UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo de dispersão hiper-Poisson para variáveis discretas observáveis e não observáveis

Daiane de Souza Santos

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Daiane de Souza Santos

Modelo de dispersão hiper-Poisson para variáveis discretas observáveis e não observáveis

Tese 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 Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *EXEMPLAR DE DEFESA*

Área de Concentração: Estatística

Orientadora: Profa. Dr. Vicente Garibay Cancho
Coorientador: Prof. Dr. Josemar Rodrigues

USP – São Carlos
Outubro de 2019

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

de Souza Santos, Daiane
d278m Modelo de dispersão hiper-Poisson para variáveis discretas observáveis e não observáveis / Daiane de Souza Santos; orientador Vicente Garibay Cancho; coorientador Josemar Rodrigues. -- São Carlos, 2019.
116 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2019.

1. Distribuição hiper-Poisson. 2. Inferência estatística. 3. Modelos de sobrevivência. 4. Modelos de fragilidade. 5. Inferência bayesiana. I. Garibay Cancho, Vicente, orient. II. Rodrigues, Josemar, 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

Daiane de Souza Santos

Hyper-Poisson dispersion model for observable and unobservable discrete variables

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

Concentration Area: Statistics

Advisor: Profa. Dr. Vicente Garibay Cancho
Co-advisor: Prof. Dr. Josemar Rodrigues

USP – São Carlos
October 2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado da candidata Daiane de Souza Santos, realizada em 06/12/2019:

Prof. Dr. Vicente Garibay Cancho
USP

Profa. Dra. Elizabeth Mie Hashimoto
UTFPR

Prof. Dr. Edson Zangiacomi Martinez
USP

Profa. Dra. Andréia da Silva Meyer
UNESP

Profa. Dra. Cibele Maria Russo Novelli
USP

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Edson Zangiacomi Martinez e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Vicente Garibay Cancho

Este trabalho é dedicado com muito amor e carinho
ao meu companheiro Roberto e aos meus filhos
Leonardo e Elisa.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Vicente Garibay Cancho, por todo empenho, boas ideias e compreensão com que me conduziu ao longo deste trabalho.

Ao meu coorientador, Josemar Rodrigues, por toda sabedoria compartilhada.

Aos professores Dr. Mário de Castro e Dr. Arthur Lemonte, membros da banca examinadora do exame de qualificação, pelas sugestões apresentadas e pelas correções apontadas.

Ao corpo docente de estatística da UFSCar e do ICMC, pelas disciplinas ministradas.

Aos funcionários do ICMC e da UFSCar, especialmente à Maria Isabel de Araujo, pela dedicação e convívio.

Ao meu companheiro de vida, Roberto, por todo incentivo ao longo destes anos.

Aos meus colegas de doutorado, principalmente José Clelto e George, por toda companhia durante as disciplinas.

Finalmente, agradeço à CAPES pelo apoio financeiro durante o desenvolvimento deste trabalho.

RESUMO

SANTOS, D. S. **Modelo de dispersão hiper-Poisson para variáveis discretas observáveis e não observáveis**. 2019. 117 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A distribuição Poisson é amplamente utilizada para modelar dados de contagem, no entanto tem como desvantagem a suposição de que os dados precisam ter média e variância iguais, o que nem sempre é verdade, pois em muitas situações é comum o fenômeno de sobredispersão (variância maior do que a média) ou subdispersão (variância menor do que a média). Desta forma, trabalhamos com a distribuição hiper-Poisson, que permite analisar dados com sobredispersão ou subdispersão. O modelo hiper-Poisson é investigado aqui em dois cenários distintos, primeiramente modelando variáveis aleatórias observáveis em problemas de contagem, e em um segundo momento representando uma variável não observável (latente) utilizada em modelos de análise de sobrevivência. No primeiro cenário, realizamos uma abordagem clássica para a estimação dos parâmetros da distribuição hiper-Poisson e empregamos o usual teste da razão de verossimilhanças, juntamente com o teste gradiente para testar o parâmetro de dispersão do modelo. Por outro lado, na análise de sobrevivência, propomos um novo modelo com fração de cura induzido por fragilidade discreta com distribuição de probabilidade hiper-Poisson, uma vez que é importante a escolha de uma distribuição que leve em conta a dispersão dos fatores de risco. Para este novo modelo desenvolvemos procedimentos inferenciais sob as perspectivas clássica e bayesiana. Todos os modelos trabalhados foram analisados por meio de estudos de simulação e aplicados a conjuntos de dados reais.

Palavras-chave: Distribuição hiper-Poisson, Teste gradiente, Modelos de fragilidade, Modelos com fração de cura, Algoritmo EM, Inferência bayesiana.

ABSTRACT

SANTOS, D. S. **Hyper-Poisson dispersion model for observable and unobservable discrete variables**. 2019. 117 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Poisson distribution is widely used to model count data, however it has the disadvantage the assumption that the data must have equal mean and variance, which is not always true, since in many situations the phenomenon of overdispersion (variance greater than average) or underdispersion (variance lower than average) is common. Thus, we work with the hyper-Poisson distribution, which may accomodate data with overdispersion or underdispersion. The hyper-Poisson model is investigated here in two distinct scenarios, first modeling observable random variables in counting problems, and secondly representing an unobservable (latent) variable used in survival analysis models. In the first scenario, we take a classic approach for the estimation of the parameters of the hyper-Poisson distribution and we developed the usual likelihood ratio test, together with the gradient test to test the model dispersion parameter. In the survival analysis, we propose a new cure rate model induced by frailty discrete with hyper-Poisson probability distribution, since it is important to choose a distribution that takes into account the dispersion of risk factors. For this new model we developed inferential procedures from the classical and bayesian perspectives. All the models worked were analyzed through simulation studies and applied to real data sets.

Keywords: Hyper-Poisson distribution, Gradient test, Frailty models, Cure rate models, EM-algorithm, Bayesian inference.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Aplicações que envolvem dados de contagem têm sido encontradas há diversos anos em diferentes áreas e ainda hoje são bastante comuns, como por exemplo, o número de visitas em um sítio web, o número de chamadas em um *call center* e o número de itens com defeito em uma linha de produção.

Quando se fala em dados de contagem, certamente a distribuição mais popular para modelar este tipo de dados é o modelo Poisson, que tem a propriedade de que a média e a variância dos dados devem ser iguais (equidispersão). Porém, muitos conjuntos de dados reais apresentam sobredispersão ou subdispersão em relação à distribuição Poisson, que ocorre quando a variância excede a média ou quando a variância é menor do que a média, respectivamente.

Heterogeneidade e agregação são as possíveis causas para a sobredispersão, enquanto que o efeito de repulsão resulta em subdispersão. Diante disto, diversos modelos foram propostos ao longo dos anos a fim de contornar este problema, como as distribuições Binomial Negativa, Poisson Generalizada, Poisson Dupla, Conway-Maxwell-Poisson (COM-Poisson) (CONWAY; MAXWELL, 1962) e hiper-Poisson (BARDWELL; CROW, 1964), que modelam dados com subdispersão ou sobredispersão.

Entre as distribuições listadas, a COM-Poisson é uma das que mais vem sendo utilizada recentemente na literatura, por se tratar de uma generalização elegante e flexível do modelo Poisson. Contudo, não apresenta forma fechada para as expressões da média e da variância, o que acaba sendo visto como um aspecto negativo por alguns autores (FRANCIS *et al.*, 2012). Diante disto, conduzimos a primeira parte deste trabalho investigando a distribuição hiper-Poisson, que modela dados de contagem para sobredispersão e subdispersão e, apresenta a grande vantagem de permitir a ligação de regressores à média e ao parâmetro de dispersão.

Interessados em testar o parâmetro de dispersão no modelo hiper-Poisson, desenvolvemos o clássico teste da razão de verossimilhanças e, como novidade neste tipo de modelo, apresentamos ainda o teste gradiente, que foi proposto na literatura por Terrell (2002). O teste

gradiente é muito simples de ser calculado, pois não envolve o cálculo de matrizes e tampouco de suas inversas. Além disso, ele preserva algumas propriedades assintóticas dos testes de escore, Wald e razão de verossimilhanças, parecendo ser uma alternativa atrativa a estas estatísticas usuais.

Como segunda parte do nosso trabalho, percorremos a área de análise de dados de sobrevivência, um campo da estatística formado por um conjunto de técnicas cujo principal objetivo está em analisar conjunto de dados que envolvem o tempo até a ocorrência de um determinado evento de interesse, que pode ser exemplificado como a morte de um paciente com câncer, a falha de um componente eletrônico, o aparecimento de um tumor ou até mesmo a cura de um indivíduo após um certo período de tratamento.

Existe uma parte da literatura sobre análise de sobrevivência que diz respeito aos denominados modelos de fragilidade (WIENKE, 2010), que considera a existência de uma heterogeneidade não observada entre os indivíduos em estudo, sendo caracterizada por uma variável aleatória não observada (variável latente), denominada fragilidade.

Usualmente, a variável de fragilidade assume uma distribuição contínua, sendo a distribuição Gama a mais utilizada, por ser não negativa, flexível e algebricamente conveniente. Contudo, cada vez mais outras distribuições têm sido empregadas e tem se tornado apropriado em certas situações considerar a fragilidade com distribuições discretas. Tal fato acontece quando, por exemplo, a heterogeneidade no tempo de vida surge devido à presença de um número desconhecido de defeitos em uma unidade de teste ou o número desconhecido de causas que levam à exposição a determinado dano (CARONI; CROWDER; KIMBER, 2010). Fragilidade com distribuição discreta se faz necessário ainda quando precisamos permitir que a fragilidade assuma o valor zero, correspondendo assim a um modelo contendo uma proporção de unidades que nunca falham, que em um contexto médico indica os indivíduos imunues ou curados, entre os quais o evento de interesse não ocorreu mesmo após longo período de observação.

O evento de interesse em muitos estudos de análise de sobrevivência pode ser a morte de um paciente ou a recorrência de um tumor. No entanto, devido aos recentes avanços nos tratamentos médicos, uma grande parte dos indivíduos é esperada ser “curada”, ou seja, permanecer livre da doença após prolongados período de observação. Em vista disso, existe uma vasta literatura sobre modelos de fração de cura. Muitos desses modelos foram obtidos no cenário de riscos competitivos, mas podem ser obtidos também por meio dos modelos de risco proporcional com distribuições de fragilidade discreta.

Neste panorama em que as distribuições de fragilidade contínuas não permitem a possibilidade de um indivíduo apresentar risco zero para a ocorrência de um evento é que propomos um modelo de sobrevivência com fração de cura induzido por fragilidade discreta com distribuição de probabilidade hiper-Poisson. A escolha de um modelo que expresse a dispersão dos fatores de risco é fundamental para explicar a dispersão do tempo de vida dos pacientes. Ademais, pouco tem se falado sobre essa distribuição nessa parte da análise de sobrevivência. Para o modelo

proposto, procedimentos inferenciais sob uma perspectiva clássica e bayesiana são considerados.

1.1 Descrição dos objetivos

Em meio ao contexto apresentado anteriormente, o principal objetivo desta tese é trabalhar com o modelo hiper-Poisson em duas perspectivas distintas. Em um primeiro momento, como variáveis observáveis em problemas de contagem e em um segundo momento, como variáveis latentes (fragilidade) em um novo modelo de sobrevivência com fração de cura. Assim, podemos destacar os seguintes objetivos específicos que pretendemos alcançar no decorrer deste trabalho:

(i) analisar os procedimentos inferenciais de estimação e desenvolver os testes de hipóteses da razão de verossimilhanças e gradiente para o parâmetro de dispersão no modelo hiper-Poisson;

(ii) avaliar o desempenho dos procedimentos de estimação dos parâmetros e dos testes de hipóteses por meio de estudos de simulação;

(iii) propor um novo modelo de sobrevivência com fração de cura induzido por fragilidade discreta seguindo distribuição hiper-Poisson;

(iv) estudar as propriedades estruturais do novo modelo induzido por fragilidade hiper-Poisson e desenvolver os procedimentos inferenciais por meio das perspectivas clássica e bayesiana;

(v) validar as metodologias propostas por meio de aplicações dos modelos trabalhados a conjunto de dados reais.

1.2 Apresentação dos capítulos

Este primeiro capítulo traz a motivação e a contextualização deste trabalho, bem como a descrição de seus principais objetivos. O Capítulo 2 é iniciado com um referencial teórico sobre o problema de subdispersão e sobredispersão presentes nos dados de contagem. A distribuição hiper-Poisson é apresentada e um modelo bastante geral chamado COM-Poisson estendido é descrito. São exibidos ainda os testes de hipóteses empregados. Uma abordagem inferencial clássica utilizando métodos de máxima verossimilhança é considerada para a estimação dos parâmetros do modelo hiper-Poisson com e sem covariáveis e os testes de hipóteses para o parâmetro de dispersão são realizados e avaliados por meio de um estudo de simulação. Por fim, duas aplicações a conjuntos de dados reais da literatura são conduzidas. Resultou do Capítulo 2 o artigo, Santos, Cancho e Rodrigues (2019), publicado no periódico *Journal of Statistical Computation and Simulation*. No Capítulo 3 são tratados dois conceitos da análise de sobrevivência, modelos com fração de cura e modelos de fragilidade, e a partir disso, é exibido o novo modelo

de sobrevivência com fração de cura induzido por fragilidade hiper-Poisson proposto nesta tese. Procedimentos inferenciais sob uma perspectiva clássica são considerados para a estimação dos parâmetros, bem como o algoritmo EM e a técnica da verossimilhança perfilada. Além disso, são apresentados um estudo de simulação com o objetivo de avaliar o desempenho do método de estimação proposto e uma aplicação do modelo a um conjunto de dados reais. Procedimentos inferenciais sob a perspectiva bayesiana para o modelo de sobrevivência com fração de cura proposto são realizados no Capítulo 4, e são ainda discutidos critérios bayesianos de comparação de modelos e medidas de diagnóstico. Novamente a metodologia é averiguada por meio de conjuntos de dados simulados e uma aplicação do modelo em um conjunto de dados reais é executada. Descendeu do Capítulo 4 o artigo Souza *et al.* (2017), publicado na revista *Statistical Methods in Medical Research*. Por fim, no Capítulo 5 são discutidas as principais conclusões desta tese com base nos resultados obtidos, bem como algumas perspectivas futuras de pesquisa.

CAPÍTULO

2

O MODELO HIPER- POISSON

Neste capítulo trabalhamos com o modelo de dispersão hiper-Poisson (hP) com e sem covariáveis para variáveis discretas observáveis. Uma breve revisão de literatura sobre os principais modelos que abordam a sobredispersão e subdispersão é feita. Os testes de hipóteses razão de verossimilhança e gradiente são enunciados. O modelo hP é apresentado e então realizamos procedimentos inferenciais sob uma perspectiva clássica e os testes de hipóteses da razão de verossimilhanças e gradiente foram desenvolvidos para testar o parâmetro de dispersão.

2.1 Subdispersão e sobredispersão em dados de contagem

Os dados de contagem são geralmente definidos como números de eventos por intervalo. Para que os dados de contagem sejam registrados, os eventos precisam ser especificados previamente, e depois reconhecidos e contados durante o experimento. É esperado que os eventos ocorram várias vezes para que os dados reunidos sejam qualificados e identificados como contagem, o que implica então que a contagem precisa ser um número inteiro não negativo ou zero. Diferente do que pode ser pensado, os dados de contagem não estão restritos à biomedicina. De fato, numerosos avanços estatísticos deste assunto ocorreram dentro de outras áreas. As primeiras aplicações variavam entre o número de soldados na cavalaria prussiana mortos por coices de cavalo (PLAN, 2014). Em estudos médicos, os dados de contagem estão muito presentes, como pode ser visto nos testes clínicos em que as contagens são reportadas para cada indivíduo, como por exemplo, o número de episódios de incontinência urinária por semana ou o número de crises de epilepsia por mês.

Por outro lado, para analisar dados dessa natureza, não há dúvidas de que a distribuição mais utilizada para modelar dados de contagem é a Poisson. Porém, muitos conjuntos de dados reais não satisfazem o pressuposto de que a média deve ser igual à variância, subjacente à

distribuição de Poisson. Muitas vezes, os dados exibem sobredispersão, que ocorre quando a variância excede a média; ou subdispersão, que acontece quando a variância é menor do que a média. Portanto, uma vez que detectamos que um conjunto de dados particular possa exibir subdispersão ou sobredispersão, precisamos pensar sobre como estender nosso modelo básico, pois ao ignorar um destes fatos podemos chegar a resultados e interpretações equivocados.

A distribuição Binomial Negativa foi a primeira alternativa proposta à distribuição Poisson para tratar dados com sobredispersão. No entanto, esta distribuição não inclui a família Poisson como caso particular dentro do seu espaço paramétrico. Nessa situação, existe uma literatura vasta sobre problemas que envolvem sobredispersão, dos quais podemos mencionar Breslow (1984), Lawless (1987), Lindsey (1995) e os trabalhos propostos por Gschlößl e Czado (2008) e Lee e Durbán (2009). No que diz respeito à subdispersão, embora seja pouco mencionada na literatura, ela também é frequente na prática. Ridout e Besbeas (2004) trataram dados de subdispersão com um modelo de Poisson ponderado, usando como exemplos o número de focos de greve na indústria de carvão do Reino Unido durante períodos sucessivos entre 1948 e 1959 e o número de ovos por ninho para uma espécie de ave.

Assim, várias distribuições têm sido propostas com o intuito de modelar ambas, as estruturas de dispersão, das quais muitas são generalizações do modelo de Poisson, que são obtidas pela inclusão de um parâmetro adicional na expressão da função de probabilidade Poisson. Podemos mencionar aqui os trabalhos de Castillo e Pérez-Casany (2005), que propuseram uma distribuição de Poisson ponderada, Consul e Famoye (1992), Wang e Famoye (1997) e Famoye e Wang (2004), que estudaram a distribuição de Poisson Generalizada, Efron (1986) e Ridout e Besbeas (2004) analisaram a distribuição Poisson Dupla, que embora tenha sido formulada em um contexto de subdispersão, ambas formas de dispersão podem ser contempladas. Destacamos ainda o modelo Conway-Maxwell-Poisson (COM-Poisson), proposto por Conway e Maxwell (1962) e o modelo hiper-Poisson, conhecido também como distribuição Crow-Bardwell, apresentado por Bardwell e Crow (1964).

A distribuição COM-Poisson foi modelada em conexão com o problema de teoria de filas e estudada recentemente em diversos trabalhos dos quais podemos citar Shmueli *et al.* (2005), Kadane *et al.* (2006), Sellers e Shmueli (2010) e Sellers, Borle e Shmueli (2012). O modelo COM-Poisson generaliza a distribuição Poisson de forma elegante e flexível, sendo provavelmente o modelo mais utilizado nos últimos anos para lidar com a subdispersão e a sobredispersão. No entanto, neste modelo, a constante de normalização é aproximada pelo truncamento da série que o define e não há expressões fechadas para a média e para a variância. Embora haja aproximações para a média e para a variância que exibem uma alta precisão em uma vasta região do espaço paramétrico, alguns autores alertam sobre o problema do uso das aproximações desses momentos e para a constante normalizadora (FRANCIS *et al.*, 2012).

