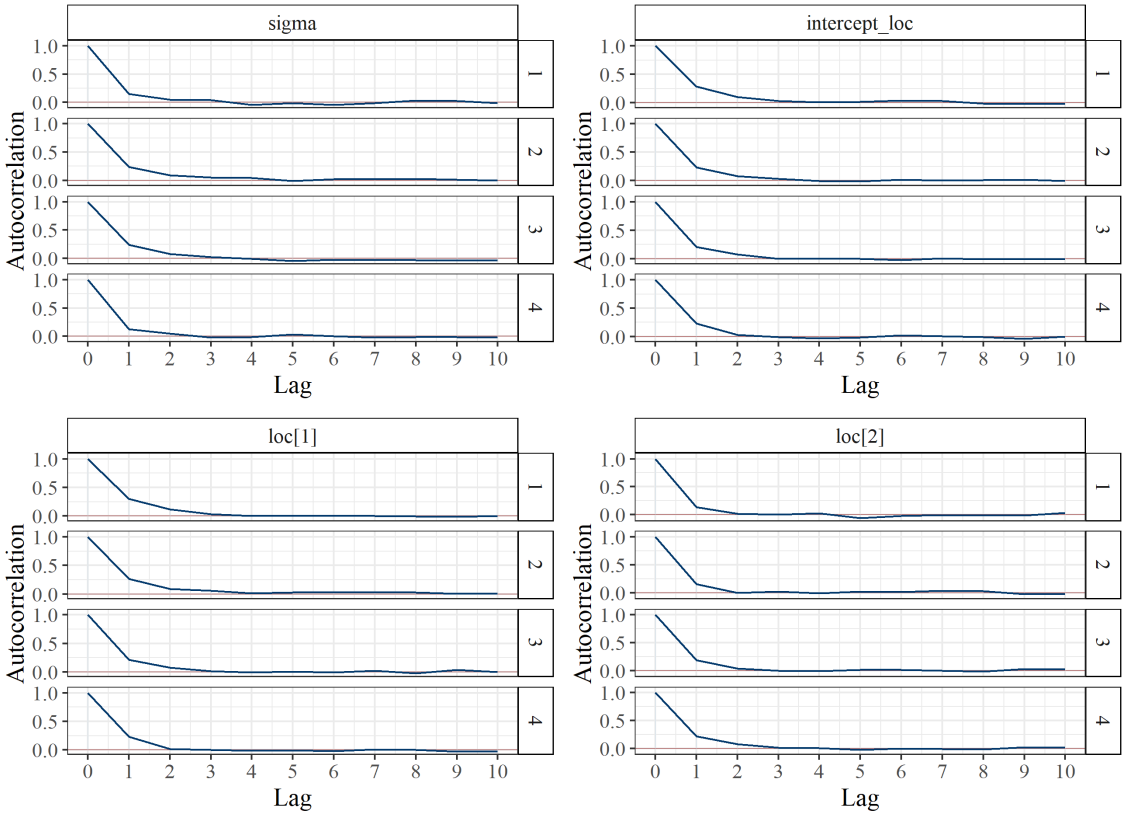

Figura 20 – Autocorrelação dos valores amostrados para cada um dos parâmetros do modelo com acoplação no μ.

A Figura 20 mostra que, para os parâmetros sigma, intercept_loc, loc[1], loc[2] e σ a autocorrelação diminui significativamente já no lag 1, indicando uma independência inicial, mesmo que fraca. Isso sugere que, mesmo com um pequeno deslocamento, as amostras sucessivas nas cadeias MCMC começam a ser aproximadamente independentes, o que é um sinal positivo da qualidade da amostragem e da eficiência do modelo. Com isto é visto as estimativas do modelo estimado, assim como seu erro padrão e intervalo de credibilidade.

Tabela 23 – Estimativas obtidas para o Modelo Log-Normal com covariáveis acopladas no μ.

Parâmetro	**Média**	**Desv. Pad.**	**Lim. Inf.**	**Lim. Sup**
Intercepto	2.9534	0.1134	-13.0676	18.2822
CD4	-0.0093	2.00e-4	-0.0328	0.0124
Tratamento	-7.9719	0.0953	-21.9436	6.1762
σ	29.8623	0.0742	20.4545	41.2659

A Tabela 23 apresenta as estimativas dos parâmetros de um Modelo Log-Normal com covariáveis acopladas no parâmetro μ. O intercepto apresentou uma média estimada de 2.9534, com ampla variação. O parâmetro CD4 teve uma média de -0.0093, mas não foi estatisticamente significativo, pois seu intervalo de credibilidade (-0.0328 a 0.0124) inclui o zero. O tratamento apresentou uma estimativa média de -7.9719, porém também não sendo significativo. O parâmetro

de dispersão σ apresentou uma média elevada (29.8623), sugerindo grande variabilidade nos dados.

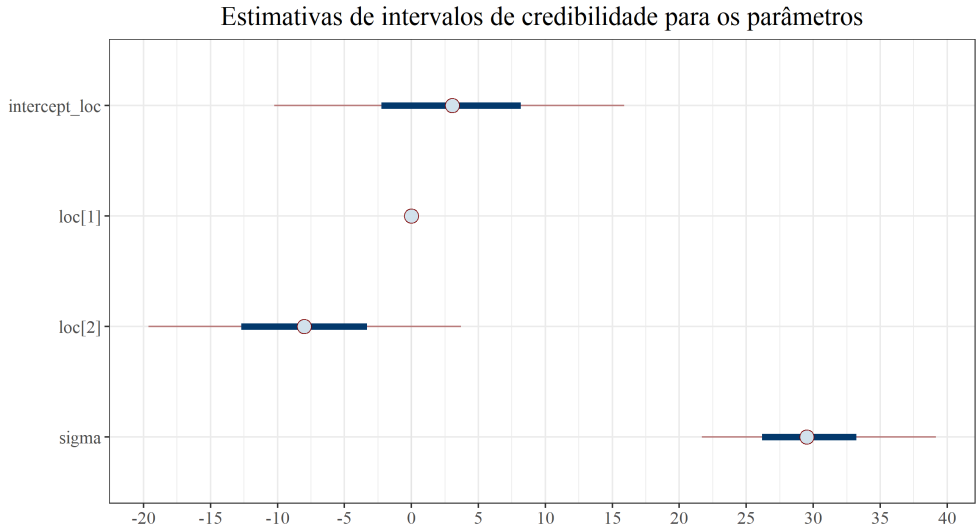

Figura 21 – Intervalos de confiança dos parâmetros estimados.

A Figura 21 mostra os intervalos de credibilidade para os parâmetros estimados do modelo bayesiano, revelando a incerteza associada a cada estimativa. Com isto nota-se que com exceção do parâmetro sigma, todos os outros obtiveram uma distribuição mais simetrica em torno do ponto médio, conmtudo todos contém o valor zero, mostrando não significância deles ao modelo. Além disso conseguimos ver a relação bivariadamente com a matriz de dispersão.