A distribuição hiper-Poisson, generaliza a distribuição de Poisson com apenas dois parâmetros e é adequada para ambos tipos de dispersão (subdispersão e sobredispersão). Um

modelo de regressão com variável resposta seguindo distribuição hiper-Poisson foi estudado recentemente por Sáez-Castillo e Conde-Sánchez (2013), que consideraram como vantagem a escolha desse modelo o fato da existência de expressões explícitas da média em termos dos parâmetros, facilitando assim a construção dos modelos de regressão, em que o efeito das covariáveis sobre a média pode ser avaliado diretamente, como no modelo Poisson.

Recentemente foi estudado em Rodrigues, Prado e Balakrishnan (2016) o modelo Conway-Maxwell-Poisson estendido, que tem como casos particulares as distribuições COM-Poisson, Poisson, Geométrica e a hiper-Poisson, permitindo assim a modelagem com dados com sobredispersão ou subdispersão.

2.2 A distribuição hiper-Poisson

A distribuição hiper-Poisson, denotada aqui por hP(θ,η), foi proposta por Bardwell e Crow (1964) e tem função massa de probabilidade dada por

$$P(Z=z;\eta,\theta)=p(z;\eta,\theta)=\frac{1}{{}_1F_1(1;\eta;\theta)}\frac{\theta^z}{(\eta)_z},\quad z=0,1,2,\ldots, \tag{2.1}$$

em que $\eta,\theta>0$ e $(a)_r=a(a+1)\ldots(a+r-1)=\Gamma(a+r)/\Gamma(a)$ para $a>0$, r um inteiro positivo e

$${}_1F_1(a;b;w)=\sum_{r=0}^{\infty}\frac{(a)_r}{(b)_r}\frac{w^r}{r!}$$

é a série hipergeométrica confluente, que pode ser vista com mais detalhes em Johnson, Kemp e Kotz (2005).

A função geradora de probabilidade é dada por

$$G_Z(s)=\sum_{z=0}^{\infty}p(z;\eta,\theta)s^z=\frac{{}_1F_1(1;\eta;\theta s)}{{}_1F_1(1;\eta;\theta)},\quad 0\leq s<1. \tag{2.2}$$

Logo, a distribuição hP pode ser considerada como um membro da família de distribuições de probabilidade hipergeométrica generalizada (SÁEZ-CASTILLO; CONDE-SÁNCHEZ, 2013). Além disso, como a função geradora de probabilidade está relacionada à função hipergeométrica confluente, também pode ser considerada como um caso particular da família de distribuições hipergeométrica confluente (JOHNSON; KEMP; KOTZ, 2005).

A distribuição hP pode ser vista ainda como um caso particular da distribuição Conway-Maxwell-Poisson estendida estudada dentro dos problemas da teoria de filas por Rodrigues, Prado e Balakrishnan (2016) e enunciada por meio do seguinte resultado:

Teorema 1. Dado um sistema de filas M/M/1 com taxa de serviço $\mu_z=w(z,\eta)\mu$, em que $w(z,\eta)$ é uma função positiva com $w(0,\eta)=1$, a probabilidade de ter m unidades no sistema, denominada como a distribuição COM- Poisson estendida, é dada por

$$P(Z=z;\eta,\theta)=\frac{\theta^z}{\prod_{i=0}^{z}w(i;\eta)K(\theta,\eta)}\Leftrightarrow w(1,\eta)=1, \tag{2.3}$$

com $z = 0, 1, 2, \ldots$, e

$$K(\theta, \eta) = \sum_{z=0}^{\infty} \frac{\theta^z}{\prod_{i=0}^{z} w(i; \eta)}.$$

Demonstração. O resultado pode ser visto em Rodrigues, Prado e Balakrishnan (2016). □

Observação 1. Alguns casos especiais da distribuição COM-Poisson estendida são os seguintes:

- Distribuição COM–Poisson com parâmetros θ e η: $w(z, \eta) = z^{\eta}, \quad z \geq 1$;
- Distribuição Poisson com parâmetro θ: $w(z, \eta) = z, \quad z \geq 1$;
- Distribuição Geométrica com parâmetro $1 - \theta$: $w(z, \eta) = 1, \quad z \geq 0$;
- Distribuição hiper-Poisson com parâmetros θ e η se considerarmos

$$w(i, \eta) = \begin{cases} \eta + i - 1, & \text{se} \quad i \geq 2 \\ 1, & \text{se} \quad i = 1 \\ 1, & \text{se} \quad i = 0. \end{cases}$$

A média e a variância de uma variável Z com distribuição COM-Poisson estendida com parâmetros η e θ são dadas respectivamente por

$$E(Z) = \theta \frac{d \log(K_w(\theta, \eta))}{d\theta}$$

e

$$Var(Z) = E(Z) + \theta^2 \frac{d^2 \log E_\theta[w^\star(Z, \eta)]}{d\theta^2},$$

em que $w^\star(z; \eta) = z! / \prod_{i=1}^{z} w(i; \eta)$. Para a distribuição hP, especificamente, a média e a variância se reduzem a

$$E(Z) = \mu = \theta - (\eta - 1) \frac{{}_1F_1(1; \eta, \theta) - 1}{{}_1F_1(1; \eta; \theta)} \tag{2.4}$$

e

$$Var(Z) = \theta + (\theta - (\eta - 1))\mu - \mu^2,$$

respectivamente.

Rodrigues, Prado e Balakrishnan (2016) mostra que existe uma conexão entre a função $w^\star(z; \eta)$ com a sobredispersão e a subdispersão por meio do seguinte teorema:

Teorema 2. A distribuição COM-Poisson estendida com parâmetros η e θ admite sobredispersão (subdispersão) se a função peso Poisson $z \Longrightarrow w^\star(z, \eta)$ em (2.3) não depender de θ e for log-convexa (log-côncava).

Demonstração. O resultado segue de Rodrigues, Prado e Balakrishnan (2016). □

Para o modelo hP(θ,η), a função de peso é $w^{\star}(z;\eta)=\dfrac{z!}{(\eta)_z}$ e

$$\frac{d^2\log(w^{\star}(z;\eta))}{dz^2}=\sum_{k\in\mathbf{N}}\frac{1}{(k+z+1)^2}-\sum_{k\in\mathbf{N}}\frac{1}{(z+\eta+k)^2}.$$

Notemos que a expressão anterior assume valores positivos quando $\eta>1$ e assume valores negativos quando $\eta<1$. Assim, pelo Teorema 2, o modelo hP admite sobredispersão quando $\eta>1$. Se $\eta<1$, exibe subdispersão em relação a distribuição Poisson, de modo que o parâmetro η é interpretado como parâmetro de dispersão. Se $\eta=1$, temos o modelo Poisson com média θ.

2.3 Testes de hipóteses

A literatura estatística clássica quando aborda o problema de testes de hipóteses em modelos paramétricos para grandes amostras, utiliza alguns métodos bastante conhecidos baseados na função de verossimilhança, como o teste da razão de verossimilhanças, teste de Wald e o teste do escore, todos propostos por volta de 1940. Um pouco mais recente, uma nova estatística de teste que compartilha as mesmas propriedades assintóticas de primeira ordem com as três estatísticas mencionadas anteriormente foi proposta por Terrell (2002). Esta estatística, referida na literatura como estatística gradiente e muito estudada em Lemonte (2016), apresenta a vantagem de ser notoriamente simples de ser obtida, pois não envolve a estimação da matriz de informação de Fisher e nem o cálculo de sua inversa.

Optamos neste trabalho em utilizar as estatísticas razão de verossimilhanças e gradiente pelo fato de ambas não envolverem o cálculo da matriz de informação de Fisher, uma vez que para os modelos estudados não conseguimos calcular esta matriz de maneira exata. Dessa forma, apresentamos na sequência a formulação da estatística razão de verossimilhanças e exibimos a construção da estatística gradiente. Além disso, trazemos ainda as estatísticas de Wald e escore que são utilizadas para se obter a estatística gradiente.

Seja $\boldsymbol{y}=(y_1,\ldots,y_n)^{\top}$ uma amostra de n observações independentes com função de densidade de probabilidade (ou função massa de probabilidade) $f(\boldsymbol{y};\boldsymbol{\vartheta})$, que depende de um vetor com dimensão p de parâmetros desconhecidos $\boldsymbol{\vartheta}=(\vartheta_1,\ldots,\vartheta_p)^{\top}$. Assumimos que $\boldsymbol{\vartheta}\in\boldsymbol{\Theta}$, em que $\boldsymbol{\Theta}\subset\mathbb{R}^p$ é um subconjunto aberto do espaço Euclidiano. Seja $L(\boldsymbol{\vartheta};\boldsymbol{y})$ a função de verossimilhança, expressa como

$$L(\boldsymbol{\vartheta};\boldsymbol{y})=\prod_{i=1}^{n}f(y_i;\boldsymbol{\vartheta}).$$

Entretanto, é mais conveniente numericamente trabalhar com o logaritmo da função de verossimilhança, a função denominada de log-verossimilhança, escrita como

$$\ell(\boldsymbol{\vartheta};\boldsymbol{y})=\sum_{i=1}^{n}\log f(y_i;\boldsymbol{\vartheta}).$$

A primeira derivada da função de log-verossimilhança é chamada de função escore e é representada como

$$\boldsymbol{U}(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta};\boldsymbol{y})}{\partial \boldsymbol{\vartheta}}.$$

A matriz de informação de Fisher é dada por $\boldsymbol{K}(\boldsymbol{\vartheta}) = \mathrm{E}\left[\boldsymbol{U}(\boldsymbol{\vartheta})\boldsymbol{U}(\boldsymbol{\vartheta})^\top\right]$, que sob certas condições de regularidade (ver por exemplo, Sen e Singer (1993)), pode ser mostrado que é equivalente a $\boldsymbol{K}(\boldsymbol{\vartheta}) = -E\left[\dfrac{\partial^2 \ell(\boldsymbol{\vartheta};\boldsymbol{y})}{\partial \boldsymbol{\vartheta}\partial \boldsymbol{\vartheta}^\top}\right]$.

Consideramos o problema de testar a hipótese nula simples

$$H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0,$$

contra a hipótese alternativa bilateral

$$H_{\mathrm{a}} : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0,$$

em que $\boldsymbol{\vartheta}_0$ é um vetor com dimensão p fixado. As estatísticas da razão de verossimilhanças (RV), escore (S) e Wald (W) para testar H_0 são expressas, respectivamente, como

$$\mathrm{RV} = 2\left\{\ell(\widehat{\boldsymbol{\vartheta}};\boldsymbol{y}) - \ell(\boldsymbol{\vartheta}_0;\boldsymbol{y})\right\},$$

$$\mathrm{S} = \boldsymbol{U}(\boldsymbol{\vartheta}_0)^\top \boldsymbol{K}^{-1}(\boldsymbol{\vartheta}_0)\boldsymbol{U}(\boldsymbol{\vartheta}_0)$$

e

$$\mathrm{W} = (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^\top \boldsymbol{K}(\widehat{\boldsymbol{\vartheta}})(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0),$$

em que $\widehat{\boldsymbol{\vartheta}}$ é o estimador de máxima verossimilhança de $\boldsymbol{\vartheta}$, o qual pode ser obtido como solução de $\boldsymbol{U}(\widehat{\boldsymbol{\vartheta}}) = \boldsymbol{0}$, sob certas condições.

É sabido que as estatísticas da razão de verossimilhanças, escore e Wald possuem distribuição assintótica qui-quadrado com p graus de liberdade (χ^2_p) sob a hipótese nula $H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, perante condições de regularidade (BICKEL; DOKSUM, 2015). Rejeitamos H_0 se o valor observado da estatística ultrapassa o quantil $100(1-\alpha)\%$ da distribuição χ^2_p, sendo α o nível nominal do teste.

A estatística proposta por Terrell (2002) é obtida por meio das estatísticas de escore e Wald modificado (WM), que é expressa da seguinte maneira:

$$\mathrm{WM} = (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^\top \boldsymbol{K}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0).$$

Assim, para construir a estatística gradiente precisamos inicialmente de uma matriz quadrada $\boldsymbol{A}(\boldsymbol{\vartheta})$, tal que $\boldsymbol{A}(\boldsymbol{\vartheta})^\top \boldsymbol{A}(\boldsymbol{\vartheta}) = \boldsymbol{K}(\boldsymbol{\vartheta})$. Dessa maneira, podemos reescrever os testes de escore e Wald modificado, respectivamente como

$$\begin{aligned}
\mathrm{S} &= \boldsymbol{U}(\boldsymbol{\vartheta}_0)^\top (\boldsymbol{A}(\boldsymbol{\vartheta}_0)^\top \boldsymbol{A}(\boldsymbol{\vartheta}_0))^{-1} \boldsymbol{U}(\boldsymbol{\vartheta}_0) \\
&= \boldsymbol{U}(\boldsymbol{\vartheta}_0)^\top \boldsymbol{A}(\boldsymbol{\vartheta}_0)^{-1} (\boldsymbol{A}(\boldsymbol{\vartheta}_0)^\top)^{-1} \boldsymbol{U}(\boldsymbol{\vartheta}_0) \\
&= [(\boldsymbol{A}(\boldsymbol{\vartheta}_0)^{-1})^\top \boldsymbol{U}(\boldsymbol{\vartheta}_0)]^\top (\boldsymbol{A}(\boldsymbol{\vartheta}_0)^{-1})^\top \boldsymbol{U}(\boldsymbol{\vartheta}_0)
\end{aligned}$$

e

$$\begin{aligned} \text{WM} &= (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^\top (\boldsymbol{A}(\boldsymbol{\vartheta}_0)^\top \boldsymbol{A}(\boldsymbol{\vartheta}_0))(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \\ &= [\boldsymbol{A}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)]^\top \boldsymbol{A}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0). \end{aligned}$$

De modo análogo ao feito em Lemonte (2016), vamos definir $\boldsymbol{P}_1 = (\boldsymbol{A}(\boldsymbol{\vartheta}_0)^{-1})^\top \boldsymbol{U}(\boldsymbol{\vartheta}_0)$ e $\boldsymbol{P}_2 = \boldsymbol{A}(\boldsymbol{\vartheta}_0)(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)$. Padronizando os vetores $\boldsymbol{P}_1$ e $\boldsymbol{P}_2$, cada um tem distribuição assintótica normal multivariada: $\boldsymbol{P}_1 \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$ e $\boldsymbol{P}_2 \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$, quando n é grande, em que $\boldsymbol{I}_p$ é a matriz identidade de ordem p. Além disso, $\boldsymbol{P}_1$ e $\boldsymbol{P}_2$ convergem um para o outro em probabilidade, ou seja, $\boldsymbol{P}_1 - \boldsymbol{P}_2 \xrightarrow{\mathbb{P}} \boldsymbol{0}_p$. Assim, o produto interno entre esses dois vetores padronizados resulta em

$$\begin{aligned} \boldsymbol{P}_1^\top \boldsymbol{P}_2 &= [(\boldsymbol{A}(\boldsymbol{\vartheta}_0)^{-1})^\top \boldsymbol{U}(\boldsymbol{\vartheta}_0)]^\top \boldsymbol{A}(\boldsymbol{\vartheta}_0)[(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)] \\ &= \boldsymbol{U}(\boldsymbol{\vartheta}_0)^\top (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0), \end{aligned}$$

que sob H_0, apresenta distribuição assintótica χ^2_p. Com isso, podemos definir a estatística gradiente.

Definição 1. (TERRELL, 2002) A estatística gradiente, G, para testar a hipótese nula $H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$ contra a hipótese alternativa bilateral $H_a : \boldsymbol{\vartheta} \neq \boldsymbol{\vartheta}_0$ é dada por

$$\text{G} = \boldsymbol{U}(\boldsymbol{\vartheta}_0)^\top (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0).$$

Segundo Terrell (2002), a estatística gradiente apresenta a peculiaridade de não ser evidentemente não negativa, embora deva ser não negativa assintoticamente. Uma vez que as estatísticas RV, escore e Wald são claramente não negativas, esta questão torna-se natural. O próximo teorema descreve esta peculiaridade.

Teorema 3. Se a função de log-verossimilhanças $\ell(\boldsymbol{\vartheta}; \boldsymbol{y})$ é unimodal e diferenciável para algum $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$, então

$$\text{G} = \boldsymbol{U}(\boldsymbol{\vartheta}_0)^\top (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \geq 0.$$

Demonstração. A demonstração pode ser vista em Terrell (2002). □

Discutiremos agora, de modo similar ao feito em Lemonte (2016), o caso em que a hipótese nula H_0 é composta, que corresponde de fato ao nosso interesse neste trabalho. Seja o vetor de parâmetros $\boldsymbol{\vartheta}$ particionado como $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1^\top, \boldsymbol{\vartheta}_2^\top)^\top$, em que $\boldsymbol{\vartheta}_1$ e $\boldsymbol{\vartheta}_2$ são vetores de parâmetros de dimensões q e $p-q$, respectivamente. Suponhamos que o interesse agora esteja em testar a hipótese nula composta

$$H_0 : \boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_{10}$$

contra a hipótese alternativa bilateral

$$H_a : \boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_{10},$$

em que $\boldsymbol{\vartheta}_{10}$ é um vetor de constantes conhecidas.

Seja $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\vartheta}}_1^\top, \widehat{\boldsymbol{\vartheta}}_2^\top)^\top$ e $\widetilde{\boldsymbol{\vartheta}} = (\boldsymbol{\vartheta}_{10}^\top, \widetilde{\boldsymbol{\vartheta}}_2^\top)^\top$ os estimadores de máxima verossimilhança de $\boldsymbol{\vartheta}$ irrestritos e restritos (obtidos sob a hipótese nula). As estatísticas razão de verossimilhanças e gradiente para testar a hipótese nula composta $H_0 : \boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_{10}$ são definidas, respectivamente, como

$$\mathrm{RV} = 2\left\{\ell(\widehat{\boldsymbol{\vartheta}}; \boldsymbol{y}) - \ell(\widetilde{\boldsymbol{\vartheta}}; \boldsymbol{y})\right\}$$

e

$$\mathrm{G} = \boldsymbol{U}(\widetilde{\boldsymbol{\vartheta}})^\top (\widehat{\boldsymbol{\vartheta}} - \widetilde{\boldsymbol{\vartheta}}). \tag{2.5}$$

Estas estatísticas têm distribuição assintótica χ^2_q sob a hipótese nula (perante condições de regularidade), sendo q o número de restrições impostas por H_0.