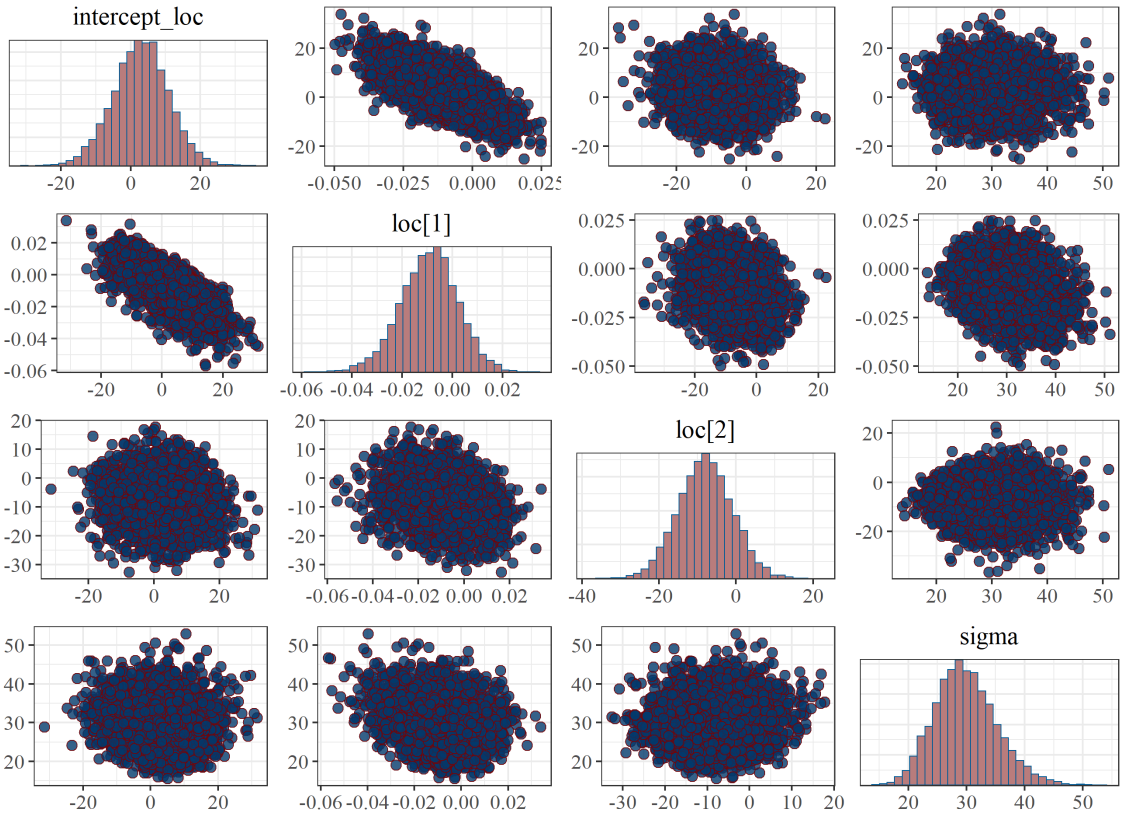

Figura 22 – Matriz de gráficos das cadeias simuladas.

A Figura 22 apresenta uma matriz de gráficos das cadeias simuladas para os parâmetros do modelo bayesiano: intercept_loc, loc[1], loc[2], e σ. Os histogramas na diagonal mostram distribuições bem definidas para cada parâmetro. Os gráficos de dispersão indicam que a maioria dos parâmetros são aproximadamente independentes, exceto por uma leve correlação negativa entre intercept_loc e loc[1]. Em geral, as distribuições e relações indicam que o modelo está bem ajustado e os parâmetros são estimados de forma confiável.

Outra informação interessante que podemos retirar é o tempo levado para gerar o *burn-in* e amostragem para cada uma das 4 cadeias geradas, o que pode ser visto pela Tabela 24.

Tabela 24 – Tempo de amostragem e *burn-in* para cada uma das cadeias.

	Burn-in	**Amostragem**
Cadeia 1	0.8580	0.2730
Cadeia 2	0.8040	0.3250
Cadeia 3	1.0110	0.2780
Cadeia 4	0.8290	0.2610

É possível verificar pela Tabela 24 que o tempo de aquecimento de todas as cadeias é bem superior ao tempo de amostragem em todos os casos, o que é normal também, visto que no começo ainda não existe a convergência da cadeia, fazendo assim os passos serem mais lentos.

Além disso, como se trata de uma estimação via métodos MCMC, é interessante verificar a taxa de aceitação da cadeia.

Tabela 25 – Taxa de aceitação para cada uma das cadeias simuladas após o *burn-in*.

Cadeia 1	Cadeia 2	Cadeia 3	Cadeia 4
0,9321	0,9464	0,9487	0,9296

Nota-se pelos valores expostos na Tabela 25 que houve uma taxa de aceitação bem alta para todas as cadeias com valores acima de 90%, o que nos da uma boa noção de que o modelo foi proposto de uma maneira eficiente

4.3.3 Modelo Log-Normal com covariáveis nos parâmetros de locação μ e escala σ

Dado o modelo *Log-Normal* proposto anteriormente (Seção 3.4) e sua respectiva verossimilhança, foram definidas *prioris* gaussianas para os parâmetros regressores do modelo, sendo o intercepto $\beta_{\mu 0} \sim N(0, 10^2)$ e $\beta_{\mu i} \sim N(0, 10^2)$, $i = 1, 2$, sendo estes últimos ligadas as variáveis grupo e número de células CD4. Para o outro parâmetro da Log-Normal, para este caso foi feita a inclusão de covariáveis, definindo as mesmas prioris definidas para aquelas acopladas ao μ, tendo alta variância em todas as distribuições *prioris*, dado que não temos informações substanciais sobre esses parâmetros.

Além disso, também foi escolhido um *burn-in* de 2 mil iterações, seguido pela amostragem de mais 2 mil valores. Na aplicação bayesiana, será inicialmente verificada a convergência do modelo para, em seguida, apresentar as inferências, como as estimativas dos parâmetros e os intervalos de credibilidade. Para fins de facilidade na interpretação a Tabela 26 mostra um depara entre as variáveis utilizadas no modelo e como serão apresentadas nas figuras, as quais não foi-se possível alterar os valores dos eixos.

Tabela 26 – Correspondência entre os Parâmetros do Modelo Log-Normal com acoplação no μ e σ e sua Representação nas Figuras

Parâmetro	Figuras
Intercepto (σ)	intercept_disp
Intercepto (μ)	intercept_loc
CD4 (σ)	disp[1]
Trat (σ)	disp[2]
CD4 (μ)	loc[1]
Trat (μ)	loc[2]

Com isso, o gráfico de trajetória das cadeias simuladas em um modelo bayesiano é crucial para verificar a convergência das cadeias MCMC. Esses gráficos mostram se as cadeias

estão explorando o espaço de parâmetros de maneira adequada e convergindo para a mesma distribuição. Caso as cadeias se misturem bem e não apresentem comportamentos anômalos, podemos confiar nas estimativas dos parâmetros do modelo. Assim, esses gráficos desempenham um papel importante na garantia de que as inferências bayesianas sejam precisas e confiáveis.

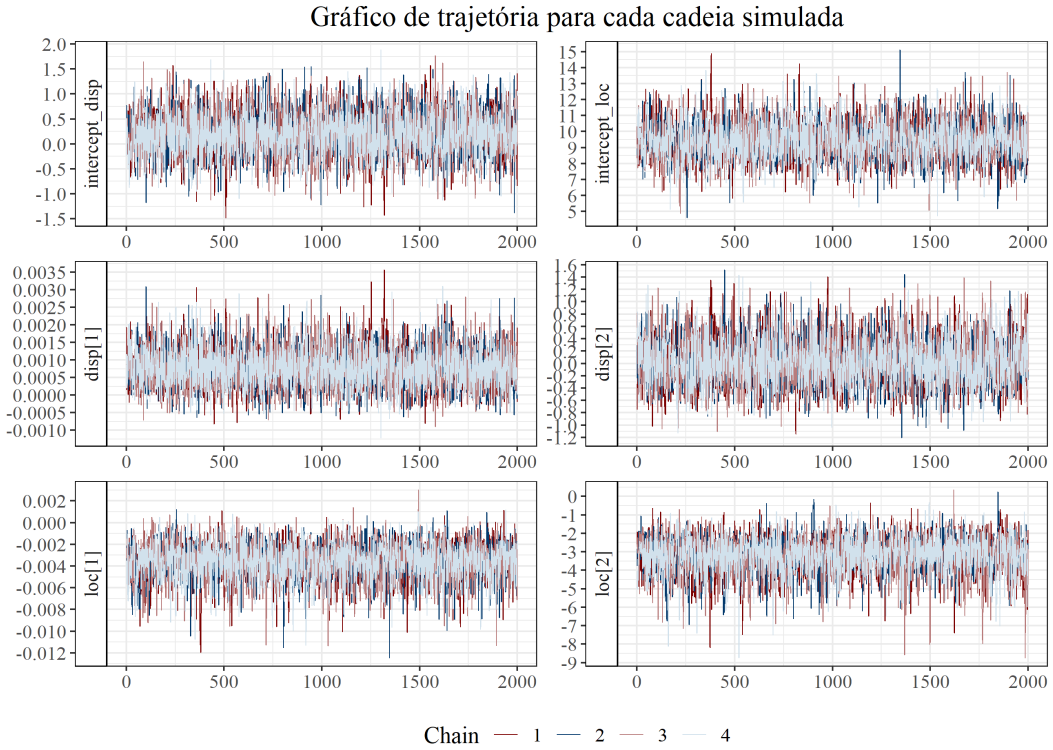

Figura 23 – Gráfico de trajetória de todas as cadeias simuladas para os parâmetros do modelo.

A Figura 19, mostra a trajetória das cadeias simuladas para os parâmetros do modelo bayesiano mostra um comportamento referente a uma boa convergência. As cadeias para cada parâmetro (intercept_loc, intercept_disp, loc[1], loc[2], disp[1] e disp[2]) se misturam bem e exploram o espaço de parâmetros de maneira semelhante, sem grandes desvios ou padrões anômalos. Isto é um indicativo que o modelo está bem ajustado e as inferências são confiáveis, mas para confirmar ainda realizaremos outras análises.

Calcular o $\hat{R}$ é essencial para verificar a convergência das cadeias MCMC. Esse índice mede a razão entre a variação entre as cadeias e a variação dentro de cada cadeia, permitindo avaliar se todas as cadeias convergiram para a mesma distribuição estacionária. Um valor de $\hat{R}$ próximo de 1 indica boa convergência, enquanto valores significativamente maiores sugerem que algumas cadeias ainda não exploraram adequadamente o espaço de parâmetros. Assim, o cálculo do $\hat{R}$ é uma etapa fundamental para garantir que as inferências obtidas sejam confiáveis e baseadas em cadeias bem misturadas e estabilizadas.

Tabela 27 – $\hat{R}$ estimado para cada parâmetro (Modelo acoplado no μ e σ).

Intercepto (σ)	Intercepto (μ)	CD4 (σ)	Trat (σ)	CD4 (μ)	Trat (μ)
1.0014	1.0002	1.0006	1.0023	1.0001	1.0013

O gráfico mostra que os valores de R chapéu para todos os parâmetros estão próximos de 1, indicando boa convergência das cadeias MCMC. Valores próximos de 1 sugerem que as cadeias se misturaram bem e que as estimativas dos parâmetros são confiáveis (Tabela 27).

Calcular o tamanho efetivo da amostra é importante para avaliar a qualidade das estimativas das cadeias MCMC. Ele mede a quantidade de amostras independentes equivalentes, ajudando a determinar se as cadeias exploraram bem o espaço de parâmetros.

Tabela 28 – Tamanho efetivo da amostra para cada parâmetro (Modelo acoplado no μ e σ).

Intercepto (σ)	Intercepto (μ)	CD4 (σ)	Trat (σ)	CD4 (μ)	Trat (μ)
0.4911	0.3913	0.5627	0.4136	0.4233	0.3818

A Tabela 28 apresenta os tamanhos efetivos da amostra para os parâmetros de um modelo com covariáveis acopladas nos parâmetros μ e σ. Os valores indicam boa eficiência nas estimativas, com destaque para o parâmetro Intercepto associado a σ (0.4911), que apresenta a maior proporção relativa de informações utilizáveis. Apesar de o tamanho efetivo variar entre os parâmetros, mesmo os menores valores, como o do Tratamento em μ (0.3818), são competitivos, evidenciando que este modelo oferece maior equilíbrio e eficiência geral em comparação aos modelos previamente apresentados. Isso reforça sua adequação em capturar a variabilidade e os efeitos analisados, mostrando-se mais robusto para a análise proposta.

Por fim, queremos calcular as medidas de autocorrelação da série de valores estimados para cada um dos parâmetros, afim de visualizar de existe ou não independência entre os valores estimados, ou seja, se existe ou não correlação serial entre os valores estimados.