Na sequência apresentamos uma expressão muito conveniente matematicamente e computacionalmente. A partição de $\boldsymbol{\vartheta}$ induz a partição correspondente: $\boldsymbol{U}(\boldsymbol{\vartheta}) = (\boldsymbol{U}_1(\boldsymbol{\vartheta})^\top, \boldsymbol{U}_2(\boldsymbol{\vartheta})^\top)^\top$, em que

$$\boldsymbol{U}_1(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta}; \boldsymbol{z})}{\partial \boldsymbol{\vartheta}_1} \quad \text{e} \quad \boldsymbol{U}_2(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta}; \boldsymbol{z})}{\partial \boldsymbol{\vartheta}_2}.$$

Uma vez que $\boldsymbol{U}_2(\widetilde{\boldsymbol{\vartheta}}) = \mathbf{0}_{p-q}$, a estatística gradiente em 2.5 pode ser expressa como

$$\mathrm{G} = \boldsymbol{U}_1(\widetilde{\boldsymbol{\vartheta}})^\top (\widehat{\boldsymbol{\vartheta}}_1 - \boldsymbol{\vartheta}_{10}).$$

Vários trabalhos têm explorado a estatística gradiente, dos quais podemos mencionar aqui Lemonte e Ferrari (2011), que compararam por meio de estudos de simulação o teste da razão de verossimilhanças e o teste gradiente em modelos de regressão Birbaun-Saunders em amostras censuradas. Lemonte e Ferrari (2012b) analisaram o poder local apresentando inicialmente a expansão assintótica da função de poder local do teste gradiente sob uma sequência de hipóteses alternativas. Realizaram ainda um estudo de poder local do teste gradiente comparando-o com o poder local dos testes da razão de verossimilhança, Wald e escore na família exponencial uniparamétrica, em que constataram que nenhuma estatística é uniformemente superior às outras. Outros trabalhos recentes envolvendo a estatística gradiente que podemos mencionar aqui são Lemonte e Ferrari (2012a), Vargas *et al.* (2013) e Lemonte (2013).

2.4 Inferência no modelo hiper-Poisson

Seja uma amostra aleatória $Z_1, \ldots, Z_n$ de tamanho n de uma população com distribuição hP. A função de verossimilhança é dada por

$$L(\boldsymbol{\vartheta}|\boldsymbol{z}) = \prod_{i=1}^{n} \frac{\theta^{z_i}}{{}_1F_1(1; \eta; \theta)} \frac{\Gamma(\eta)}{\Gamma(\eta + z_i)},$$

em que $\boldsymbol{\vartheta} = (\eta, \theta)^\top$ é o vetor de parâmetros de interesse, $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ e ${}_1F_1(1; \eta; \theta)$ é a série hipergeométrica confluente. Consequentemente, a função de log-verossimilhança é expressa

como

$$\ell(\boldsymbol{\vartheta}|\boldsymbol{z}) = -\sum_{i=1}^{n} \log(\Gamma(\eta + z_i)) + n\bar{z}\log(\theta) + n[\log(\Gamma(\eta)) - \log({}_1F_1(1;\eta;\theta))], \tag{2.6}$$

com $\bar{z} = \sum_{i=1}^{n} z_i/n$.

Com a primeira derivada da função de log-verossimilhança $\ell(\eta,\theta|\boldsymbol{z})$ em relação a cada parâmetro do modelo, definimos o vetor escore $\boldsymbol{U(\vartheta)}$, cujos elementos são dados por

$$U_\eta(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta}|\boldsymbol{z})}{\partial \eta} = n\left(\frac{-1}{{}_1F_1(1;\eta;\theta)}H^{(1)} + \Psi(\eta)\right) - \sum_{i=1}^{n}\Psi(\eta + z_i) \tag{2.7}$$

e

$$U_\theta(\boldsymbol{\vartheta}) = \frac{\partial \ell(\boldsymbol{\vartheta}|\boldsymbol{z})}{\partial \theta} = n\left(\frac{\bar{z}}{\theta} - \frac{{}_1F_1(2;\eta+1;\theta)}{{}_1F_1(1;\eta;\theta)}\frac{1}{\eta}\right), \tag{2.8}$$

em que

$$H^{(1)} = \frac{\partial\ {}_1F_1(1;\eta;\theta)}{\partial \eta} = -\sum_{m=0}^{\infty}\frac{(1)_m}{(\eta)_m}\Psi(\eta+m)\frac{\theta^m}{m!} + \Psi(\eta){}_1F_1(1;\eta;\theta),$$

como pode ser visto em Ancarani e Gasaneo (2008), $\Psi(x)$ é a função digama dada por

$$\Psi(x) = \frac{d}{dx}\log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

e utilizamos ainda o fato de que

$$\frac{\partial\ {}_1F_1(a;b;w)}{\partial w} = \frac{a}{b}{}_1F_1(a+1;b+1;w).$$

Os estimadores de máxima verossimilhança (EMVs) de $\boldsymbol{\vartheta} = (\eta,\theta)^\top$, denotados como $\widehat{\eta}$ e $\widehat{\theta}$, podem ser encontrados pela maximização da função de log-verossimilhança (2.6) resolvendo as equações dadas por $U_\eta(\boldsymbol{\vartheta}) = 0$ e $U_\theta(\boldsymbol{\vartheta}) = 0$. A solução destas equações não possui forma fechada, tornando-se necessário o uso de métodos iterativos para a obtenção das estimativas. O método BFGS (PRESS *et al.*, 2007) foi utilizado para maximizar a função de log-verossimilhança.

Nosso objetivo ao testar hipóteses está em investigar a equidispersão. Dessa forma, as hipóteses de interesse são

$$H_0 : \eta = 1 \quad \text{contra} \quad H_a : \eta \neq 1. \tag{2.9}$$

Seja $\widehat{\boldsymbol{\vartheta}} = \arg\max_{(\eta,\theta)}\ell(\boldsymbol{\vartheta}|\boldsymbol{z})$ o EMV obtido a partir do ajuste do modelo completo e $\widetilde{\boldsymbol{\vartheta}} = \arg\max_{(\eta=1,\theta)}\ell(\boldsymbol{\theta}|\boldsymbol{z})$ o correspondente EMV sob a hipótese nula H_0. Como dito anteriormente, não há forma fechada para os EMVs do modelo completo, sendo assim, denotaremos $\widehat{\boldsymbol{\vartheta}} = (\hat{\eta},\hat{\theta})$. Porém, sob H_0, com $\tilde{\theta}$ representando o EMV de θ sob a hipótese nula, podemos reescrever a expressão (2.8) como

$$U_\theta(\widetilde{\boldsymbol{\vartheta}}) = n\left(-\frac{{}_1F_1(2;2;\tilde{\theta})}{{}_1F_1(1;1;\tilde{\theta})} + \frac{\bar{z}}{\tilde{\theta}}\right) = n\left(-1 + \frac{\bar{z}}{\tilde{\theta}}\right),$$

implicando que $\tilde{\theta} = \bar{z}$. Logo, $\widetilde{\boldsymbol{\vartheta}} = (1, \bar{z})$. Com isso, podemos determinar as duas estatísticas de testes escolhidas para testar as hipóteses dadas em (2.9) no modelo hP.

Dessa maneira, as estatísticas da razão de verossimilhanças (RV) e gradiente (G) são dadas, respectivamente, por

$$\mathrm{RV} = 2\left\{\sum_{i=1}^{n}\log\left(\frac{\Gamma(1+z_i)}{\Gamma(\hat{\eta}+z_i)}\right) + n\bar{z}\left(\log\frac{\hat{\theta}}{\bar{z}}\right) + n\left[\log\Gamma(\hat{\eta}) + \log\left(\frac{{}_1F_1(1;1;\bar{z})}{{}_1F_1(1;\hat{\eta};\hat{\theta})}\right)\right]\right\}$$

e

$$\mathrm{G} = U_\eta(\widetilde{\boldsymbol{\vartheta}})(\hat{\eta} - 1),$$

em que

$$U_\eta(\widetilde{\boldsymbol{\vartheta}}) = n\left(\frac{1}{{}_1F_1(1;1,\bar{z})}\sum_{m=0}^{\infty}\Psi(1+m)\frac{\bar{z}^m}{m!}\right) - \sum_{i=1}^{n}\Psi(1+z_i).$$

2.5 Inferência no modelo de regressão hiper-Poisson

Em muitas aplicações práticas é comum assumir que os parâmetros do modelo dependem de covariáveis. Dessa forma, consideramos duas situações: (i) o modelo de regressão hP com covariáveis no parâmetro θ e (ii) o modelo de regressão hp com covariáveis na média μ.

Trabalhamos primeiramente com o modelo em que z_i, o valor da variável reposta do i-ésimo indivíduo da amostra segue uma distribuição hP e associamos o parâmetro θ com as covariáveis $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ por meio da expressão $\theta_i = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denota o vetor de coeficientes de regressão.

O vetor de parâmetros de interesse agora é dado por $\boldsymbol{\vartheta} = (\eta, \beta_0, \ldots, \beta_p)^\top$. Como já mencionado, com a primeira derivada da função de log-verossimilhança em relação a cada parâmetro do modelo obtemos o vetor escore $\boldsymbol{U}(\boldsymbol{\vartheta})$. Para tal, reescrevemos aqui a função de log-verossimilhança dada em (2.6) em função das covariáveis como

$$\ell_\theta(\boldsymbol{\vartheta};\boldsymbol{z}) = -\sum_{i=1}^{n}\Big[\log(\Gamma(\eta+z_i)) + z_i\log(\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})) - \log({}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})))\Big] + n\log(\Gamma(\eta)). \tag{2.10}$$

Assim, os elementos do vetor escore $\boldsymbol{U}(\boldsymbol{\vartheta})$ agora são dados por

$$U_\eta(\boldsymbol{\vartheta}) = \frac{\partial\ell_\theta(\boldsymbol{\vartheta};\boldsymbol{z})}{\partial\eta} = -\sum_{i=1}^{n}\Psi(\eta+z_i) + n\Psi(\eta) - \sum_{i=1}^{n}\frac{1}{{}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))}H^{(1)} \tag{2.11}$$

e

$$U_{\beta_r}(\boldsymbol{\vartheta}) = \frac{\partial\ell_\theta(\boldsymbol{\vartheta};\boldsymbol{z})}{\partial\beta_r} = \left[\sum_{i=1}^{n}\frac{z_i}{\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})} - \frac{{}_1F_1(2;\eta+1;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))}{{}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))}\frac{1}{\eta}\right]\frac{\partial(\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))}{\partial\beta_r}, \tag{2.12}$$

em que

$$H^{(1)} = -\sum_{m=0}^{\infty}\frac{(1)_m}{(\eta)_m}\Psi(\eta+m)\frac{(\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))^m}{m!} + \Psi(\eta)\,{}_1F_1(1;\eta;\theta_i)$$

e $r = 0, 1, \ldots, p$. Portanto, $\boldsymbol{U}(\boldsymbol{\vartheta})$ é um vetor $k \times 1$ em que k corresponde ao número de parâmetros do modelo ($k = p + 2$) e é dado por

$$\boldsymbol{U}(\boldsymbol{\vartheta}) = (U_\eta(\boldsymbol{\vartheta}), U_{\beta_0}(\boldsymbol{\vartheta}), \ldots, U_{\beta_p}(\boldsymbol{\vartheta}))^\top.$$

Assim como no caso anterior, métodos iterativos são necessários para se obter os EMVs.

Novamente, nosso interesse nos testes de hipóteses está em averiguar a equidispersão. Para tal utilizamos mais uma vez as estatísticas de razão de verossimilhanças e gradiente para testar as hipóteses dadas em (2.9).

Seja $\widehat{\boldsymbol{\vartheta}} = \arg\max_{(\eta,\boldsymbol{\beta})} \ell_\theta(\boldsymbol{\vartheta}; \boldsymbol{z})$ o estimador de máxima verossimilhança obtido a partir do ajuste do modelo completo e $\widetilde{\boldsymbol{\vartheta}} = \arg\max_{(\eta=1,\boldsymbol{\beta})} \ell_\theta(\boldsymbol{\vartheta}; \boldsymbol{z})$ o correspondente EMV sob a hipótese nula H_0. As estatísticas razão de verossimilhanças e gradiente são dadas, respectivamente, por

$$\text{RV} = 2\left\{\ell_\theta(\widehat{\boldsymbol{\vartheta}}|\boldsymbol{z}) - \ell_\theta(\widetilde{\boldsymbol{\vartheta}}|\boldsymbol{z})\right\}. \tag{2.13}$$

e

$$\text{G} = U_\eta(\widetilde{\boldsymbol{\vartheta}})(\hat{\eta} - 1), \tag{2.14}$$

em que $\hat{\eta}$ é o EMV do parâmetro η obtido do modelo completo.

Como feito recentemente em Sáez-Castillo e Conde-Sánchez (2013), consideramos agora o modelo em que z_i segue uma distribuição hP com média dada por

$$\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Desse modo, lidaremos agora com o modelo hP com parâmetros η e μ, em que podemos analisar o efeito da covariável sobre a média.

A estimação dos coeficientes de regressão $\boldsymbol{\beta}$ e do parâmetro η é obtida por meio da maximização da função de log-verossimilhança dada pela expressão (2.6), que depende de η e θ. Porém, estamos agora modelando com η e μ, então precisamos substituir o parâmetro θ pela sua expressão em termos da média μ, que pode ser deduzida de (2.4), em cada etapa do processo de otimização. O problema é que não existe uma expressão direta e fechada da expressão θ em termos de μ e η, mas podemos superar essa dificuldade resolvendo a equação resultante por meio de métodos numéricos em cada avaliação da função de log-verossimilhança no processo de otimização (SÁEZ-CASTILLO; CONDE-SÁNCHEZ, 2013).

Em suma, devemos aplicar o método de máxima verossimilhança de maneira usual, maximizando a função de log-verossimilhança considerando os coeficientes de regressão na expressão de μ e calculando o valor de θ em cada avaliação da função de verossimilhança como solução de (2.4). Neste contexto, utilizamos as funções `optim` da linguagem R (R Core Team, 2017) para maximizar a função de log-verossimilhança e `uniroot` para resolver a equação (2.4) numericamente.

Ainda como notado em Sáez-Castillo e Conde-Sánchez (2013), é importante destacar que para resolver (2.4) em cada avaliação da função de log-verossimilhança no processo de otimização aumenta-se fortemente o esforço computacional para obter um ajuste. Diante disso, a fim de minimizar esse tempo computacional, é aconselhável fornecer limites para o parâmetro θ bem próximos do valor real, reduzindo assim o tamanho do intervalo da função `uniroot` para encontrar a solução. Os limites para θ fornecidos por Sáez-Castillo e Conde-Sánchez (2013) são dados por

$$\theta \geq \min\{\mu, \max(\mu + (\eta - 1), \eta\mu)\}$$

e

$$\theta \leq \max\{\mu, \min(\mu + (\eta - 1), \eta\mu)\}.$$

Mais uma vez o vetor de parâmetros de interesse é dado por $\boldsymbol{\vartheta} = (\eta, \beta_0, \ldots, \beta_p)$. Os elementos do vetor escore $\boldsymbol{U}(\boldsymbol{\vartheta})$ não possuem forma explícita, mas iremos representá-los da seguinte maneira:

$$U_\eta(\boldsymbol{\vartheta}) = \frac{\partial \ell_\mu(\boldsymbol{\vartheta}|\boldsymbol{z})}{\partial \eta} \quad \text{e} \quad U_{\beta_r}(\boldsymbol{\vartheta}) = \frac{\partial \ell_\mu(\boldsymbol{\vartheta}|\boldsymbol{z})}{\partial \beta_r},$$

em que $\ell_\mu(\boldsymbol{\vartheta}|\boldsymbol{z})$ representa a função de log-verossimilhança com parâmetros μ e η que não conseguimos exibir.

Averiguamos mais uma vez com os testes de hipóteses a equidispersão. As estatísticas de razão de verossimilhanças e gradiente foram obtidas de maneira análoga ao caso anterior dado pelas expressões (2.13) e (2.14), porém com a função de log-verossimilhança em termos dos parâmetros μ e η.

2.6 Estudo de simulação

Nesta seção apresentamos um estudo de simulação para o modelo hP com e sem covariáveis, que tem como propósitos estudar o desempenho do método de máxima verossimilhança no processo de estimação dos parâmetros e analisar o comportamento dos testes razão de verossimilhanças e gradiente para testar a equidispersão. Os principais interesses nos testes de hipóteses estão em investigar a distribuição assintótica sob a hipótese nula e o poder para detectar a hipótese alternativa dessas duas estatísticas de teste.

2.6.1 Simulações sem covariáveis

A fim de examinar algumas propriedades dos estimadores de máxima verossimilhança, foram geradas 1.000 amostras de tamanhos $n = 100, 200, 400$ e 600 do modelo de dispersão hP com $\theta = 4$ e três escolhas de parâmetros para η, (i) $\eta = 0,5$, (ii) $\eta = 1$ e (iii) $\eta = 2$, representando assim em (i) dados com subdispersão, em (ii) dados com equidispersão e em (iii) dados com sobredispersão.

A implementação computacional foi desenvolvida em linguagem R R Core Team (2017). As estimativas de máxima verossimilhança foram obtidas pela maximização direta da função de log-verossimilhança utilizando a função `optim` em R.

Para avaliar a eficiência do estimador de cada parâmetro, estimativas de Monte Carlo da raiz do erro quadrático médio (REQM) e do desvio padrão do estimador foram calculadas por meio das seguintes equações:

$$\text{REQM}(\widehat{\vartheta}_k) = \left(\frac{1}{Q} \sum_{q=1}^{Q} (\widehat{\vartheta}_{kq} - \vartheta_k)^2 \right)^{1/2} \quad \text{e} \quad \text{DP}(\widehat{\vartheta}_k) = \left(\frac{1}{Q-1} \sum_{q=1}^{Q} (\widehat{\vartheta}_{kq} - \bar{\vartheta}_k)^2 \right)^{1/2},$$

em que ϑ_k é o k-ésimo componente do vetor de parâmetros $\backslash\vartheta$, $\widehat{\vartheta}_k$ o estimador de máxima verossimilhança de ϑ_k, Q o número de amostras geradas ($Q = 1.000$) e $\bar{\vartheta}_k = \frac{1}{Q}\sum_{q=1}^{Q} \widehat{\vartheta}_{kq}$.

Calculamos ainda as probabilidades de cobertura (PC) dos intervalos de confiança assintóticos para os parâmetros para o nível de confiança nominal fixado em 95%. Tais estimativas foram obtidas por meio do cálculo da proporção (baseado na simulação de 1.000 amostras) de intervalos que continham o verdadeiro valor dos parâmetros fixados na geração dos dados. Os resultados dessa simulação estão reportados na Tabela 1. Os histogramas e os gráficos QQ-normal dos parâmetros estimados são apresentados no Apêndice A.