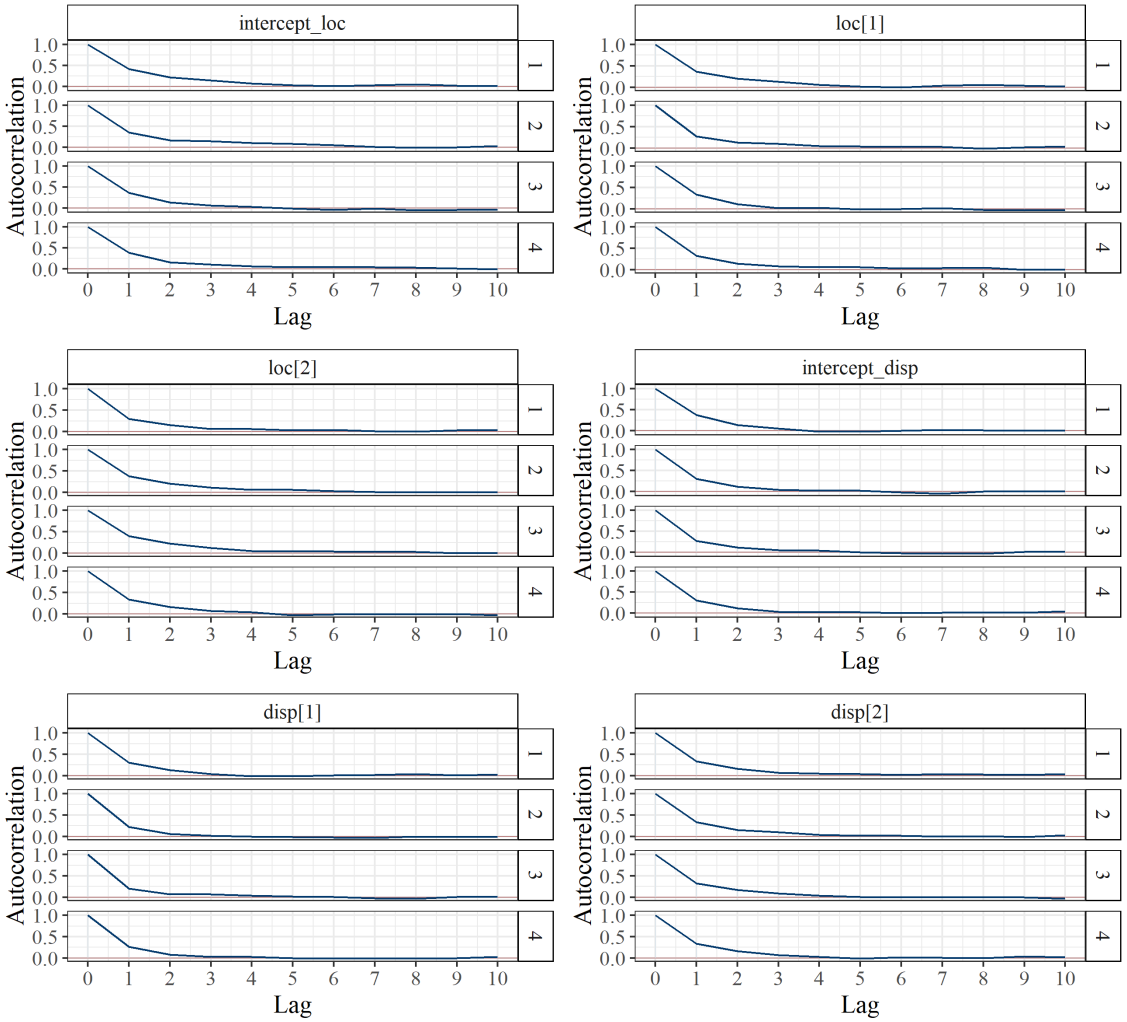

Figura 24 – Autocorrelação dos valores amostrados para cada um dos parâmetros do modelo com acoplação no μ e σ.

A Figura 24 mostra que, para todos os parâmetros, a autocorrelação diminui significativamente já no lag 1, indicando uma independência inicial, mesmo que fraca. Isso sugere que, mesmo com um pequeno deslocamento, as amostras sucessivas nas cadeias MCMC começam a ser aproximadamente independentes, o que é um sinal positivo da qualidade da amostragem e da eficiência do modelo. Com isto é visto as estimativas do modelo estimado, assim como seu erro padrão e intervalo de credibilidade.

Tabela 29 – Estimativas obtidas para do modelo com acoplação no μ e σ.

Parâmetro	**Média**	**Desv. Pad.**	**Lim. Inf.**	**Lim. Sup**
Intercepto (σ)	0.2188	0.0070	-0.6550	1.1072
Intercepto (μ)	9.4318	0.0213	7.2238	11.6436
CD4 (σ)	0.0008	0.0000	-0.0002	0.0020
Trat (σ)	0.0330	0.0066	-0.6582	0.8276
CD4 (μ)	-0.0036	0.0000	-0.0074	-0.0007
Trat (μ)	-3.2468	0.0182	-5.4733	-1.5228

A Tabela 29 apresenta as estimativas do modelo com acoplamento nos parâmetros μ e σ, com média, desvio padrão e intervalos de credibilidade. O Intercepto em μ (média: 9.4318) e o Tratamento em μ (média: -3.2468) destacam-se por seus intervalos de credibilidade que não incluem zero, indicando significância estatística. O parâmetro CD4 em μ também é significativo, com um efeito pequeno, mas negativo (-0.0036). Por outro lado, os parâmetros associados a σ — Intercepto, CD4 e Tratamento — apresentam intervalos de credibilidade que incluem zero, sugerindo que seus efeitos não são estatisticamente significativos neste modelo. Esses resultados evidenciam que os efeitos mais relevantes estão associados aos parâmetros ligados ao μ, especialmente ao Tratamento, que mostra um impacto marcante no modelo.

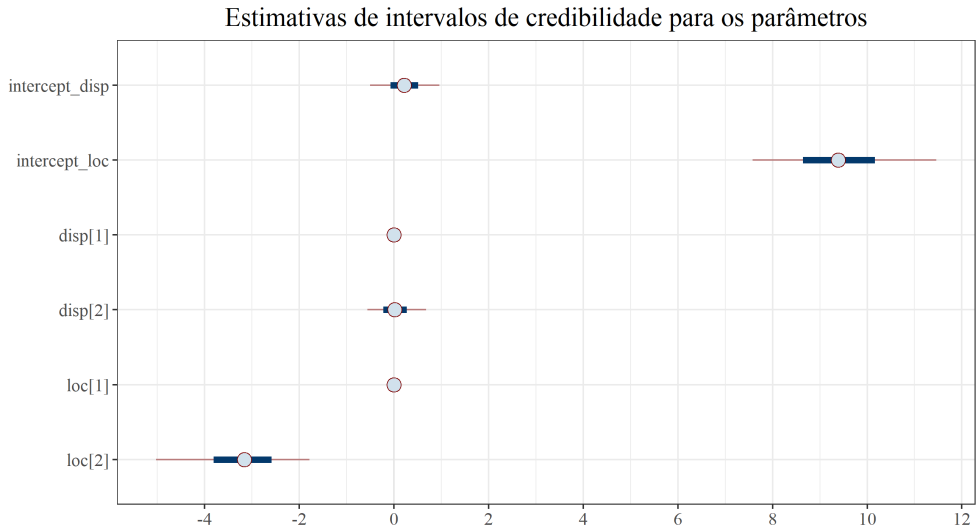

Figura 25 – Intervalos de confiança dos parâmetros estimados do modelo com acoplação no μ e σ.

A Figura 25 complementa a análise ao ilustrar graficamente os intervalos de credibilidade para os parâmetros. Observa-se que os parâmetros associados ao μ, como o Intercepto e loc[2], possuem intervalos que não incluem zero, reforçando sua significância. Por outro lado, os parâmetros relacionados ao σ, como disp[1] e disp[2], apresentam intervalos que incluem zero, corroborando sua menor relevância no modelo. Essa visualização destaca a robustez dos parâmetros ligados ao μ.

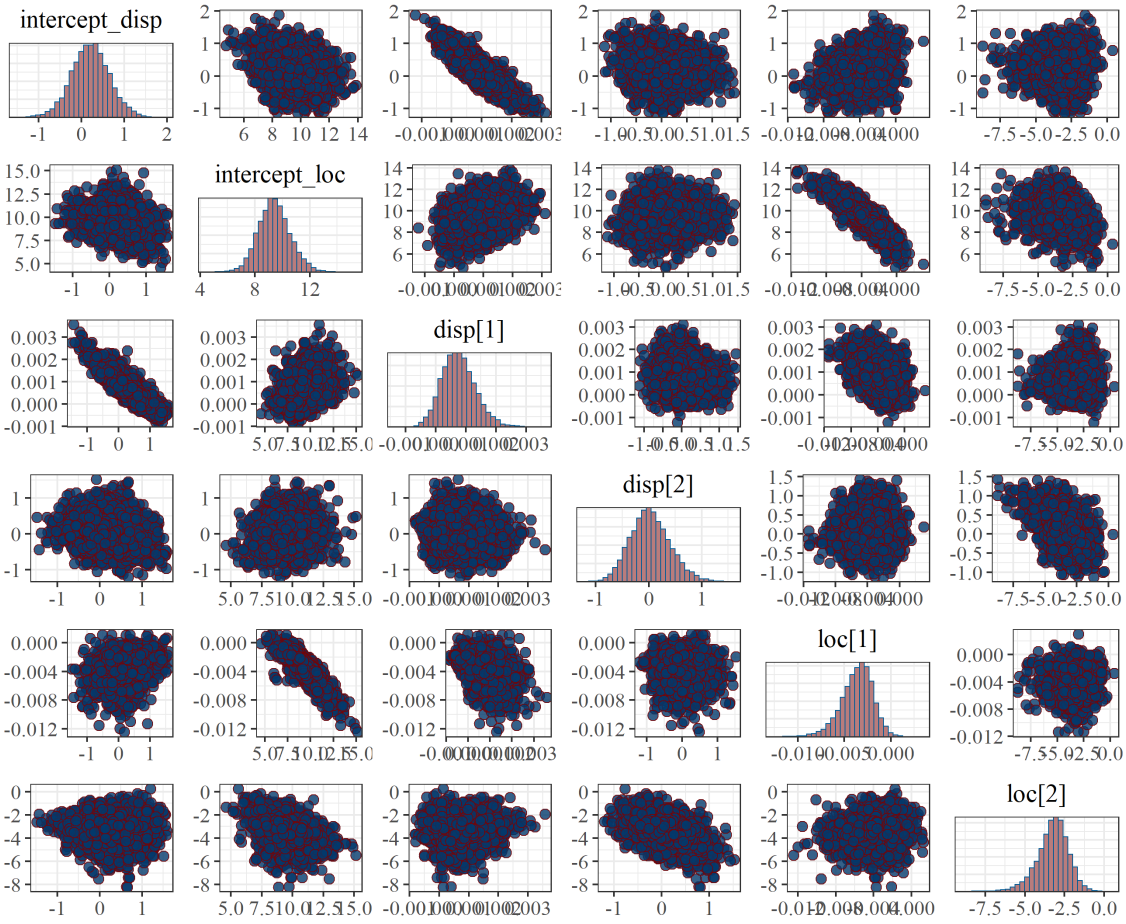

Figura 26 – Matriz de gráficos das cadeias simuladas do modelo com acoplação no μ e σ.

A Figura 26 mostra a distribuição posterior e as correlações entre os parâmetros intercept_disp, intercept_loc, disp[1], disp[2], loc[1], e loc[2]. Os histogramas na diagonal indicam distribuições bem definidas para cada parâmetro. Os gráficos de dispersão sugerem que a maioria dos parâmetros são aproximadamente independentes, com algumas correlações notáveis, como entre intercept_disp e intercept_loc, e entre disp[1] e loc[1], ou seja seus interceptos com o número de células CD4. Isso sugere que, em geral, os parâmetros estão bem estimados e a interação entre eles é baixa, indicando um bom ajuste do modelo.

Outra informação interessante que podemos retirar é o tempo levado para gerar o *burn-in* e amostragem para cada uma das 4 cadeias geradas, o que pode ser visto pela Tabela 30.

Tabela 30 – Tempo de amostragem e *burn-in* para cada uma das cadeias do modelo com acoplação no μ e σ.

	Burn-in	**Amostragem**
Cadeia 1	7.319	1.405
Cadeia 2	3.771	1.591
Cadeia 3	3.334	1.611
Cadeia 4	3.668	1.372

É possível verificar pela Tabela 30 que o tempo de aquecimento de todas as cadeias é bem superior ao tempo de amostragem em todos os casos, o que é normal também, visto que no começo ainda não existe a convergência da cadeia, fazendo assim os passos serem mais lentos. Contudo, nota-se que por existir mais parâmetros no modelo, o tempo é superior aos outros modelos apresentados.

Além disso, como se trata de uma estimação via métodos MCMC, é interessante verificar a taxa de aceitação da cadeia.

Tabela 31 – Taxa de aceitação para cada uma das cadeias simuladas após o *burn-in* do modelo com acoplação no μ e σ.

Cadeia 1	**Cadeia 2**	**Cadeia 4**	**Cadeia 5**
0.915	0.928	0.9425	0.9357

Nota-se pelos valores expostos pela Tabela 31 que houve uma taxa de aceitação bem alta para todas as cadeias com valores acima de 90%, o que nos da uma boa noção de que o modelo foi proposto de uma maneira eficiente

4.4 Considerações finais

Este capítulo apresentou uma visão abrangente sobre os métodos bayesianos, incluindo a definição de prioris (subjetivas, objetivas e suas misturas) e técnicas de amostragem baseadas em métodos Monte Carlo via Cadeia de Markov (MCMC), como os algoritmos Metropolis-Hastings e Hamiltoniano (HMC). Além disso, foi introduzida a utilização da plataforma Stan para a modelagem estatística, destacando sua eficiência na implementação de modelos complexos.

Na aplicação prática, os modelos Weibull e Log-Normal foram comparados em termos de convergência e significância dos parâmetros. O modelo Weibull apresentou dificuldades de convergência, com cadeias caindo em estados absorventes. Por outro lado, os modelos Log-Normais demonstraram melhor desempenho, com destaque para o modelo com acoplação nos dois parâmetros (μ, σ), que teve a melhor convergência geral. Embora os parâmetros associados ao σ não tenham mostrado significância estatística, os parâmetros ligados ao μ foram significativos, reforçando a adequação deste modelo. Assim, o modelo Log-Normal com acoplação nos dois parâmetros se destacou como o mais robusto e confiável, sendo o mais indicado para o contexto analisado.

CAPÍTULO

5

CONCLUSÕES E PROPOSTAS FUTURAS

Os pacientes que possuem o vírus HIV precisam realizar o controle do número de células CD4 infectadas pelo vírus, ou seja, o controle da carga viral, sendo este monitoramento realizado por equipamentos laboratoriais. No entanto, esses equipamentos possuem um limite mínimo de detecção, de modo que, abaixo de determinado valor, não é possível obter medições exatas, apenas indicando que o valor é menor do que um limite, como, por exemplo, "$<$400". Para lidar com esta característica, foi desenvolvido um modelo estatístico para dados com múltiplos níveis de censura à esquerda.

Neste trabalho, foi estudada a modelagem estatística baseada em modelos de regressão com a inserção de covariáveis em distribuições Weibull e Log-Normal, explorando a peculiaridade de diferentes níveis de censura à esquerda incorporados por meio da função de verossimilhança, que considera a contribuição de cada observação, seja pelo produto das funções de distribuição acumulada para dados censurados, seja pela função densidade para observações completas.