Pela Tabela 1 podemos observar, conforme esperado, que as estimativas dos dois parâmetros ficam cada vez mais próximas dos seus verdadeiros valores conforme o tamanho da amostra aumenta. Ainda como previsto, o desvio padrão e o REQM das estimativas decrescem quando o número de observações na amostra aumenta e a razão entre essas duas medidas é um em quase todas as configurações. Além disso, no caso de subdispersão, o desvio padrão e o REQM apresentam os menores valores. No que diz respeito às estimativas da PC, estas estão próximas do nível nominal de 95% para tamanhos de amostras superiores a 400 para os dois parâmetros em todos os cenários.

A fim de avaliar a performance dos testes de hipóteses razão de verossimilhanças (RV) e gradiente (G) aplicados ao modelo hP, construímos 10.000 amostras de tamanhos $n = 200, 400$ e 600 sob a hipótese nula dada em (2.9) e analisamos primeiramente as taxas de rejeição de H_0. Adotamos como níveis de significância nominais $\alpha = 1\%, 5\%$ e 10% e assumimos três diferentes valores para o parâmetro θ, $\theta = 1, 2$ e 4. Os resultados estão na Tabela 2, em que observamos que os dois testes tiveram taxas de rejeição bem próximas aos níveis nominais adotados em todos os cenários e para os diferentes tamanhos de amostras escolhidos.

A fim de examinar ainda mais a distribuição sob a hipótese nula dos testes RV e gradiente, apresentamos as distribuições simuladas dos dois testes com $\theta = 4$ e com os 4 diferentes tamanhos de amostras trabalhados ($n = 100, 200, 400$ e 600), que podem ser vistas por meio dos gráficos quantil-quantil na Figura 1. Os gráficos mostram que a distribuição qui-quadrado com um grau de liberdade fornece uma aproximação razoável para a distribuição sob a hipótese nula

Tabela 1 – Médias das EMVs, desvio padrão (DP), raiz do erro quadrático médio (REQM) e a probabilidade de cobertura (PC) dos parâmetros do modelo hP sem covariáveis.

		$\eta = 0,5$		$\eta = 1$		$\eta = 2$	
n		$\hat{\eta}$	$\hat{\theta}$	$\hat{\eta}$	$\hat{\theta}$	$\hat{\eta}$	$\hat{\theta}$
100	Média	0,579	4,073	1,007	3,997	2,028	4,001
	DP	0,377	0,440	0,511	0,542	0,796	0,740
	REQM	0,385	0,446	0,511	0,542	0,796	0,739
	PC	1,000	0,992	0,939	0,946	0,909	0,923
200	Média	0,530	4,029	0,996	3,993	2,025	4,016
	DP	0,293	0,329	0,384	0,403	0,590	0,548
	REQM	0,294	0,330	0,384	0,403	0,590	0,548
	PC	0,991	0,977	0,923	0,933	0,929	0,931
400	Média	0,500	3,999	1,004	4,004	2,018	4,013
	DP	0,217	0,244	0,266	0,282	0,405	0,379
	REQM	0,217	0,244	0,266	0,282	0,405	0,379
	PC	0,949	0,946	0,944	0,953	0,944	0,956
600	Média	0,496	3,996	1,004	4,002	2,009	4,005
	DP	0,177	0,195	0,214	0,225	0,322	0,302
	REQM	0,177	0,195	0,214	0,225	0,322	0,302
	PC	0,948	0,947	0,945	0,954	0,950	0,948

Fonte: Elaborada pelo autor.

dos testes RV e gradiente, sendo essa aproximação melhorada com o aumento do tamanho da amostra.

Para finalizar, realizamos um estudo de simulação a fim de analisar o poder dos dois testes ao detectar a hipótese alternativa dada em (2.9). Para tais simulações, geramos 10.000 amostras com tamanhos $n = 200$ e $n = 400$ sob a hipótese alternativa $H_a : \eta = \delta$, para diferentes valores de δ, e então calculamos as taxas de rejeição. Adotamos como nível de significância nominal $\alpha = 5\%$. O poder dos testes RV e gradiente estão exibidos na Figura 2, da qual podemos ver claramente, conforme esperado, que o poder de ambos os testes é maior com a amostra de tamanho 400. Podemos notar ainda que, quando δ é menor que um, o teste RV apresenta um poder ligeiramente maior em relação ao teste gradiente e quando $\delta > 1$, esta situação se inverte e o teste gradiente apresenta um poder ligeiramente superior.

2.6.2 Simulações com covariáveis

Realizamos aqui um estudo de simulação análogo ao realizado na seção anterior, porém consideramos agora o modelo hP com covariáveis no parâmetro θ e com covariáveis na média μ como descrito na Seção 2.5.

Iniciamos mais uma vez analisando as propriedades frequentistas dos EMVs. Para cada i,

Tabela 2 – Taxas de rejeição de H_0 dos testes razão de verossimilhanças (RV) e gradiente (G) no modelo hP.

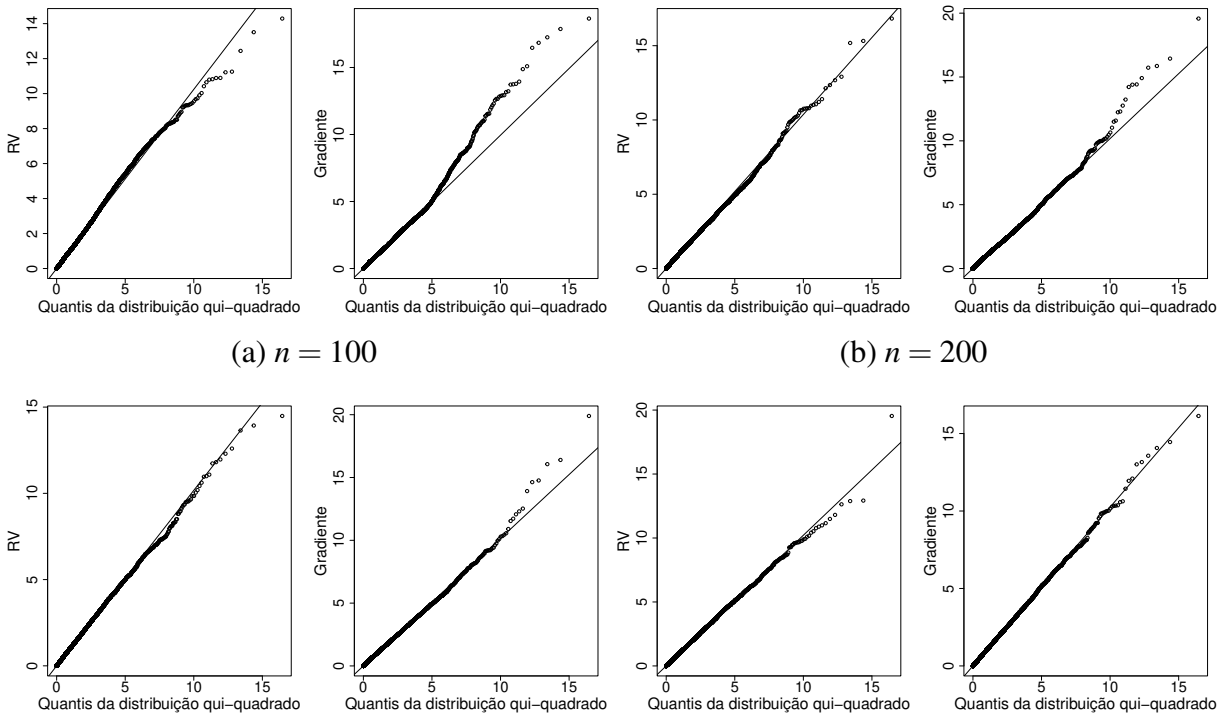
	Nível de Significância α					
	0,1		0,05		0,01	
n	RV	G	RV	G	RV	G
			$\theta = 1$			
100	0,107	0,102	0,052	0,063	0,010	0,029
200	0,101	0,102	0,050	0,054	0,009	0,016
400	0,100	0,100	0,049	0,053	0,010	0,015
600	0,102	0,101	0,051	0,051	0,010	0,015
n	RV	G	RV	G	RV	G
			$\theta = 2$			
100	0,104	0,093	0,052	0,053	0,011	0,019
200	0,100	0,097	0,049	0,048	0,010	0,013
400	0,096	0,096	0,049	0,048	0,012	0,013
600	0,098	0,097	0,050	0,051	0,011	0,013
n	RV	G	RV	G	RV	G
			$\theta = 4$			
100	0,108	0,102	0,060	0,048	0,013	0,015
200	0,107	0,102	0,054	0,052	0,011	0,011
400	0,099	0,097	0,050	0,049	0,011	0,012
600	0,098	0,961	0,049	0,048	0,010	0,010

Fonte: Elaborada pelo autor.

$i = 1, \ldots, n$, geramos variáveis aleatórias do modelo hP com três configurações para o parâmetro η, $\eta = 0,5$, $\eta = 1$ e $\eta = 2$ e $\theta_i = \exp(\beta_0 + \beta_1 x_i)$, em que $\beta_0 = 0,1$ e $\beta_1 = 1$ e as covariáveis x_i foram geradas de uma distribuição Bernoulli com parâmetro $0,5$. A simulação está baseada em 1.000 amostras de tamanhos $n = 100, 200, 400$ e 600. Os resultados dessa simulação estão organizados na Tabela 3. Os histogramas e os gráficos QQ-normal dos parâmetros estimados são apresentados no Apêndice B.

Pela Tabela 3 podemos ver que as estimativas dos parâmetros η e β_1 são muito próximas dos seus verdadeiros valores nos três cenários e para todos os tamanhos de amostra considerados. Já a estimativa do parâmetro β_0 vai se aproximando do seu verdadeiro valor à medida que o tamanho da amostra aumenta. Mais uma vez como esperado, o desvio padrão e o REQM das estimativas decrescem à medida que o tamanho da amostra aumenta e assim como observado no caso sem covariáveis, essas medidas apresentam os menores valores na configuração em que se tem subdispersão $(\eta = 0,5)$. No que diz respeito às PCs, estas estão bem próximas do valor nominal de 95% para os parâmetros β_0 e β_1 em todos os casos, contudo para o parâmetro η, estas estimativas são um pouco inferiores ao nível nominal, mas parece ir se aproximando à medida que o tamanho da amostra cresce.

Conduzimos mais uma vez um estudo com o objetivo de averiguar o comportamento

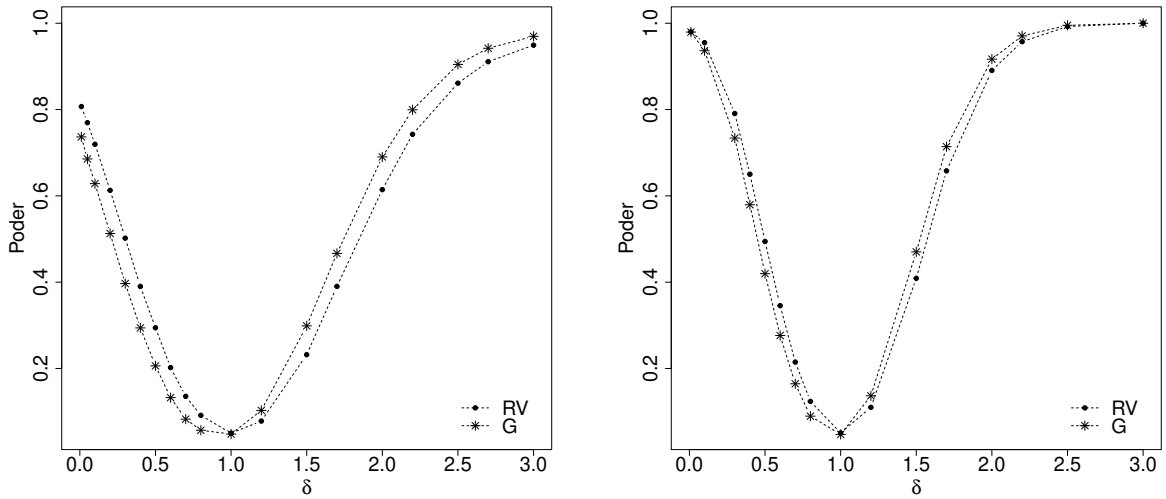


(a) $n = 100$ (b) $n = 200$

(c) $n = 400$ (d) $n = 600$

Figura 1 – Gráficos Q-Q dos testes razão de verossimilhanças e gradiente no modelo hP com $\theta = 4$ contra a distribuição qui-quadrado com 1 grau de liberdade.

Fonte: Elaborada pelo autor.



(a) $n = 200$ (b) $n = 400$

Figura 2 – Poder dos testes razão de verossimilhanças e gradiente para $\alpha = 5\%$ no modelo hP sem covariáveis.

Fonte: Elaborada pelo autor.

Tabela 3 – Médias das EMVs, desvio padrão (DP), raiz do erro quadrático médio (REQM) e a probabilidade de cobertura (PC) dos parâmetros do modelo de regressão hP com covariáveis em θ.

		$\eta = 0,5$			$\eta = 1$			$\eta = 2$		
n		$\hat{\eta}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\eta}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\eta}$	$\hat{\beta}_0$	$\hat{\beta}_1$
100	Média	0,502	0,473	1,024	0,998	0,458	1,034	1,994	0,421	1,059
	DP	0,291	0,200	0,168	0,480	0,258	0,189	0,965	0,364	0,219
	REQM	0,291	0,202	0,170	0,479	0,261	0,192	0,965	0,372	0,225
	PC	0,885	0,947	0,942	0,888	0,949	0,953	0,878	0,935	0,953
200	Média	0,500	0,487	1,013	0,999	0,478	1,019	2,039	0,474	1,023
	DP	0,204	0,136	0,112	0,350	0,182	0,129	0,704	0,249	0,140
	REQM	0,204	0,137	0,113	0,350	0,183	0,130	0,704	0,251	0,141
	PC	0,922	0,958	0,955	0,917	0,951	0,954	0,914	0,956	0,958
400	Média	0,502	0,495	1,002	1,006	0,493	1,005	2,013	0,487	1,008
	DP	0,146	0,102	0,084	0,246	0,134	0,094	0,477	0,179	0,102
	REQM	0,145	0,102	0,084	0,246	0,134	0,094	0,477	0,179	0,102
	PC	0,938	0,948	0,947	0,945	0,945	0,948	0,940	0,946	0,949
600	Média	0,498	0,491	1,001	0,998	0,487	1,012	2,001	0,488	1,014
	DP	0,120	0,083	0,067	0,201	0,128	0,107	0,392	0,146	0,082
	REQM	0,120	0,083	0,068	0,200	0,129	0,108	0,392	0,146	0,083
	PC	0,940	0,949	0,951	0,943	0,951	0,953	0,944	0,948	0,952

Fonte: Elaborada pelo autor.

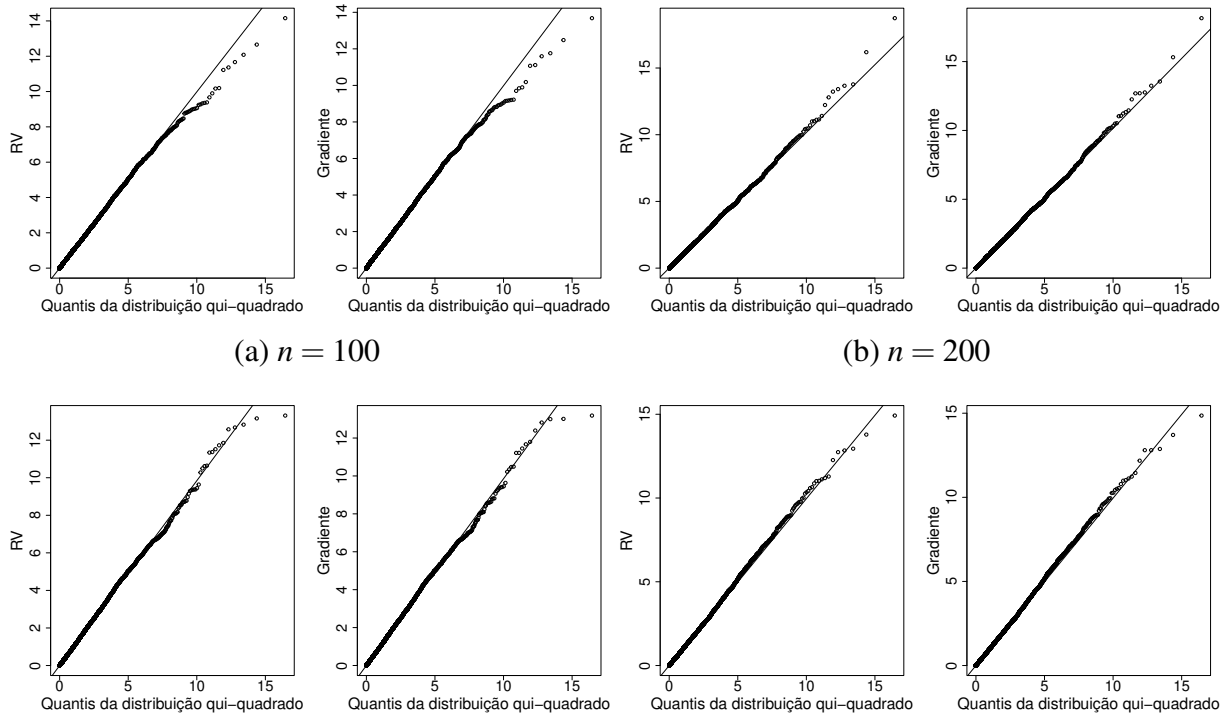
dos testes de hipóteses RV e gradiente no modelo hP, mas agora na presença de covariáveis no parâmetro θ. As análises foram novamente baseadas em 10.000 amostras de tamanhos $n = 100, 200, 400$ e 600 e inicialmente foram observadas as taxas de rejeição de H_0 com níveis de significância nominais de $1\%, 5\%$ e 10%. Os resultados estão dispostos na Tabela 4, dos quais podemos ver claramente que ambos os testes apresentaram taxas de rejeição muito próximas aos níveis nominais adotados para todos os tamanhos de amostras. Os gráficos quantil-quantil apresentados na Figura 3 mostram que a distribuição qui-quadrado com um grau de liberdade fornece uma aproximação razoável para as distribuições sob H_0 das estatísticas RV e gradiente, principalmente com tamanho de amostra superior ou igual a 400.

Na sequência, averiguamos o poder dos testes razão de verossimilhanças e gradiente ao detectar a hipótese alternativa no modelo hP com covariáveis no parâmetro θ. De modo análogo ao caso sem covariáveis, geramos 10.000 amostras com tamanhos $n = 200$ e $n = 400$ sob a hipótese alternativa $H_a : \eta = \delta$, para diferentes valores de δ e adotamos mais uma vez $\alpha = 5\%$. Pela Figura 4, vemos que para o tamanho de amostra $n = 200$ ambos os testes atingem um poder inferior a 80% mesmo com $\eta = 3,5$, isto é, apresentam um poder muito inferior ao desejado quando $\eta > 1$, sendo o o poder do teste gradiente um pouco mais elevado. Já quando $\eta < 1$, os

Tabela 4 – Taxas de rejeição de H_0 dos testes razão de verossimilhanças (RV) e gradiente (G) no modelo de regressão hP com covariáveis em θ.

	Nível de Significância α					
	0,1		0,05		0,01	
n	RV	G	RV	G	RV	G
100	0,102	0,103	0,055	0,054	0,010	0,009
200	0,100	0,099	0,052	0,051	0,009	0,009
400	0,100	0,100	0,051	0,051	0,010	0,009
600	0,098	0,099	0,052	0,052	0,011	0,011

Fonte: Elaborada pelo autor.



(a) $n = 100$ (b) $n = 200$

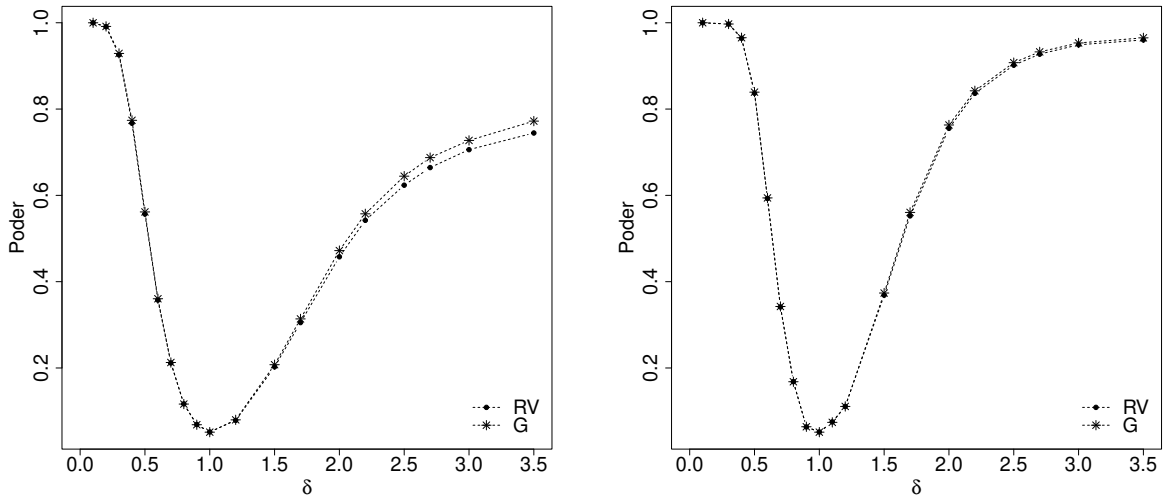
(c) $n = 400$ (d) $n = 600$

Figura 3 – Gráficos Q-Q dos testes razão de verossimilhanças e gradiente no modelo de regressão hP com covariáveis em θ contra a distribuição qui-quadrado com 1 grau de liberdade.

Fonte: Elaborada pelo autor.

testes apresentam praticamente o mesmo valor em todos os pontos e atinge o valor um quando $\eta = 0,1$. Conforme esperado, para o tamanho de amostra $n = 400$ o poder é mais elevado, ambos os testes atingiram um poder superior a 90% quando $\eta = 0,4$ e quando $\eta = 2,5$. Diferente do que observado no caso sem covariáveis, os testes RV e gradiente obtiveram praticamente o mesmo poder com $n = 400$, quando $\eta > 1$ e quando $\eta < 1$.

Para encerrar os estudos de simulação, trabalhamos com o modelo hP parametrizado na média μ. A fim de analisar as propriedades frequentistas dos EMVs, para cada i, $i = 1, \ldots, n$,



(a) $n = 200$ (b) $n = 400$

Figura 4 – Poder dos testes razão de verossimilhanças e gradiente para $\alpha = 5\%$ no modelo de regressão hP com covariáveis em θ.

Fonte: Elaborada pelo autor.

geramos variáveis aleatórias do modelo hP com parâmetros $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, com $\beta_0 = 0,5$, $\beta_1 = 1$ e covariáveis x_i geradas de uma distribuição Bernoulli com parâmetro $0,5$. Tomamos mais uma vez $\eta = 0,5$, $\eta = 1$ e $\eta = 2$ e trabalhamos com 1.000 amostras de tamanhos $n = 100, 200, 400$ e 600. A Tabela 5 traz os resultados dessa simulação. Os histogramas juntamente com os gráficos QQ-normal dos parâmetros estimados podem ser vistos no Apêndice B.

Pela Tabela 5 podemos ver que as estimativas de todos os parâmetros são muito próximas dos valores verdadeiros para todos os tamanhos de amostras considerados. Assim como nos casos anteiros (sem covariáveis e com covariáveis em θ), os desvio padrão e os REQM são menores quando $\eta < 0,5$ e essas medidas vão diminuindo conforme o tamanho da amostra aumenta. Mais uma vez as probabilidades de cobertura dos parâmetros β_0 e β_1 são bem próximas do valor nominal de 95% em todos os cenários e para todos os tamanhos de amostras. Já para o parâmetro η, esta quantidade vai se aproximando de 95% conforme o tamanho da amostra vai aumentando.

De modo análogo aos casos sem covariáveis e com covariáveis em θ, averiguamos os comportamento dos testes RV e gradiente no que diz respeito ao erro do tipo I e ao poder. A Tabela 6 traz as taxas de rejeição de H_0 baseadas em 10.000 simulações dos testes RV e gradiente, da qual podemos ver que ambos os testes apresentam taxas muito próximas aos níveis de significância nominais adotados ($10\%, 5\%$ e 1%) para os quatro tamanhos de amostras considerados.

A Figura 5 traz os gráficos quantil-quantil, pelo qual vemos que a distribuição qui-quadrado com um grau de liberdade fornece uma aproximação razoável para as distribuição sob

Tabela 5 – Médias das EMVs, desvio padrão (DP), raiz do erro quadrático médio (REQM) e a probabilidade de cobertura (PC) dos parâmetros do modelo de regressão hP parametrizado na média.

		$\eta = 0,5$			$\eta = 1$			$\eta = 2$		
n		$\hat{\eta}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\eta}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\eta}$	$\hat{\beta}_0$	$\hat{\beta}_1$
100	Média	0,497	0,495	1,004	0,999	0,494	1,005	1,975	0,491	1,007
	DP	0,268	0,105	0,123	0,479	0,115	0,134	0,874	0,127	0,148
	REQM	0,268	0,105	0,123	0,479	0,115	0,134	0,874	0,127	0,148
	PC	0,898	0,940	0,939	0,886	0,940	0,943	0,894	0,942	0,946
200	Média	0,504	0,498	1,003	1,000	0,497	1,004	2,000	0,496	1,005
	DP	0,192	0,070	0,083	0,348	0,078	0,092	0,636	0,086	0,100
	REQM	0,192	0,070	0,083	0,348	0,078	0,092	0,636	0,086	0,101
	PC	0,914	0,949	0,947	0,917	0,948	0,947	0,927	0,951	0,954
400	Média	0,504	0,500	0,998	1,007	0,499	0,999	2,001	0,499	0,998
	DP	0,136	0,050	0,060	0,247	0,055	0,066	0,446	0,061	0,060
	REQM	0,136	0,050	0,060	0,247	0,055	0,066	0,446	0,061	0,060
	PC	0,941	0,950	0,947	0,945	0,947	0,947	0,931	0,950	0,947
600	Média	0,500	0,496	1,005	0,996	0,496	1,005	1,988	0,495	1,006
	DP	0,110	0,042	0,049	0,200	0,046	0,053	0,357	0,051	0,059
	REQM	0,110	0,042	0,049	0,200	0,046	0,054	0,357	0,051	0,059
	PC	0,941	0,943	0,954	0,946	0,944	0,954	0,943	0,953	0,946

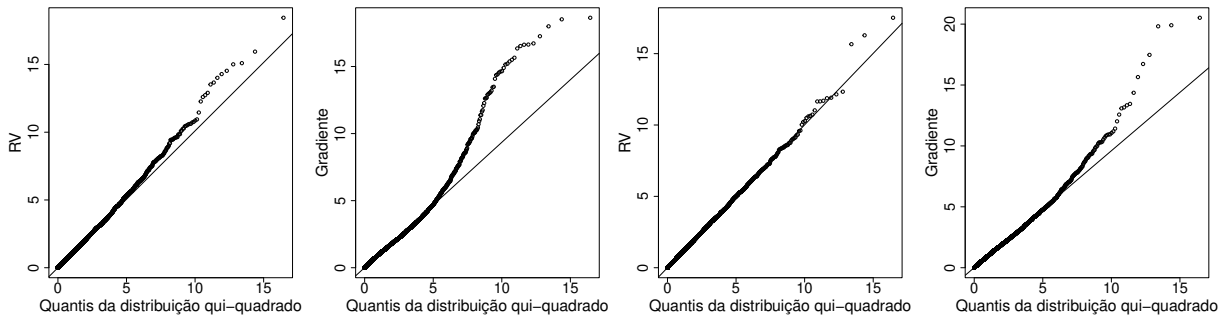
Fonte: Elaborada pelo autor.

Tabela 6 – Taxas de rejeição de H_0 dos testes razão de verossimilhanças (RV) e gradiente (G) no modelo de regressão hP parametrizado na média.

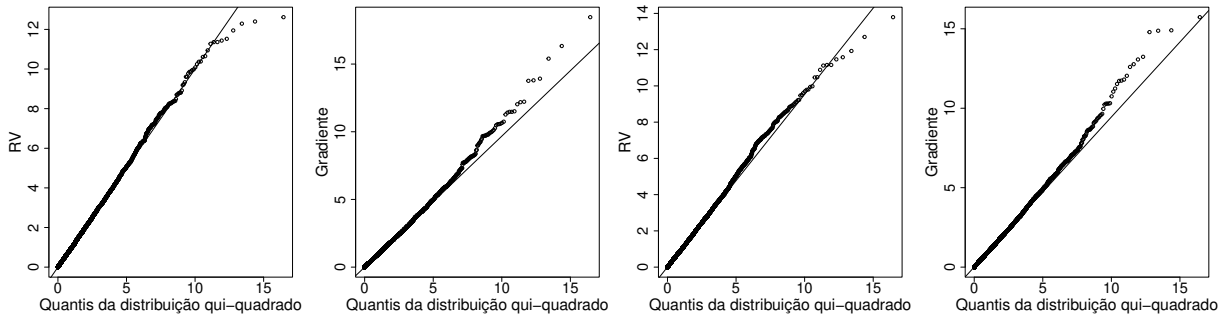
	Nível de Significância α					
	0,1		0,05		0,01	
n	RV	G	RV	G	RV	G
100	0,106	0,083	0,052	0,041	0,013	0,013
200	0,103	0,094	0,052	0,047	0,009	0,011
400	0,097	0,094	0,049	0,048	0,010	0,011
600	0,098	0,095	0,049	0,048	0,010	0,011

Fonte: Elaborada pelo autor.

H_0 das duas estatísticas, sendo a aproximação da estatística razão de verossimilhanças um pouco mais satisfatória que a estatística gradiente.



(a) $n = 100$ (b) $n = 200$



(c) $n = 400$ (d) $n = 600$

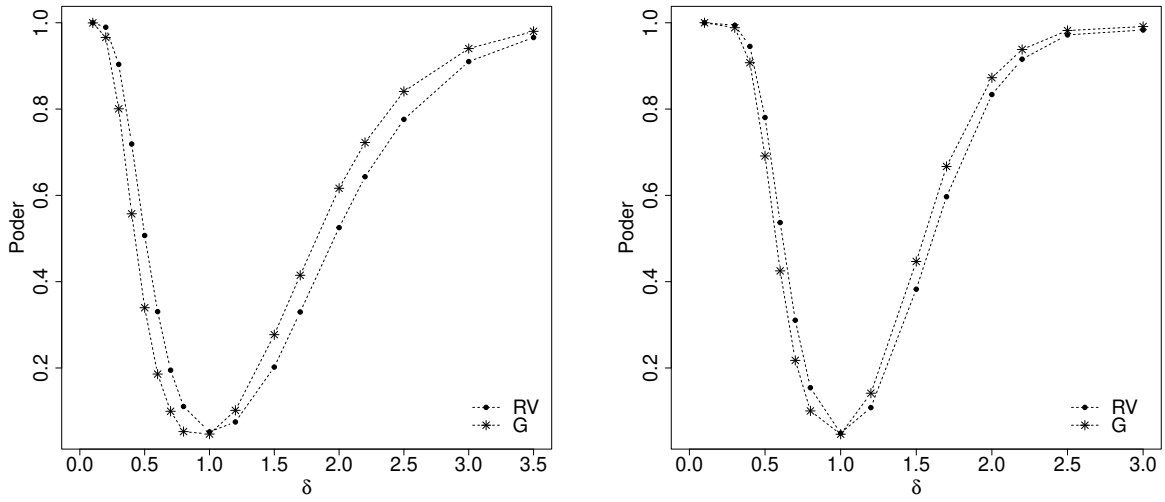
Figura 5 – Gráficos Q-Q dos testes razão de verossimilhanças e gradiente no modelo de regressão hP parametrizado na média contra a distribuição qui-quadrado com 1 grau de liberdade.

Fonte: Elaborada pelo autor.

Para encerrar, mostramos na Figura 6 o comportamento do poder dos testes RV e gradiente obtido de maneira análoga aos casos anteriores. Mais uma vez vemos que, conforme esperado, o poder de ambos os testes é maior com o tamanho de amostra 400. Vemos ainda que, assim como no caso sem covariáveis, o poder do teste gradiente é um pouco maior que o do teste RV quando $\eta > 1$ e isto se inverte quando $\eta < 1$. Podemos observar ainda, que com o tamanho de amostra 200 o poder de ambos os testes é maior quando comparado aos casos sem covariáveis e com com covariáveis no parâmetro θ.

2.7 Aplicação a dados reais

Nesta seção ajustamos o modelo hiper-Poisson a dois conjuntos de dados reais com o intuito de ilustrar a performance dos testes razão de verossimilhanças e gradiente ao testar o parâmetro de dispersão. O primeiro conjunto de dados é um exemplo de dados com subdispersão e o segundo com sobredispersão.



(a) $n = 200$ (b) $n = 400$

Figura 6 – Poder dos testes razão de verossimilhanças e gradiente com $\alpha = 5\%$ no modelo de regressão hP parametrizado na média.

Fonte: Elaborada pelo autor.

2.7.1 Aplicação 1: Bids

Trabalhamos primeiramente com os dados de Cameron, Johansson *et al.* (1997) (utilizados recentemente em Sáez-Castillo e Conde-Sánchez (2013)) sobre o número de lances recebidos por 125 empresas norte americanas que foram alvo de ofertas de leilão entre os anos de 1978 e 1985, e que foram efetivamente assumidas no prazo de 52 semanas após o lance inicial. Esse conjunto de dados está no pacote `Ecdat` do *software R* nomeado por *Bids*. A variável de contagem é o número de lances depois do lance inicial (NUMBIDS) recebidos pelas empresas alvo, cuja distribuição pode ser vista na Figura 7. Os regressores são:

- Ações defensivas tomadas pela administração da empresa alvo: variáveis indicadoras para defesa legal por processo (LEGLREST), alterações propostas na estrutura dos ativos (REALREST), alteração proposta na estrutura de propriedade (FINREST) e convite de gestão para oferta amigável de terceiros (WHITEKNT);
- Características específicas das empresas: preço do lance dividido pelo preço do lance 14 dias úteis antes do lance (BIDPREM), porcentagem de estoque mantido pelas empresas (INSTHOLD) e valor contábil total dos ativos em bilhões de dólares (SIZE);
- Intervenção por reguladores federais: uma variável indicadora para intervenção do Departamento de justiça (REGULATN).

A fim de reproduzir alguns dos cenários do estudo de simulação, primeiramente ajustamos os dados *Bids* ao modelo hP sem covariável e depois ajustamos com todas as covariáveis presentes

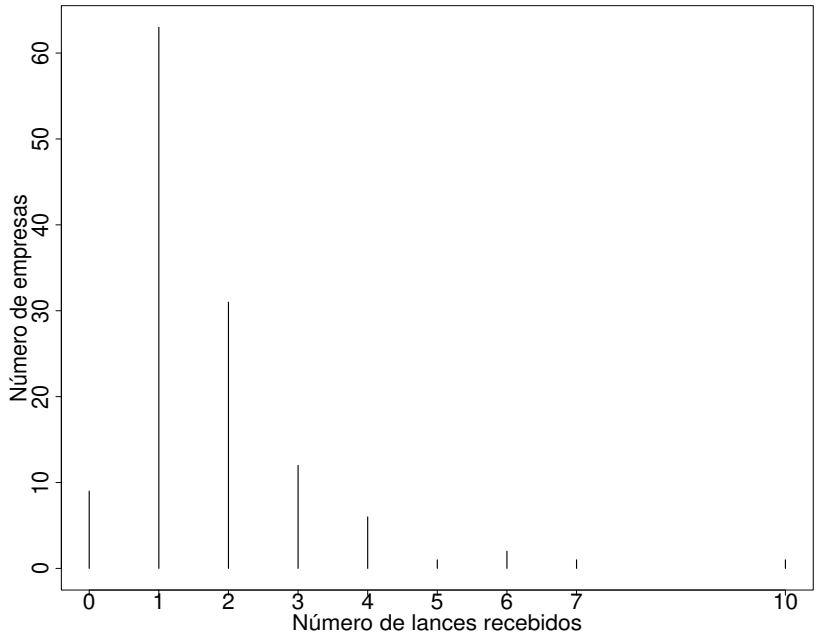


Figura 7 – Distribuição do número de lances recebidos pelas 125 empresas.

Fonte: Elaborada pelo autor.

no parâmetro θ e com as covariáveis na média μ.

A Tabela 7 traz os valores das estimativas de máxima verossimilhança (EMVs) dos parâmetros η e θ do modelo hP sem covariáveis e seus respectivos erros padrão. As estimativas de η são inferiores a um, indicando assim a subdispersão dos dados. Na Tabela 8 estão organizadas as estimativas das estatísticas razão de verossimilhanças e gradiente juntamente com seus respectivos valores-p ao testar a hipótese $H_0 : \eta = 1$ contra $H_1 : \eta \neq 1$. Assim, adotando um nível de significância nominal de 5%, vemos que a hipótese nula é rejeitada por ambos os testes, contudo a estimativa da estatística gradiente apresenta um valor inferior à estimativa da estatística RV, como já observado no estudo de simulação quando $\eta < 1$.

Tabela 7 – EMVs e os erros padrão dos parâmetros do modelo hP ajustado ao conjunto de dados *Bids* sem os regressores.

Parâmetro	Estimativa	Erro padrão
η	0,2553	0,1217
θ	1,0591	0,1566

Fonte: Elaborada pelo autor.