Tradicionalmente, a transformação logarítmica era amplamente utilizada na variável resposta "carga viral"em estudos relacionados, visando estabilizar a variância e facilitar a modelagem. No entanto, o objetivo deste trabalho foi desenvolver um modelo que permitisse a análise diretamente na escala original dos dados, sem a necessidade dessa transformação. Esse objetivo foi alcançado com sucesso por meio da utilização do modelo Log-Normal, que se mostrou adequado para lidar com os dados na escala original, preservando as características intrínsecas da variável e permitindo interpretações mais diretas e intuitivas dos resultados.

Foi possível demonstrar a aplicabilidade dos modelos utilizando dados reais de um estudo clínico conduzido no Laboratório de Investigação em Dermatologia e Imunodeficiências da Faculdade de Medicina da Universidade de São Paulo, no âmbito do projeto desenvolvido pelo Dr. Duarte e Dr. Silva. Nesta etapa, com base na metodologia frequentista, também foi conduzido um estudo de simulação para investigar a consistência e eficiência dos estimadores

de máxima verossimilhança (EMVs), considerando diferentes tamanhos amostrais. Para isso, foram analisados três cenários e avaliados três critérios: viés, raiz quadrada do erro quadrático médio (EQM) e a probabilidade de cobertura (PC). Os resultados das simulações frequentistas corroboraram a eficiência da metodologia, destacando sua robustez. Além disso, três modelos foram comparados, e os resultados indicaram que o modelo *Log-Normal* (μ, σ) é o mais adequado, apresentando os menores valores de *AIC*, *BIC* e *AICc*, bem como melhor adequação dos resíduos. Isso demonstra o equilíbrio ideal entre ajuste e complexidade, reforçando a aplicabilidade do modelo proposto.

Além disso, foi realizada uma análise de convergência via inferência bayesiana, destacando-se a superioridade dos modelos Log-Normais em relação ao Weibull, que apresentou dificuldades de convergência devido a cadeias caindo em estados absorventes. O modelo Log-Normal com acoplação nos dois parâmetros (μ, σ) mostrou-se mais robusto, com melhor convergência geral, especialmente para os parâmetros associados ao μ, embora os parâmetros ligados ao σ não tenham apresentado significância. Este modelo também apresentou melhores ajustes segundo os critérios *AIC*, *BIC* e *AICc*, além de maior adequação dos resíduos.

Em síntese, este trabalho contribuiu significativamente ao apresentar um modelo estatístico que permite a análise de dados na escala original, algo de grande relevância em estudos clínicos, como no controle da carga viral de pacientes com HIV. A superação da necessidade de transformações logarítmicas, aliada à robustez metodológica evidenciada pelos resultados do modelo Log-Normal com múltiplos níveis de censura, reforça a aplicabilidade e a precisão da abordagem proposta. Além disso, o desenvolvimento e a avaliação de novas metodologias para lidar com diferentes tipos de dados censurados mostram-se essenciais para expandir as fronteiras do conhecimento estatístico. Este trabalho, ao mesmo tempo em que resolve um problema específico, oferece uma base sólida para futuros estudos e aplicações, incentivando a criação de soluções estatísticas mais adaptadas a contextos diversos e desafiadores.

5.1 Propostas futuras

Como proposta para trabalhos futuros, sugere-se a incorporação de uma estrutura longitudinal ao modelo, considerando que os dados disponíveis apresentam históricos das observações dos pacientes ao longo do tempo. Diferentemente deste trabalho, que utilizou apenas um recorte referente à última observação de cada paciente, a análise longitudinal permitiria explorar as dinâmicas e padrões temporais das variáveis, oferecendo uma visão mais detalhada sobre a evolução da carga viral e suas relações com covariáveis, como o número de células CD4 e o grupo de tratamento. Essa abordagem não apenas aumentaria o poder explicativo e preditivo do modelo, mas também poderia trazer *insights* importantes sobre a progressão do HIV e a efetividade das intervenções ao longo do tempo, contribuindo para análises clínicas mais robustas e completas.

REFERÊNCIAS

BERGER, J. O. **Statistical decision theory and Bayesian analysis**. [S.l.]: Springer Science & Business Media, 2013. Citado na página 77.

BICKEL, P. J.; DOKSUM, K. A. **Mathematical statistics: basic ideas and selected topics, volumes I-II package**. [S.l.]: Chapman and Hall/CRC, 2015. Citado na página 37.

BOLFARINE, H.; SANTOS, B.; CORREIA, L.; MARTINEZ, G.; GOMEZ, H.; BAZAN, J. Modelos de regressao com respostas limitadas e censuradas. **13a Escola de Modelos de Regressao, Maresias, Sao Sebastiao, SP**, 2013. Citado na página 24.

BRASIL. Manual técnico para o diagnóstico da infecção pelo hiv. **Brasília: Ministério da Saúde**, Ministério da Saúde, 2013. Citado nas páginas 23 e 63.

BURDA, M.; MAHEU, J. M. Bayesian adaptively updated hamiltonian monte carlo with an application to high-dimensional bekk garch models. **Studies in Nonlinear Dynamics & Econometrics**, De Gruyter, v. 17, n. 4, p. 345–372, 2013. Citado na página 81.

CALDERHEAD, B. **Differential geometric MCMC methods and applications**. Tese (Doutorado) — University of Glasgow, 2011. Citado na página 79.

CAMERON, A. C.; TRIVEDI, P. K. **Microeconometrics: methods and applications**. [S.l.]: Cambridge university press, 2005. 137 p. Citado nas páginas 36 e 37.

CANINI, S. R. M. d. S.; REIS, R. B. d.; PEREIRA, L. A.; GIR, E.; PELÁ, N. T. R. Qualidade de vida de indivíduos com hiv/aids: uma revisão de literatura. **Revista Latino-Americana de Enfermagem**, SciELO Brasil, v. 12, p. 940–945, 2004. Citado na página 23.

CASELLA, G. **Berger. RL, Statistical Inference**. [S.l.]: Duxbury, Thomson Learning Inc., Pacific Grove, CA,, 2002. Citado nas páginas 45 e 46.

CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Cengage Learning, 2021. Citado na página 37.

CECCOTTI, T. B. Intervalos de confiança baseados em deviance para os hiperparâmetros em modelos estruturais. Universidade Federal de Minas Gerais, 2015. Citado nas páginas 35 e 36.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006. Citado nas páginas 27, 30, 32, 33 e 34.

____. **Análise de Sobrevivência Aplicada**. São Paulo: Edgard Blucher, 2006. 137 p. (Série do livro, 15). Bibliografia: p. 131–132. ISSN XXXX-XXXX. ISBN XX-XXX-XXXX-X. Citado nas páginas 25 e 52.

DANILEVICZ, I. M.; FILHO, W. dos R. M. Monte carlo hamiltoniano e stan. Citado na página 79.

DEGROOT, M. H. Probability and statistics. **(No Title)**, 2002. Citado na página 76.

DIAS, D. de S. Inferência bayesiana em modelos de volatilidade estocástica usando métodos de monte carlo hamiltoniano. 2018. Citado nas páginas 79 e 82.