Na Tabela 9 estão listados as estimativas do parâmetro η e dos coeficientes de regressão, e os respectivos erros padrão do modelo de regressão hP com covariáveis no parâmetro θ e do modelo de regressão hP parametrizado na média ajustado aos dados *Bids*.

Tabela 8 – Estimativas das estatísticas RV e gradiente e os respectivos valor-p ao testar a hipótese $H_0 : \eta = 1$ para o conjunto de dados *Bids* no modelo hP.

Estatísticas	Estimativas	Valor-p
RV	7,6661	0,0056
Gradiente	3,9045	0,0481

Fonte: Elaborada pelo autor.

Tabela 9 – EMVs e os erros padrão dos parâmetros dos modelos de regressão hP com covariáveis em θ e parametrizado na média ajustados aos dados *Bids*.

	Covariáveis em θ		Covariáveis na média μ	
Parâmetro	Estimativa	Erro padrão	Estimativa	Erro padrão
η	0,0836	0,0404	0,0797	0,0399
$\beta_{\text{Intercepto}}$	0,7663	0,7306	1,1451	0,4051
β_{LEGLREST}	0,4629	0,2087	0,2519	0,1117
β_{REALREST}	-0,2380	0,2332	-0,1230	0,1391
β_{FINREST}	0,1443	0,2828	0,1103	0,1642
β_{WHITEKNT}	0,9793	0,2519	0,5056	0,1117
β_{BIDPREM}	-1,3964	0,5232	-0,7895	0,2869
β_{INSTHOLD}	-0,2835	0,5403	-0,1656	0,3076
β_{SIZE}	0,0517	0,0227	0,0380	0,0154
β_{REGULATN}	-0,0007	0,2195	0,0147	0,1227

Fonte: Elaborada pelo autor.

Calculamos as estimativas das estatísticas RV e gradiente ao testar o parâmetro de dispersão nos dois modelos de regressão trabalhados. Os resultados obtidos estão na Tabela 10, em que observamos que as estimativas das duas estatísticas são bastante próximas nos dois casos. Além disso, vemos que a estimativa da estatística gradiente é mais uma vez bem menor que a estimativa da estatística RV, mas ambas apresentam valor-p inferior a 5%.

Tabela 10 – Estimativas das estatísticas e os respectivos valores-p ao testar a hipótese $H_0 : \eta = 1$ para o conjunto de dados *Bids* nos modelos de regressão hP.

	Covariáveis em θ		Covariáveis na média μ	
Estatísticas	Estimativas	Valor-p	Estimativas	Valor-p
RV	25,6682	$< 0,0001$	26,2198	$< 0,0001$
Gradiente	8,48446	0,0035	8,6192	0,0033

Fonte: Elaborada pelo autor.

2.7.2 Aplicação 2: Store

Como segunda ilustração consideramos os dados apresentados em Paula (2004) sobre o perfil dos clientes de uma determinada loja oriundos de 109 áreas de uma cidade. A variável de contagem é o número esperado de clientes em cada área, cuja média é igual a 11,22 e a variância igual a 44,45. As variáveis explicativas em cada área utilizadas aqui foram:

- Renda: renda média anual (em dezenas de milhar de dólares);
- Idade: idade média dos domicílios (em décadas);
- Dist 1: distância entre a área e o concorrente mais próximo (em milhas);
- Dist 2: distância entre a área e a loja (em milhas);

Ajustamos mais uma vez os modelos de regressão hP com covariáveis no parâmetro θ e parametrizado na média μ. A Tabela 11 traz as EMVs do parâmetro η e dos coeficientes de regressão juntamente com os respectivos erros padrão dos dois modelos de regressão hP tratados. Vemos pelos resultados do ajuste que a estimativa de η é superior a um, indicando assim a sobredispersão dos dados.

Tabela 11 – EMVs e os erros padrão dos parâmetros dos modelos de regressão hP com covariáveis em θ e parametrizado na média ajustados aos dados *Store*.

	Covariáveis em θ		Covariáveis na média μ	
Parâmetro	Estimativa	Erro padrão	Estimativa	Erro padrão
η	3,3174	1,2768	2,6875	1,2918
$\beta_{\text{Intercepto}}$	3,1600	0,1947	3,0666	0,2349
β_{Renda}	-0,0527	0,0164	-0,0634	0,0199
β_{Idade}	-0,0212	0,0162	-0,0264	0,0192
$\beta_{\text{Dist 1}}$	0,1408	0,0263	0,1654	0,0281
$\beta_{\text{Dist 2}}$	-0,1096	0,0172	-0,1288	0,0181

Fonte: Elaborada pelo autor.

Para finalizar, calculamos as estimativas das estatísticas razão de verossimilhanças e gradiente ao testar a hipótese sobre o parâmetro de dispersão. Os resultados obtidos estão na Tabela 12. Adotando mais uma vez um nível de significância de 5%, vemos que no modelo com covariáveis em θ, ambos os testes rejeitam a hipótese nula. Já no modelo parametrizado na média, o teste RV apresenta o nível descritivo superior ao nível nominal de 5%. Observamos ainda que, em ambos os casos a estimativa da estatística gradiente é superior à estatística RV.

Tabela 12 – Estimativas das estatísticas e os respectivos valor-p ao testar a hipótese $H_0 : \eta = 1$ para o conjunto de dados *Store* com todas as covariáveis.

	Covariáveis em θ		Covariáveis na média μ	
Estatísticas	Estimativas	Valor-p	Estimativas	Valor-p
RV	6,2306	0,0125	3,4201	0,0644
G	9,3387	0,0022	6,7404	0,0094

Fonte: Elaborada pelo autor.

2.8 Alguns comentários

Neste Capítulo a distribuição hiper-Poisson foi apresentada como um modelo muito atrativo por permitir dados com subdispersão ou sobredispersão. De fato, a existência de expressões para a média em termos dos parâmetros simplifica a formulação dos modelos de regressão e o efeito das covariáveis podem ser diretamente avaliados. A principal dificuldade do modelo de regressão hP parametrizado na média está na estimação dos coeficientes, uma vez que o método de máxima verossimilhança implica em um grande esforço computacional.

Trabalhamos com duas distintas parametrizações do modelo de regressão hP. Uma com covariáveis ligadas a um dos parâmetros do modelo (modelo 1) e a outra com as covariáveis na média (modelo 2). No que diz respeito ao processo de estimação dos parâmetros, ambos modelos de regressão apresentaram comportamentos parecidos. Em relação aos testes de hipóteses, podemos ver que o poder dos dois testes se aproximam de um mais rapidamente no modelo 2. As estatísticas da razão de verossimilhanças e gradiente mostraram um comportamento similar quando observamos as taxas de rejeição da hipótese nula. Podemos notar ainda que o poder da estatística da razão de verossimilhanças é ligeiramente superior no caso de subdispersão e menor no caso de sobredispersão, em relação à estatística gradiente.

Uma questão que merece mais atenção aqui e que também foi discutida em Lemonte e Ferrari (2011) é a seguinte: por que os testes de Wald e escore não foram incluídos nos estudos de simulação anteriores? Lembremos que as estatísticas de Wald e escore envolvem a matriz de informações de Fisher, que não pode ser obtida analiticamente para o modelo de regressão hP. Uma prática comum é usar a matriz de informação observada em vez da matriz de informação esperada, então seguimos essa abordagem executando vários experimentos de simulação, incluindo os testes de Wald e escore. Nossa conclusão é que esses testes não podem ser recomendados pelos seguintes motivos: primeiro, o teste de escore no modelo 2 foi muito conservador. Por exemplo, para $n = 200$ e $n = 600$ com $\alpha = 0,05$, as taxas de rejeição da hipótese nula do teste de escore para testar $H_0 : \eta = 1$ são $0,025$ e $0,020$, respectivamente. Outro motivo está no fato de que no caso de subdispersão, em ambos modelos de regressão, o poder do teste de Wald é muito baixo, mesmo para $\eta = \delta = 0,1$, em que os testes da razão de verossimilhanças e gradiente têm poder próximo a 1.

CAPÍTULO

3

UM MODELO DE SOBREVIVÊNCIA COM FRAGILIDADE HIPER-POISSON: ENFOQUE CLÁSSICO

Neste capítulo um novo modelo de sobrevivência com fração de cura é introduzido. Este modelo é obtido a partir de uma variável de fragilidade com distribuição hiper-Poisson. Dois tópicos da análise de sobrevivência essenciais para esta etapa do trabalho são descritos. Discutimos aspectos de inferência para o modelo proposto em uma abordagem clássica, em que exploramos as ferramentas de máxima verossimilhança por meio do algoritmo EM (*Expectation-Maximization*) e utilizamos ainda a técnica da verossimilhança perfilada para estimar um dos parâmetros do modelo.

3.1 Análise de sobrevivência

Nesta seção falamos um pouco sobre os modelos com fração de cura e modelos de fragilidade, ambos dentro da área de análise de sobrevivência. Esses modelos embasaram a pesquisa para esta parte do desenvolvimento da tese, na qual a distribuição hiper-Poisson foi empregada para modelar variáveis latentes.

3.1.1 Modelos com fração de cura

Na análise de sobrevivência é comum assumir que cada indivíduo da amostra está suscetível ao evento de interesse e, nestes casos, quando os indivíduos não experimentam o evento de interesse até certo momento do estudo, eles são considerados censurados. Dessa maneira, estes estudos não levam em conta o caso em que os indivíduos podem simplesmente não apresentar o evento de interesse. Por exemplo, muitas vezes em sobrevivência o evento de interesse é a morte de um paciente ou a recorrência de um tumor, mas devido aos grandes avanços

na medicina é esperado que uma elevada parcela dos indivíduos seja curada, ou seja, deixe de ser suscetível ao evento de interesse. Modelos que consideram que uma parte da população pode ser ou se tornar suscetível a certo evento de interesse são chamados modelos de sobrevivência com fração de cura ou ainda modelos de longa duração.

Os modelos com fração de cura foram abordados inicialmente por Boag (1949) e Berkson e Gage (1952), e ficaram conhecidos na literatura como modelos de mistura padrão. Estes modelos supõem a existência de uma possível causa interferindo para a ocorrência do evento, e que tal causa se manifesta ou não segundo uma probabilidade a ser calculada. Tomando como T a variável aleatória que representa o tempo até a ocorrência do evento de interesse, descrevemos brevemente o procedimento probabilístico deste modelo. Seja Z_i uma variável aleatória Bernoulli associada ao indivíduo i, com probabilidade de sucesso p, de tal forma que Z_i assume o valor 1 se o indivíduo i é suscetível ao evento, e assume o valor 0 se o indivíduo é imune. Assim, p indica a proporção de indivíduos suscetíveis na população. Z_i não é observada, pois não se sabe se um indivíduo é imune ou não. Os indivíduos suscetíveis em algum momento do estudo apresentarão o evento de interesse, com função de sobrevivência $S(t)$ própria, enquanto que os indivíduos com $Z_i = 0$ não apresentarão o evento de interesse. Assim sendo,

$$P(T > t|Z_i = 1) = S(t) \quad \text{e} \quad P(T > t|Z_i = 0) = 1.$$

Considerando uma população em que existe a possibilidade de indivíduos imunes, estas probabilidades implicam que a probabilidade de o evento ocorrer após o tempo t para um indivíduo qualquer é dada por

$$S_{\text{pop}}(t) = P(T > t) = 1 - p + pS(t),$$

em que S_{pop} representa a função de sobrevivência da população, que é imprópria, pois $S_{\text{pop}}(\infty) = 1 - p$ e corresponde à proporção de indivíduos curados.

Yakovlev e Tsodikov (1996) e Chen, Ibrahim e Sinha (1999) estudaram um outro modelo de longa duração que ficou conhecido como modelo de tempo de promoção. Neste modelo a função de sobrevivência da população é dada por

$$S_{\text{pop}}(t) = \exp\{-\theta F(t)\}, \quad \theta > 0,$$

em que $F(t)$ é uma função de distribuição própria. A proporção de indivíduos imunes aqui é dada por $S_{\text{pop}}(\infty) = \exp(-\theta)$.

A literatura sobre modelos de fração de cura é extensa e está em rápido desenvolvimento, destacando-se como referências fundamentais os livros de Maller e Zhou (1996a) e Ibrahim, Chen e Sinha (2001b), além dos trabalhos de Tsodikov, Ibrahim e Yakovlev (2003), Yin e Ibrahim (2005), Cooner *et al.* (2007), Rodrigues *et al.* (2009a), Cancho, Rodrigues e Castro (2011), Cancho, Castro e Rodrigues (2012) e Cordeiro *et al.* (2016), entre outros.

3.1.2 Modelos de fragilidade

Frequentemente, a análise de dados de sobrevivência é baseada no pressuposto de que a população em estudo é homogênea. Em outras palavras, condicionados às covariáveis, cada indivíduo apresenta o mesmo risco de experimentar um evento de interesse, tal como a morte ou a recorrência de uma doença. Porém, em estudos médicos, por exemplo, é muito fácil concordar com o pressuposto de que mesmo os indivíduos que exibem os mesmos valores de covariáveis não são previstos a apresentar qualquer resposta de tratamento exatamente no mesmo tempo, pois existem variações biológicas não mensuráveis ou até mesmo não observáveis, como fatores genéticos, que justificam essa heterogeneidade entre os indivíduos. Além disso, é muito comum assumir que os tempos de eventos dos indivíduos na população, condicionados às covariáveis observáveis, são também independentes. No entanto, essa suposição não está sempre adequada à realidade dos objetos em estudo, visto que algumas vezes os tempos de sobrevivência são observados em grupos, e tais tempos podem não ser independentes. Por exemplo, quando as análises trazem sujeitos de uma mesma família, é razoável supor que haja alguma associação entre os tempos de interesse deste grupo familiar. São nestas conjunturas que aparecem os modelos de fragilidade (*frailty models*), sendo caracterizados pela inserção de uma variável aleatória não observada, a fragilidade, que no contexto de sobrevivência univariada, representa uma medida de heterogeneidade, enquanto que no contexto multivariado, é ainda uma medida de associação entre os indivíduos.

Para solucionar o problema da heterogeneidade não observada devido às covariáveis não observadas, Vaupel, Manton e Stallard (1979) e Lancaster (1979) sugeriram independentemente a inclusão de um efeito aleatório no modelo. O termo fragilidade foi referido pioneiramente por Vaupel, Manton e Stallard (1979) para a heterogeneidade não observada em modelos de sobrevivência univariados, que inseriu esse conceito de fragilidade em bioestatística com aplicações em dados de mortalidade populacional. Lancaster (1979) trabalhou com dados sobre desemprego, e então incorporou os modelos de fragilidade em econometria, que são conhecidos como modelos de riscos proporcionais mistos. Os primeiros modelos de fragilidades para dados de sobrevivência multivariados foram apresentados por Clayton (1978) e Oakes (1982).

O modelo de fragilidade mais tradicional assume uma estrutura de risco proporcional condicionado ao efeito aleatório (fragilidade). Este efeito aleatório, Z, que em geral é uma variável aleatória contínua não negativa, indica o nível individual de risco e é adicionado na função de risco como um fator multiplicativo. Em outras palavras, o modelo de fragilidade é basicamente especificado por meio da função de risco

$$h(t|Z) = Z\, h_B(t), \tag{3.1}$$

em que $h_B(t)$ é a função de risco basal. Na presença de covariáveis o modelo é usualmente estendido como

$$h(t|Z) = Z\exp(\boldsymbol{x}^\top\boldsymbol{\beta})h_B(t).$$

A função de sobrevivência correspondente, condicionada em Z é dada por

$$S(t|Z) = P(T > t|Z) = \exp\{-ZH_B(t)\} = S_B(t)^Z,$$

em que $H_B(t) = \int_0^t h_B(u)du$ é a função de risco acumulada basal e $S_B(t)$ é a função de sobrevivência basal correspondente. A função de sobrevivência incondicional, $S(t)$, pode ser expressa como

$$S(t) = \int S(t|Z)dF(z) = \int \exp\{-ZH_B(t)\}\, dF(z) = \mathscr{L}(H_B(t)),$$

em que $F(z)$ é a função de distribuição acumulada de Z e $\mathscr{L}(H_B(t))$ é a transformada de Laplace de Z.

As distribuições mais utilizadas para modelar a variável de fragilidade são a Gama (VAUPEL; MANTON; STALLARD, 1979; HOUGAARD, 1984), Gaussiana Inversa (HOUGAARD, 1984), Log-normal (SANTOS; DAVIES; FRANCIS, 1995), Gama Generalizada (BALAKRISHNAN; PENG, 2006) e Estável Positiva (HOUGAARD, 1986). Do ponto de vista computacional, estas distribuições são convenientes devido à facilidade em derivar expressões com forma fechada para as funções de sobrevivência, densidade e de risco usando a transformada de Laplace. No entanto, a distribuição de fragilidade contínua não permite a possibilidade de risco nulo (ATA; ÖZEL, 2013).

Fragilidade nula indica que existe um subgrupo de indivíduos não suscetíveis entre os quais o evento de interesse não ocorre mesmo após um longo período de observação. Neste contexto há uma conexão com os modelos de longa duração, dos quais muitos deles foram construídos em um cenário de riscos competitivos (TSODIKOV; IBRAHIM; YAKOVLEV, 2003; RODRIGUES *et al.*, 2009b; CANCHO; RODRIGUES; CASTRO, 2011), mas eles também podem ser obtidos por meio dos modelos de riscos proporcionais com distribuição de fragilidade discreta.

Fragilidades com distribuição discreta podem ser requisitadas ainda, por exemplo, quando a heterogeneidade no tempo de vida surge devido à presença de um número desconhecido de falhas em uma unidade em teste, ou o número desconhecido de causas que levam à exposição a determinado dano (CARONI; CROWDER; KIMBER, 2010). Recentemente, alguns trabalhos têm abordado modelos de fragilidade discreta, dentre eles podemos citar Caroni, Crowder e Kimber (2010), que consideram modelos de riscos proporcionais paramétricos permitindo distribuições como a Geométrica, Poisson ou Binomial Negativa, e Wienke (2010) e Ata e Özel (2013) que trabalham com modelos de fragilidade determinadas por um processo de Poison composto discreto.

Nesse contexto, a fragilidade Z, que indica o número de fatores de risco, assume valores inteiros não negativos, isto é, Z tem distribuição discreta com suporte em $\{0,1,2,\ldots\}$ no lugar de uma distribuição contínua com suporte em $(0,\infty)$. Seja $P(Z = z)$ a distribuição de probabilidade de Z especificada para $z = 0,1,2,\ldots$. Então, assumindo um modelo de riscos proporcionais (3.1),

a função de sobrevivência incondicional de T é dada por

$$S(t) = \sum_{z=0}^{\infty} S_B(t)^z P[Z = z] = E[S_B(t)^Z] = G_Z(S_B(t)), \tag{3.2}$$

em que $G_Z(s)$ é a função geradora de probabilidade de Z (CARONI; CROWDER; KIMBER, 2010).