DUANE, S.; KENNEDY, A. D.; PENDLETON, B. J.; ROWETH, D. Hybrid monte carlo. **Physics letters B**, Elsevier, v. 195, n. 2, p. 216–222, 1987. Citado na página 79.

ESTEVAM, A. C. Modelagem estatística para análise de dados imobiliários completos e com censura à esquerda. Universidade Federal de São Carlos, 2014. Citado na página 33.

FERREIRA, D. F. **Estatística multivariada**. [S.l.]: Editora Ufla Lavras, 2008. 662 p. Citado na página 38.

HANSON, K. M.; CUNNINGHAM, G. S. Posterior sampling with improved efficiency. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 1998: Image Processing**. [S.l.], 1998. v. 3338, p. 371–382. Citado na página 78.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. Oxford University Press, 1970. Citado na página 78.

HOFFMAN, M. D.; GELMAN, A. *et al.* The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. **J. Mach. Learn. Res.**, v. 15, n. 1, p. 1593–1623, 2014. Citado na página 82.

HUGHES, J. P. Mixed effects models with censored data with application to hiv rna levels. **Biometrics**, Wiley Online Library, v. 55, n. 2, p. 625–629, 1999. Citado na página 24.

JACQMIN-GADDA, H.; THIÉBAUT, R.; CHÊNE, G.; COMMENGES, D. Analysis of left-censored longitudinal data with application to viral load in hiv infection. **Biostatistics**, Oxford University Press, v. 1, n. 4, p. 355–368, 2000. Citado na página 23.

KINAS, P. G.; ANDRADE, H. A. **Introdução à análise bayesiana (com R)**. [S.l.]: Consultor Editorial, 2017. Citado na página 75.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. [S.l.]: John Wiley & Sons, 2011. v. 362. Citado na página 30.

LYLES, R. H.; WILLIAMS, J. K.; CHUACHOOWONG, R. Correlating two viral load assays with known detection limits. **Biometrics**, Wiley Online Library, v. 57, n. 4, p. 1238–1244, 2001. Citado na página 24.

MARTZ, H. F.; WALLER, R. Bayesian reliability analysis. **JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, 1982, 704**, 1982. Citado na página 31.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **The journal of chemical physics**, American Institute of Physics, v. 21, n. 6, p. 1087–1092, 1953. Citado na página 77.

METROPOLIS, N.; ULAM, S. The monte carlo method. **Journal of the American statistical association**, Taylor & Francis, v. 44, n. 247, p. 335–341, 1949. Citado na página 78.

MIGON, H. S.; GAMERMAN, D.; LOUZADA, F. **Statistical inference: an integrated approach**. [S.l.]: CRC press, 2014. Citado na página 35.

NEAL, R. M. *et al.* Mcmc using hamiltonian dynamics. **Handbook of markov chain monte carlo**, v. 2, n. 11, 2011. Citado nas páginas 79, 80 e 81.

O'HAGAN, A.; FORSTER, J. J. **Kendall's advanced theory of statistics, volume 2B: Bayesian inference**. [S.l.]: Arnold, 2004. v. 2. Citado na página 76.

OSPINA, R.; FERRARI, S. L. A general class of zero-or-one inflated beta regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 56, n. 6, p. 1609–1623, 2012. Citado na página 35.

PAULINO, C. D. M.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística bayesiana**. [S.l.: s.n.], 2018. Citado nas páginas 76 e 77.

PORTUGAL, M. S. Notas introdutórias sobre o princípio de máxima verossimilhança: Estimação e teste de hipóteses. **DECON/UFRGS, Porto Alegre, Abril**, 1995. Citado na página 37.

RESENDE, M. D. V. de R. **Matemática e estatística na análise de experimentos e no melhoramento genético**. [S.l.]: Embrapa Florestas, 2007. Citado na página 37.

SAMSON, A.; LAVIELLE, M.; MENTRÉ, F. Extension of the saem algorithm to left-censored data in nonlinear mixed-effects model: Application to hiv dynamics model. **Computational Statistics & Data Analysis**, Elsevier, v. 51, n. 3, p. 1562–1574, 2006. Citado na página 24.

SHAH, I.; NADIGER, M. Long term non progressors (ltnp) with vertically infected hiv children-a report from western india. **The Indian Journal of Medical Research**, Wolters Kluwer–Medknow Publications, v. 137, n. 1, p. 210, 2013. Citado na página 62.

SOLOMON, G.; WEISSFELD, L. Pseudo maximum likelihood approach for the analysis of multivariate left-censored longitudinal data. **Statistics in medicine**, Wiley Online Library, v. 36, n. 1, p. 81–91, 2017. Citado na página 24.

TEAM, S. **Stan: a C++ library for probability and sampling, version 2.0**. 2014. Citado na página 83.

TEAM, S. D. **Linear Regression - Stan User's Guide**. 2018. Accessed: 2024-09-24. Disponível em: <https://mc-stan.org/docs/2_18/stan-users-guide/linear-regression.html>. Citado na página 84.

TOBIN, J. Estimation of relationships for limited dependent variables. **Econometrica: journal of the Econometric Society**, JSTOR, p. 24–36, 1958. Citado na página 24.

TORRES, M. de P.; STUTZ, L. T.; NETO, A. J. da S. Comparação entre os métodos metropolis-hastings e monte carlo hamiltoniano na identificação de propriedades termofísicas em problemas de condução de calor. **Revista Mundi Engenharia, Tecnologia e Gestão (ISSN: 2525-4782)**, v. 4, n. 5, 2019. Citado na página 78.

TURNBULL, B. W. The empirical distribution function with arbitrarily grouped, censored and truncated data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 38, n. 3, p. 290–295, 1976. Citado na página 30.

WANG, W.-L.; LIN, T.-I.; LACHOS, V. H. Extending multivariate-t linear mixed models for multiple longitudinal data with censored responses and heavy tails. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 27, n. 1, p. 48–64, 2018. Citado na página 24.

WEIBULL, W. A statistical theory of the strength of materials, 1939. **Generalstabens Litografiska Anstalts Förlag**, 1939. Citado na página 31.

____. A statistical distribution function of wide applicability. journal of applied mechanics 18: 293-297. **Statistical and Computational Analysis**, v. 291, 1951. Citado na página 31.

WILKS, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. **The annals of mathematical statistics**, JSTOR, v. 9, n. 1, p. 60–62, 1938. Citado na página 38.

WU, H.; DING, A. A.; GRUTTOLA, V. D. Estimation of hiv dynamic parameters. **Statistics in medicine**, Wiley Online Library, v. 17, n. 21, p. 2463–2485, 1998. Citado na página 23.

WU, T.; FENG, C.; LI, L. Residual analysis for regression with censored data via randomized survival probabilities. 2019. Citado nas páginas 25 e 53.

XAVIER, C. M. Métodos de monte carlo hamiltoniano aplicados em modelos garch. Universidade Federal de São Carlos, 2019. Citado na página 80.