O modelo em (3.2) é o mesmo modelo de sobrevivência com fração de cura obtido no cenário de riscos competitivos por Rodrigues *et al.* (2009b) e Tsodikov, Ibrahim e Yakovlev (2003). Além disso, uma vez que $\lim_{t\to\infty} S_B(t) = 0$ e $\lim_{t\to\infty} S(t) = G_Z(0) = P(Z = 0) > 0$, a expressão em (3.2) é uma função de sobrevivência não própria. Isto caracteriza modelos de sobrevivência com fração de cura, em que $p_0 = P(Z = 0)$ denota a proporção de indivíduos curados.

3.2 Formulação do modelo

O novo modelo de sobrevivência que vamos introduzir aqui é obtido quando a variável de fragilidade Z na expressão (3.2) segue uma distribuição hiper-Poisson com função geradora de probabilidade dada por (2.2). De acordo com essa configuração, a função de sobrevivência em (3.2) é expressa na forma

$$S(t) = \frac{{}_1F_1(1;\eta;\theta S_B(t))}{{}_1F_1(1;\eta;\theta)}. \tag{3.3}$$

A fração de indivíduos curados é dada por $p_0 = \lim_{t\to\infty} S(t)$. Pela expressão (3.3), temos que

$$p_0 = \lim_{t\to\infty} S(t) = \frac{1}{{}_1F_1(1;\eta;\theta)} > 0. \tag{3.4}$$

indicando que (3.3) não é uma função de sobrevivência própria.

A função de densidade associada a (3.3) é escrita como

$$f(t) = \frac{{}_1F_1(2;\eta+1;\theta S_B(t))}{{}_1F_1(1;\eta;\theta)} \frac{\theta}{\eta} f_B(t), \tag{3.5}$$

em que $f_B(t) = -dS_B(t)/dt$ e a função de risco correspondente é dada por

$$h(t) = \frac{{}_1F_1(2;\eta+1;\theta S_B(t))}{{}_1F_1(1;\eta;\theta S_B(t))} \frac{\theta}{\eta} f_B(t).$$

A Figura 8 mostra a flexibilidade na função de sobrevivência e na função de risco em termos do parâmetro de dispersão adicional η introduzido no modelo. Além disso, no caso especial em que $\eta = 1$, o modelo se reduz ao modelo de fração de cura investigado por Yakovlev e Tsodikov (1996) e Chen, Ibrahim e Sinha (1999).

Esse novo modelo, que chamaremos a partir de agora como modelo com fração de cura hP, é identificável, como pode ser visto pela proposição a seguir.

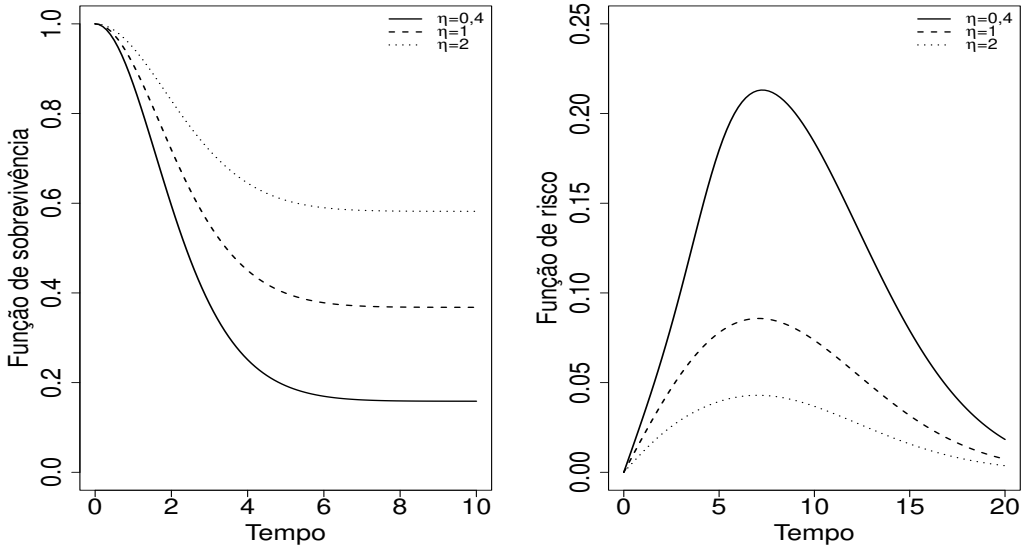


Figura 8 – Função de sobrevivência (painel esquerdo) e função de risco (painel direito) com distribuição de sobrevivência basal $S_B(t) = \exp(-0,1t^2)$ e $\theta = 1$.

Fonte: Elaborada pelo autor.

Proposição 1. O modelo com fração de cura hP representado pela função de sobrevivência dada por (3.3) é identificável.

Demonstração. Sejam $\boldsymbol{\nu}_1 = (\eta_1, \theta_1, \boldsymbol{\phi}_1)$ e $\boldsymbol{\nu}_2 = (\eta_2, \theta_2, \boldsymbol{\phi}_2)$ tal que $\boldsymbol{\nu}_1 \neq \boldsymbol{\nu}_2$, em que $\boldsymbol{\phi}_1$ e $\boldsymbol{\phi}_2$ são os vetores de parâmetros da função de sobrevivência basal. Suponhamos que $S(t;\boldsymbol{\nu}_1) = S(t;\boldsymbol{\nu}_2)$ para todo $t > 0$, que de (3.3) implica que

$$\frac{{}_1F_1(1,\eta_2,\theta_2)}{{}_1F_1(1;\eta_1;\theta_1)} = \frac{{}_1F_1(1;\eta_2;\theta_2 S_B(t;\boldsymbol{\phi}_2))}{{}_1F_1(1;\eta_1;\theta_1 S_B(t;\boldsymbol{\phi}_1))}. \tag{3.6}$$

Sabemos que ${}_1F_1(1;\eta;\theta) = \sum_{r=0}^{\infty} \frac{\theta^r}{(\eta)_r}$. Dessa maneira, se tomarmos $\theta_2 > \theta_1$ e $\eta_1 > \eta_2$, segue que

$$\frac{{}_1F_1(1;\eta_2;\theta_2)}{{}_1F_1(1;\eta_1,\theta_1)} > 1.$$

Por outro lado, como $\boldsymbol{\phi}_1 \neq \boldsymbol{\phi}_2$, é possível encontrar um t_0 tal que $\theta_2 S_B(t_0;\boldsymbol{\phi}_1) < \theta_1 S_B(t_0;\boldsymbol{\phi}_2)$, o que implica que

$$\frac{{}_1F_1(1,\eta_2,\theta_2 S_B(t_0;\boldsymbol{\phi}_2))}{{}_1F_1(1,\eta_1,\theta_1 S_B(t_0;\boldsymbol{\phi}_1))} < 1.$$

Portanto, a igualdade em (3.6) não pode ser satisfeita, concluindo a demonstração. □

A função de sobrevivência (própria) para os indivíduos em risco, denotada por S_{R} é calculada por $S_{\mathrm{R}}(t) = P(T > t \mid Z \geq 1)$ e é dada por

$$S_{\mathrm{R}}(t) = \frac{{}_1F_1(1\eta;\theta S_B(t)) - 1}{{}_1F_1(1;\eta;\theta) - 1},\ t > 0. \tag{3.7}$$

Observemos que $S_{\mathrm{R}}(0) = 1$ e $S_{\mathrm{R}}(\infty) = 0$, o que implica que é uma função de sobrevivência própria.

A relação matemática entre o modelo (3.3) e o modelo com fração de cura de mistura (BOAG, 1949; BERKSON; GAGE, 1952) é dada por

$$S(t) = \frac{1}{{}_1F_1(1,\eta,\theta)} + \left[1 - \frac{1}{{}_1F_1(1,\eta,\theta)}\right] S_{\mathrm{R}}(t), \tag{3.8}$$

em que $S_{\mathrm{R}}(t)$ é dada por (3.7). Assim, $S(t)$ é um modelo de mistura com fração de cura com proporção de curados igual a $p_0 = \frac{1}{{}_1F_1(1,\eta,\theta)}$ e função de sobrevivência $S_{\mathrm{R}}(t)$ para a população em risco. Esse resultado implica que cada modelo de fração de cura de mistura corresponde a algum modelo da forma (3.3) para quaisquer η, θ e $S_{\mathrm{B}}(\cdot)$.

3.3 Inferência

Assumimos que os tempos de falha $T_1, \ldots, T_n$ de n indivíduos não são completamente observados e estão sujeitos à censura à direita, ou seja, consideramos que os indivíduos censurados têm tempos de falha maiores que os tempos de censura. Denotamos por C_i o tempo de censura do i-ésimo indivíduo, então observamos $t_i = \min\{T_i, C_i\}$ e

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é uma observação censurada,} \end{cases}$$

para $i = 1, \ldots, n$. Incorporamos as covariáveis por meio do parâmetro θ. Para cada indivíduo i, seja $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ representando o vetor de covariáveis e seja $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denotando o vetor de coeficientes de regressão correspondente. Relacionamos θ às covariáveis por meio da função $\log(\theta_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, isto é, $\theta_i = \exp(x_i^\top \beta)$.

Baseados nos n pares $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, a função de verossimilhança marginal observada pode ser escrita como

$$L = L(\boldsymbol{\vartheta}; \boldsymbol{x}, \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} f_p(t_i, \boldsymbol{x_i}; \boldsymbol{\vartheta})^{\delta_i} S_p(t_i, \boldsymbol{x_i}; \boldsymbol{\vartheta})^{1-\delta_i},$$

em que $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \eta)^\top$ é o conjunto dos parâmetros do modelo,

$$S_p(t_i, \boldsymbol{x_i}; \boldsymbol{\vartheta}) = \frac{{}_1F_1((1; \eta; \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) S_B(t_i, ; \boldsymbol{\gamma}))}{{}_1F_1(1; \eta; \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))}$$

e

$$f_p(t_i, \boldsymbol{x_i}; \boldsymbol{\vartheta}) = \frac{{}_1F_1(2, \eta+1, \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) S_B(t_i; \boldsymbol{\gamma}))}{{}_1F_1(1; \eta; \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))} \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{\eta} f_B(t_i; \boldsymbol{\gamma}),$$

em que $S_B(t_i; \boldsymbol{\gamma})$ e $f_B(t_i; \boldsymbol{\gamma})$ são as funções de sobrevivência e densidade de probabilidade basal, respectivamente.

A função de verossimilhança pode ser escrita como

$$L \propto \prod_{i=1}^{n}(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta))^{\delta_i} f_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta})^{\delta_i} \times [p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)+(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta))S_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta})]^{1-\delta_i}, \tag{3.9}$$

em que

$$p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta) = \frac{1}{{}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}$$

e

$$S_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta}) = \frac{S_p(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta})-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)}{p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)} \quad \text{e} \quad f_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta}) = \frac{f_p(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta})}{1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)}$$

são as funções de sobrevivência e de densidade de probabilidade dos indivíduos em risco (suscetíveis), respectivamente.

Alternativamente, a função de verossimilhança pode ser expressa como

$$L \propto \prod_{i\in\Delta_1}(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)) \prod_{i\in\Delta_1} f_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta}) \prod_{i\in\Delta_0}[p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)+(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta))S_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta})],$$

em que $\Delta_i = \{i \in \{1\ldots,n\} : \delta_i = 1\}$ e $\Delta_0 = \{i \in \{1,\ldots,n\} : \delta_i = 0\}$ são os conjuntos de observações censuradas e não censuradas, respectivamente.

O i-ésimo elemento do conjunto de observações censuradas pode ser derivado de dois grupos diferentes, indivíduos sob risco ou curados. Suponha que definamos uma variável latente I que indica este evento. Seja I_i a i-ésima variável latente dada por

$$I_i = \begin{cases} 1, & \text{se suscetível,} \\ 0, & \text{se curado.} \end{cases}$$

Então, a função de verossimilhança completa pode ser escrita como

$$L_c \propto \prod_{i\in\Delta_1}(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)) \prod_{i\in\Delta_1} f_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta}) \prod_{i\in\Delta_0} p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)^{1-I_i} \prod_{i\in\Delta_0}[(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta))S_R(t_i;\boldsymbol{x}_i;\boldsymbol{\vartheta})]^{I_i}.$$

Para o passo E do algoritmo EM, precisamos calcular a esperança da função de log-verossimilhança dos dados completos com respeito à distribuição dos dados não observados I_i, dado os valores atualizados dos parâmetros e os dados observados $\boldsymbol{O}$. Notemos que I_i's são variáveis aleatórias Bernoulli na verossimilhança completa e então precisamos calcular $w_i^{(z)} = E[I_i|\boldsymbol{\vartheta}^{*(z)},\boldsymbol{O}]$, em que $\boldsymbol{\vartheta}^* = (\boldsymbol{\beta}^\top,\boldsymbol{\gamma}^\top)$ e $\boldsymbol{\vartheta}^{*(z)}$ representa o valor atual do parâmetro na z-ésima iteração. Dessa forma,

$$\begin{aligned} w_i^{(z)} = E[I_i|\boldsymbol{\vartheta}^{*(z)},\boldsymbol{O}] &= P(I_i=1|T_i>t_i) = \left.\frac{P(T_i>t_i|I_i=1)P(I_i=1)}{P(T_i>t_i)}\right|_{\boldsymbol{\vartheta}^{*(z)}} \\ &= \frac{S_p(t_i,\boldsymbol{x_i};\boldsymbol{\vartheta}^{*(z)})-p_0(\boldsymbol{x}_i;\boldsymbol{\beta}^{(z)},\eta^{(z)})}{S_p(t_i,\boldsymbol{x_i};\boldsymbol{\vartheta}^{*(z)})} \end{aligned}$$

e a esperança condicional da função de verossimilhança completa é dada por

$$Q(\boldsymbol{\vartheta}, \boldsymbol{w}^{(z)}) = \sum_{i\in\Delta_1} [\log(1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)) + \log f_R(t_i,\boldsymbol{x_i};\boldsymbol{\vartheta})] + \sum_{i\in\Delta_0} \log(p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta) + \sum_{i\in\Delta_0} w_i^{(z)} \left[\log\left(\frac{1-p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)}{p_0(\boldsymbol{x}_i;\boldsymbol{\beta},\eta)}\right) + \log S_R(t_i,\boldsymbol{x_i};\boldsymbol{\vartheta})\right],$$

em que $\boldsymbol{w}^{(z)} = \{w_i^{(z)} : i \in \Delta_0\}$.

No passo M do algoritmo EM, maximizamos a função $Q(\boldsymbol{\vartheta}, \boldsymbol{w}^{(z)})$ com respeito a $\boldsymbol{\vartheta}^*$ sobre o correspondente parâmetro do espaço paramétrico Θ^*, dado $\boldsymbol{w}^{(k)}$. Dessa maneira, obtemos uma melhor estimativa de $\boldsymbol{\vartheta}^*$ dada por

$$\boldsymbol{\vartheta}^{*(k+1)} = \arg\max_{\boldsymbol{\vartheta}^*\in\Theta^*} Q(\boldsymbol{\vartheta}^*, \boldsymbol{w}^{(z)}).$$

Para um valor fixado de η, os passos E e M são então continuados iterativamente até a convergência para encontrar os estimadores de máxima verossimilhança do parâmetro $\boldsymbol{\vartheta}^*$. Neste trabalho, como as estimativas de máxima verossimilhança não têm forma explícita, a etapa da maximização é executada por meio do algoritmo EM gradiente (LANGE, 1995), que é um método de Newton-Raphson de uma etapa e qualificado como um caso especial do algoritmo EM generalizado (DEMPSTER; LAIRD; RUBIN, 1977).

Assim como feito em Koutras e Milienos (2017), o parâmetro η é então estimado por meio da verossimilhança perfilada. Nesta técnica, fixamos um conjunto de valores admissíveis de η e para cada valor de η, estimamos o parâmetro $\boldsymbol{\vartheta}^*$. Finalmente, os valores que melhor ajustam nossos dados são aqueles que minimizam o critério de informação Akaike (AIC)= $-2\ell + 2k$, em que ℓ é o valor da log-verossimilhança maximizada do modelo e k é o número de parâmetros do modelo ajustado. As propriedades numéricas desse procedimento de estimação são estudadas na próxima seção.

Os procedimentos inferenciais discutidos neste trabalho são desenvolvidos assumindo η fixado. As estatísticas de testes e as estimativas intervalares para $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, sob condições de regularidade adequadas (MALLER; ZHOU, 1996b) pode ser mostrado que a distribuição assintótica do EMV $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ é normal multivariada com vetor de média $\boldsymbol{\vartheta}$ e matriz de covariâncias

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\vartheta}}) = \left\{ -\frac{\partial^2 \ell(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top} \right\}^{-1} = \{-J(\boldsymbol{\vartheta})\}^{-1},$$

avaliados em $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$ e $\ell(\boldsymbol{\vartheta}) = \log L(\boldsymbol{\vartheta})$, obtida por meio expressão 3.9.

Os elementos da matriz observada $J(\boldsymbol{\vartheta})$ são obtidos numericamente. O pacote `numDeriv` da linguagem R (R Core Team, 2017) fornece uma aproximação numérica precisa para essa matriz.

Um intervalo de confiança assintótico com nível de significância α para cada parâmetro ϑ_r é dado por

$$\left(\hat{\vartheta}_r - z_{\alpha/2}\sqrt{\widehat{\Sigma}^{r,r}}, \hat{\vartheta}_r + z_{\alpha/2}\sqrt{\widehat{\Sigma}^{r,r}}\right),$$

em que $\widehat{\Sigma}^{r,r}$ é o r-ésimo elemento da diagonal de $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\vartheta}})$ estimado por $\widehat{\boldsymbol{\vartheta}}$, para $r = 1, \ldots, p + \dim(\boldsymbol{\varphi}) + 1$, $\dim(\cdot)$ denotando a dimensão do espaço paramétrico e $z_{\alpha/2}$ é o quantil $1 - \alpha/2$ da distribuição normal padrão.