APÊNDICE

A

RESULTADOS DA SIMULAÇÃO

A.1 Simulação Weibull

Tabela 32 – Viés, raiz quadrada do erro quadrático médio e probabilidade de cobertura (PC) do estimador de máxima verossimilhança de ($\hat{\alpha}$) do modelo de múltiplos níveis de censura Weibull utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Métrica	Tamanho da Amostra	Cenário 1	Cenário 2	Cenário 3
Viés	50	0,009678	0,00141	-0,002161
	100	0,011658	0,002898	-0,000661
	250	-0,001839	-0,010299	-0,014084
	500	-0,001937	-0,010907	-0,014765
REQM	50	0,000977	0,000903	0,000915
	100	0,000579	0,000461	0,000457
	250	0,000182	0,000288	0,000184
	500	9,3e-05	0,00021	0,00031
PC	50	0,941	0,947	0,944
	100	0,956	0,955	0,955
	250	0,947	0,946	0,939
	500	0,954	0,956	0,955

Tabela 33 – Viés, raiz quadrada do erro quadrático médio e probabilidade de cobertura (PC) do estimador de máxima verossimilhança de ($\hat{\gamma}$) do modelo de múltiplos níveis de censura Weibull utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Métrica	Tamanho da Amostra	Cenário 1	Cenário 2	Cenário 3
Viés	50	0,138499	0,077071	0,050249
	100	0,070599	0,008162	-0,018661
	250	0,022002	-0,037786	-0,065854
	500	0,012938	-0,050314	-0,078863
REQM	50	0,37382	0,34759	0,338546
	100	0,174957	0,163867	0,161429
	250	0,066986	0,065503	0,067345
	500	0,033175	0,034331	0,037481
PC	50	0,947	0,94	0,936
	100	0,954	0,946	0,945
	250	0,956	0,94	0,932
	500	0,949	0,94	0,93

A.2 Simulação Log-Normal

Tabela 34 – Viés dos estimadores de máxima verossimilhança de ($\hat{\beta_\mu}$) do modelo de múltiplos níveis de censura Log-Normal utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Coeficiente	Tamnho da Amostra	Cenário 1	Cenário 2	Cenário 3
$\beta_{\mu 1}$	50	-0,00563	0,001967	0,004323
	100	-0,003228	0,004074	-0,000513
	250	-0,005848	-0,000508	-0,001163
	500	2,5e-05	-8,4e-05	0,000401
$\beta_{\mu 2}$	50	-0,010442	0,003622	0,00082
	100	-0,007151	0,007387	-0,003094
	250	-0,010106	0,003364	-0,003837
	500	-0,004142	0,003909	-0,00204
$\beta_{\mu 3}$	50	-0,012819	0,003624	-0,000715
	100	-0,009829	0,008231	-0,004602
	250	-0,012706	0,004911	-0,005441
	500	-0,007016	0,005598	-0,003785

Tabela 35 – Viés dos estimadores de máxima verossimilhança de ($\hat{\beta_\sigma}$) do modelo de múltiplos níveis de censura Log-Normal utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Coeficiente	Tamanho da Amostra	Cenário 1	Cenário 2	Cenário 3
$\beta_{\sigma 1}$	50	-0,060426	0,020101	0,061975
	100	-0,022346	0,006372	0,025811
	250	-0,006967	0,001932	0,011502
	500	-0,00494	0,000907	0,0055
$\beta_{\sigma 2}$	50	-0,05206	0,01749	0,068189
	100	-0,013334	0,00277	0,030929
	250	0,001754	-0,001966	0,015958
	500	0,003951	-0,00437	0,00925
$\beta_{\sigma 3}$	50	-0,046308	0,016585	0,071833
	100	-0,006975	0,001184	0,035979
	250	0,007906	-0,003755	0,020176
	500	0,010205	-0,006735	0,013307

Tabela 36 – Raiz do erro quadrático médio dos estimadores de máxima verossimilhança de ($\hat{\beta_\mu}$) do modelo de múltiplos níveis de censura Log-Normal utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Coeficiente	Tamanho da Amostra	Cenário 1	Cenário 2	Cenário 3
$\beta_{\mu 1}$	50	0,176504	0,105765	0,113073
	100	0,124713	0,073196	0,07445
	250	0,076931	0,04447	0,042606
	500	0,053641	0,030762	0,028887
$\beta_{\mu 2}$	50	0,176273	0,102662	0,112022
	100	0,124474	0,070852	0,073656
	250	0,076645	0,042726	0,042031
	500	0,053282	0,029411	0,028371
$\beta_{\mu 3}$	50	0,176272	0,101056	0,111581
	100	0,12437	0,069225	0,073237
	250	0,076379	0,041256	0,04164
	500	0,053063	0,02837	0,028086

Tabela 37 – Raiz do erro quadrático médio dos estimadores de máxima verossimilhança de ($\hat{\beta}_\sigma$) do modelo de múltiplos níveis de censura Log-Normal utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Coeficiente	Tamanho da Amostra	Cenário 1	Cenário 2	Cenário 3
$\beta_{\sigma 1}$	50	0,117434	0,132315	0,139185
	100	0,079188	0,086954	0,089228
	250	0,049345	0,052496	0,052729
	500	0,034756	0,036574	0,0365
$\beta_{\sigma 2}$	50	0,115309	0,12958	0,138861
	100	0,078168	0,085271	0,088942
	250	0,048858	0,05142	0,052561
	500	0,034437	0,035781	0,036336
$\beta_{\sigma 3}$	50	0,113817	0,127876	0,138254
	100	0,077497	0,084115	0,088699
	250	0,048574	0,050698	0,052407
	500	0,034283	0,035251	0,036231

Tabela 38 – Probabilidade de cobertura dos estimadores de máxima verossimilhança de ($\hat{\beta}_\mu$) do modelo de múltiplos níveis de censura Log-Normal utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Coeficiente	Tamanho da Amostra	Cenário 1	Cenário 2	Cenário 3
$\beta_{\mu 1}$	50	0,91	0,935	0,956
	100	0,94	0,967	0,972
	250	0,937	0,974	0,97
	500	0,951	0,979	0,97
$\beta_{\mu 2}$	50	0,914	0,93	0,957
	100	0,941	0,959	0,969
	250	0,94	0,968	0,973
	500	0,946	0,97	0,978
$\beta_{\mu 3}$	50	0,913	0,931	0,959
	100	0,936	0,956	0,972
	250	0,94	0,966	0,974
	500	0,948	0,967	0,98

Tabela 39 – Probabilidade de cobertura dos estimadores de máxima verossimilhança de ($\hat{\beta}_\sigma$) do modelo de múltiplos níveis de censura Log-Normal utilizando dados simulados sob os três cenários sob diferentes tamanhos de amostrais (n).

Coeficiente	**Tamanho da Amostra**	**Cenário 1**	**Cenário 2**	**Cenário 3**
$\beta_{\sigma 1}$	50	0,89	0,929	0,902
	100	0,932	0,954	0,937
	250	0,952	0,945	0,936
	500	0,937	0,951	0,939
$\beta_{\sigma 2}$	50	0,895	0,924	0,898
	100	0,936	0,951	0,934
	250	0,958	0,939	0,938
	500	0,93	0,951	0,93
$\beta_{\sigma 3}$	50	0,894	0,915	0,897
	100	0,944	0,946	0,927
	250	0,956	0,932	0,931
	500	0,926	0,947	0,918


ICMC
USP
SÃO CARLOS