Além das estimativas, os testes de hipóteses são outras questões importantes aqui. Seja $\boldsymbol{\vartheta}_1$ e $\boldsymbol{\vartheta}_2$ subconjuntos disjuntos próprios de $\boldsymbol{\vartheta}$. Suponhamos que temos interesse em testar $H_0 : \boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_{10}$ versus $H_1 : \boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_{10}$, em que $\boldsymbol{\vartheta}_{10}$ é um vetor especificado. O vetor $\boldsymbol{\vartheta}_2$ pode ser considerado como um vetor de parâmetros de pertubação. Seja $\widehat{\boldsymbol{\vartheta}}_0$ o EMVs sob H_0 e então podemos definir a estatística da razão da verossimilhanças dada por $\text{RV} = 2\left\{\ell(\widehat{\boldsymbol{\vartheta}}) - \ell(\widehat{\boldsymbol{\vartheta}}_0)\right\}$ e a estatística gradiente dada por $\text{G} = \boldsymbol{U}_1(\widetilde{\boldsymbol{\vartheta}})^\top(\widehat{\boldsymbol{\vartheta}}_1 - \boldsymbol{\vartheta}_{10})$, em que $\boldsymbol{U}_1(\boldsymbol{\vartheta}) = \partial\ell(\boldsymbol{\vartheta};\boldsymbol{z})/\partial\boldsymbol{\vartheta}_1$ e $\widetilde{\boldsymbol{\vartheta}} = (\boldsymbol{\vartheta}_{10}^\top, \widetilde{\boldsymbol{\vartheta}}_2^\top)^\top$ são os EMV obtidos sob a hipótese nula. Sob H_0 e sob certas condições de regularidade as estatísticas RV e gradiente convergem em distribuição para uma distribuição qui-quadrada com q grau de liberdade, sendo q a dimensão do vetor $\boldsymbol{\vartheta}_1$) (COX; HINKLEY, 1974).

3.4 Estudo de simulação

Nesta seção realizamos dois estudos de simulação a fim de avaliar o método inferencial proposto. Devido ao alto custo computacional, conduzimos um breve estudo em que o parâmetro η foi estimado por meio da verossimilhança perfilada e um segundo estudo, com mais iterações, em que mantivemos o parâmetro η fixado, assumindo valores menor do que 1, igual a 1 e maior do que 1, uma vez que este é o parâmetro de dispersão.

No primeiro estudo consideramos a distribuição de Weibull como a distribuição basal com parâmetros $\gamma_1 = 1{,}5$ e $\gamma_2 = 2$. Para cada indivíduo, $i = 1, \ldots, n$, a variável de fragilidade foi gerada de uma distribuição hP com parâmetros $\eta = 6$ e $\theta = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, com $\beta_0 = 0{,}5$, $\beta_1 = 1$ e $\beta_2 = 0{,}8$ e as covariáveis x_i e x_2 foram geradas de uma distribuição Bernoulli com parâmetro 0,5 e de uma distribuição uniforme $[0, 1]$, respectivamente.

Foram construídos três conjuntos de dados de tamanhos $n = 200, 400$ e 600 e consideramos 500 iterações. Os tempos de censura C_i foram amostrados de uma distribuição uniforme e a proporção de observações censuradas neste caso é de aproximadamente 47%. A grade de pesquisa de η foi conduzida no intervalo $[5{,}5; 6{,}5]$ com passo igual a 0,1 e os valores iniciais para o algoritmo EM foram $1{,}2; 1{,}7; 0{,}3; 0{,}8$ e $0{,}6$ para $\gamma_1, \gamma_2, \beta_0, \beta_1$ e β_2, respectivamente.

A fim de examinar as propriedades frequentistas, construímos intervalos de confiança assintóticos (derivados sob o pressuposto da normalidade dos estimadores) de 95% para os parâmetros do modelo e calculamos a probabilidade de cobertura (PC) deles. A média das 500

estimativas de máxima verossimilhança, os respectivos desvios padrão, a raiz do erro quadrático médio (REQM) e as probabilidades de cobertura para os diferentes tamanhos de amostras estão listados na Tabela 13 e os histogramas e os gráficos QQ-normal dos parâmetros estimados são apresentados no Apêndice C.

Pela Tabela 13 podemos notar que os resultados das estimativas são bastante próximos dos valores verdadeiros dos parâmetros, o que nos revela que os valores iniciais do algoritmo EM oferecem uma precisão bastante alta. Este também é o caso da grade de pesquisa para η quando a verossimilhança perfilada é realizada. Destacamos ainda que os valores do REQM e DP decrescem conforme o tamanho da amostra aumenta. A probabilidade de cobertura dos intervalos de confiança assintóticos é suficientemente próxima ao nível nominal (95%) e essa aproximação melhora à medida que o tamanho da amostra aumenta.

Tabela 13 – Médias das estimativas, DP, REQM e a PC dos parâmetros do modelo com fração de cura hP com η estimado por meio da verossimilhança perfilada.

		Parâmetros					
		η	γ_1	γ_2	β_0	β_1	β_2
	Verdadeiro	6	1,5	2,0	0,5	1	0,8
n =200	Média	6,016	1,517	2,025	0,480	1,019	0,820
	DP	0,495	0,113	0,226	0,217	0,156	0,242
	REQM	0,495	0,114	0,227	0,217	0,157	0,243
	PC	–	0,958	0,952	0,940	0,960	0,946
$n = 400$	Média	6,006	1,511	2,0139	0,497	1,011	0,798
	DP	0,492	0,081	0,155	0,172	0,119	0,168
	REQM	0,491	0,082	0,155	0,171	0,119	0,168
	PC	–	0,944	0,956	0,942	0,950	0,944
$n = 600$	Média	5,972	1,508	2,005	0,494	1,012	0,807
	DP	0,487	0,067	0,133	0,146	0,098	0,147
	REQM	0,488	0,069	0,133	0,147	0,098	0,147
	PC	–	0,948	0,948	0,946	0,948	0,947

Fonte: Elaborada pelo autor.

No segundo estudo de simulação, mantivemos o parâmetro de dispersão η da distribuição hP fixado nos valores $0,5$, 1 e 2, representando assim os casos de subdispersão, Poisson e sobredispersão, nessa ordem. Tomamos o outro parâmetro da distribuição de fragilidade hP $\theta = \exp(\beta_0 + \beta_1 x_1)$, com $\beta_0 = -0,5$ e $\beta_1 = 1$, em que a covariável x_i foi gerada de uma distribuição Bernoulli com parâmetro 0,5. Utilizamos novamente a distribuição de Weibull como a distribuição basal com parâmetros $\gamma_1 = 2$ e $\gamma_2 = 2$. Foram gerados 1000 conjuntos de dados com tamanhos $n = 200, 400$ e 600 cada um. Os valores iniciais para o algoritmo EM foram $1,6; 1,6; -0,2$ e $0,8$ para $\gamma_1, \gamma_2, \beta_0$ e β_1, respectivamente. Mais uma vez os tempos de censura foram amostrados de uma distribuição uniforme, mas a proporção de observações censuradas neste caso é de aproximadamente 38%.

De modo similar ao primeiro estudo, foram calculados a média das 1000 estimativas dos parâmetros, os respectivos desvio padrão, o viés e a raiz do erro quadrático médio. Além disso, os intervalos de confiança foram construídos (95%). Todas essas quantidades estão exibidas na Tabela 14, em que notamos mais uma vez que as médias das estimativas são bastante próximas dos valores verdadeiros dos parâmetros, o que pode ser confirmado pelos vieses, que assumem valores bem pequenos e que diminuem conforme o tamanho da amostra aumenta. Novamente, como esperado, os valores dos desvios e as RQME são bem próximos e ambas diminuem conforme o tamanho da amostra aumenta. Por fim, no que diz respeito às probabilidades de cobertura, estas estão suficientemente próximas do nível nominal em todos os cenários. Os histogramas e os gráficos QQ-normal dos parâmetros são apresentados no Apêndice C.

Interessados em analisar o comportamento dos testes de hipóteses da razão de verossimilhanças e gradiente, ambos estudados no Capítulo 2, no modelo de sobrevivência induzido por fragilidade proposto, realizamos mais um estudo de simulação, agora interessados em testar as hipóteses $H_0 : \beta_1 = 0$ contra $H_a : \beta_1 \neq 0$. A fim de avaliar as taxas de rejeição de H_0 dos dois testes e compará-los entre si, construímos inicialmente 5.000 amostras sob H_0 de tamanhos $n = 200, 400$ e 600. Adotamos como níveis de significância nominais $\alpha = 1\%, 5\%$ e 10% e mantemos o parâmetro de dispersão η fixado em três diferentes valores, $\eta = 0,5; 1$ e 2. A Tabela 15 traz os resultados desse estudo, em que observamos que os dois testes tiveram taxas de rejeição bem próximas aos níveis nominais adotados em todos os cenários construídos.

Apresentamos na Figura 9 os gráficos quantil-quantil com as distribuições simuladas dos testes RV e gradiente sob a hipótese nula para o modelo com fração de cura hP com $\eta = 0,5; 1$ e 2 e com amostras de tamanhos $n = 200$ e 400. Com os gráficos vemos, como esperado, que a distribuição qui-quadrado com um grau de liberdade fornece uma boa aproximação para a distribuição sob H_0 dos dois teses.

Encerramos os estudos de simulação analisando o poder dos testes da razão de verossimilhanças e gradiente ao detectar a hipótese alternativa $H_a : \beta_1 = \delta$. Geramos 5.000 amostras com tamanho $n = 400$ e parâmetro de dispersão fixado em $\eta = 2$, sob a hipótese alternativa para diferentes valores de δ, e então calculamos as taxas de rejeição. Adotamos como nível de significância nominal $\alpha = 0,05$. A figura 10 mostra o poder dos testes RV e gradiente, cuja curvas se sobrepõem, indicando que ambos têm o mesmo comportamento.

3.5 Aplicação a dados reais

Ilustramos nesta seção uma aplicação do modelo proposto para o conjunto de dados da fase III do ensaio clínico do melanoma cutâneo conduzido pelo Eastern Cooperative Oncology Group (KIRKWOOD *et al.*, 2000). Esses dados compõem um ensaio para a avaliação da performance de um tratamento pós operatório com alta dose de uma certa droga (*interferon alpha - 2b*) com o objetivo de prevenir a recorrência do câncer. Os pacientes foram incluídos no estudo entre

Tabela 14 – Médias das estimativas, viés, DP, REQM e a PC dos parâmetros do modelo de fração de cura hP com η fixado.

n	Parâmetro	Média	Viés	DP	REQM	PC
		$\eta = 0,5$				
200	γ_1	2,025	0,025	0,151	0,153	0,945
	γ_2	2,007	0,007	0,157	0,157	0,957
	β_0	-0,489	0,011	0,184	0,184	0,946
	β_1	1,009	0,009	0,227	0,227	0,947
400	γ_1	2,014	0,014	0,101	0,102	0,954
	γ_2	2,002	0,002	0,108	0,108	0,949
	β_0	-0,504	-0,004	0,124	0,124	0,947
	β_1	1,011	0,011	0,154	0,154	0,958
600	γ_1	2,009	0,009	0,082	0,083	0,953
	γ_2	2,008	0,008	0,090	0,090	0,952
	β_0	-0,493	0,007	0,104	0,104	0,952
	β_1	1,002	0,002	0,128	0,128	0,942
		$\eta = 1$				
200	γ_1	2,032	0,032	0,171	0,173	0,942
	γ_2	2,011	0,011	0,160	0,160	0,949
	β_0	-0,493	0,007	0,180	0,180	0,944
	β_1	1,003	0,003	0,180	0,218	0,953
400	γ_1	2,014	0,014	0,118	0,119	0,955
	γ_2	1,999	-0,001	0,111	0,111	0,953
	β_0	-0,506	-0,006	0,127	0,126	0,951
	β_1	1,001	0,001	0,148	0,148	0,955
600	γ_1	2,009	0,009	0,098	0,098	0,944
	γ_2	1,996	-0,004	0,089	0,089	0,947
	β_0	-0,506	-0,006	0,105	0,105	0,946
	β_1	1,007	0,007	0,126	0,126	0,951
		$\eta = 2$				
200	γ_1	2,035	0,035	0,205	0,208	0,955
	γ_2	2,005	0,005	0,179	0,179	0,948
	β_0	-0,514	0,014	0,225	0,225	0,946
	β_1	1,014	0,014	0,255	0,256	0,953
400	γ_1	2,023	0,023	0,143	0,144	0,957
	γ_2	2,004	0,004	0,124	0,124	0,946
	β_0	-0,509	0,009	0,159	0,159	0,947
	β_1	1,012	0,012	0,183	0,183	0,949
600	γ_1	2,014	0,014	0,115	0,116	0,959
	γ_2	2,005	0,005	0,100	0,103	0,959
	β_0	-0,504	0,004	0,121	0,121	0,952
	β_1	1,007	0,007	0,142	0,142	0,952

Fonte: Elaborada pelo autor.

Tabela 15 – Taxas de rejeição de $H_0 : \beta_1 = 0$ dos testes razão de verossimilhanças (RV) e gradiente (G) no modelo de sobrevivência induzido por fragilidade hP.

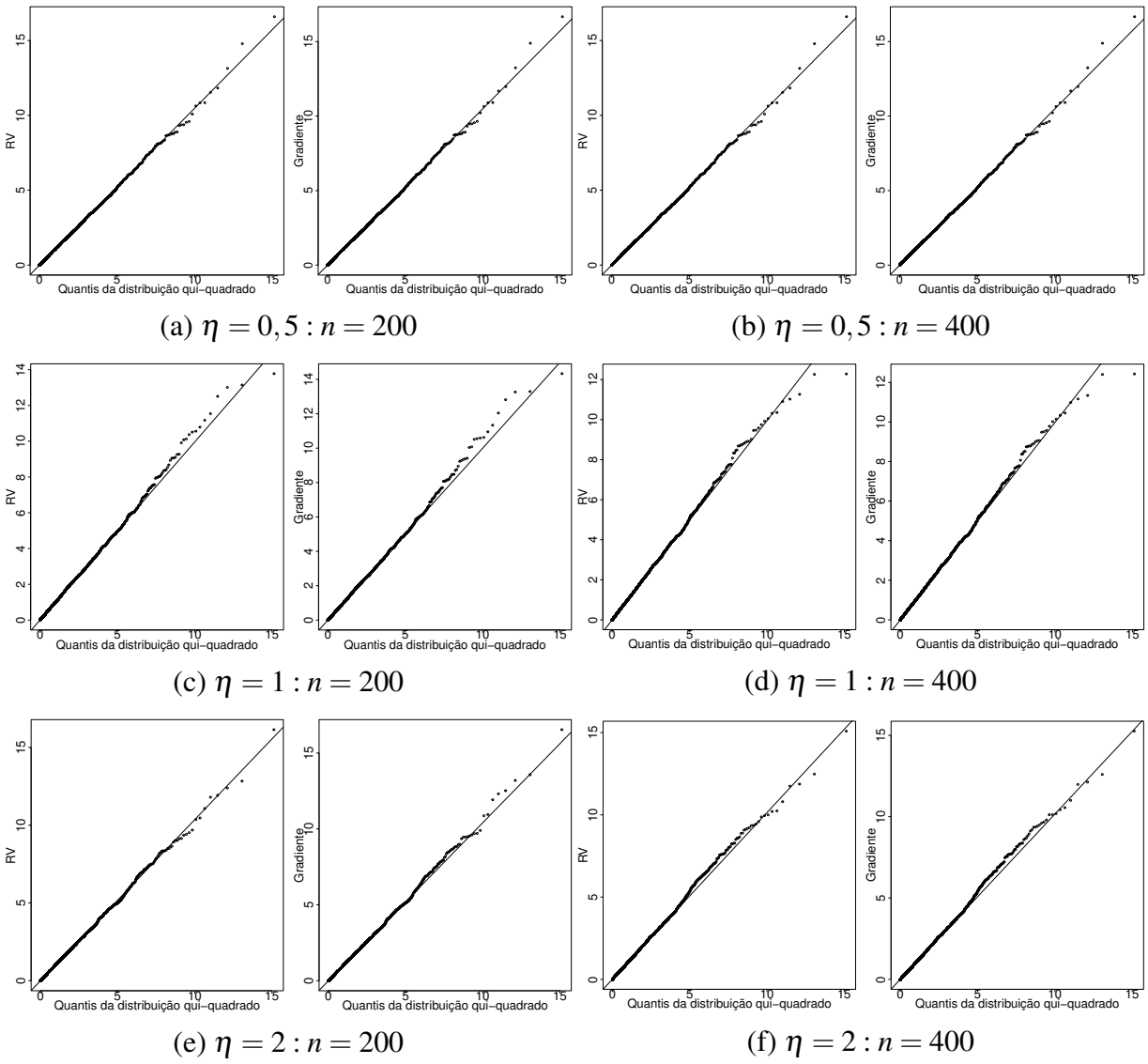
	Nível de Significância α					
	0,1		0,05		0,01	
n	RV	G	RV	G	RV	G
			$\eta = 0,5$			
200	0,106	0,106	0,053	0,053	0,011	0,012
400	0,100	0,100	0,049	0,049	0,009	0,009
600	0,102	0,100	0,048	0,048	0,010	0,010
n	RV	G	RV	G	RV	G
			$\eta = 1$			
200	0,097	0,098	0,049	0,050	0,011	0,011
400	0,104	0,104	0,051	0,051	0,011	0,011
600	0,095	0,095	0,047	0,047	0,011	0,014
n	RV	G	RV	G	RV	G
			$\eta = 2$			
200	0,104	0,106	0,054	0,054	0,012	0,013
400	0,105	0,106	0,052	0,053	0,013	0,013
600	0,106	0,106	0,053	0,054	0,010	0,011

Fonte: Elaborada pelo autor.

os anos de 1991 a 1995 e foram conduzidos até 1998. Os dados foram extraídos de Ibrahim, Chen e Sinha (2001b) (nomeado como dado E1690, disponível em http://merlot.stat.uconn.edu/ mh-chen/survbook/).

Após deletar os sujeitos com dados incompletos e tempos observados perdidos, temos um subconjunto de $n = 417$ pacientes com aproximadamente 56% de dados censurados. Os tempos observados têm média igual a 3,18 e desvio padrão igual a 1,69. Na sequência estão listadas as variáveis que foram associadas a cada participante, $i = 1, \ldots, 417$:

- t_i: tempo observado (em anos);
- x_{i1}: tratamento (0: observação, $n = 204$; 1: *interferon*, $n = 213$);
- x_{i2}: idade (0: ≥ 48 anos, $n = 197$; 1: < 48 anos, $n = 220$);
- x_{i3}: nódulo (categoria do nódulo é codificada em relação ao número de nódulos linfáticos envolvidos na doença: 1: $n = 111$; 2: $n = 137$; 3: $n = 87$; 4: $n = 82$) ;
- x_{i4}: sexo (0: masculino, $n = 263$; 1: feminino, $n = 154$);
- x_{i5}: s.p (status de performance - escala da capacidade funcional dos pacientes no que diz respeitos às suas atividades diárias: 0: atividade completa, $n = 363$; 1: outros, $n = 54$) e
- x_{i6}: tumor (espessura do tumor em décimo de milímetro).



(a) $\eta = 0,5 : n = 200$ (b) $\eta = 0,5 : n = 400$

(c) $\eta = 1 : n = 200$ (d) $\eta = 1 : n = 400$

(e) $\eta = 2 : n = 200$ (f) $\eta = 2 : n = 400$

Figura 9 – Gráficos Q-Q dos testes razão de verossimilhanças e gradiente no modelo com fração de cura hP contra a distribuição qui-quadrado com 1 grau de liberdade.

Fonte: Elaborada pelo autor.

Consideramos primeiramente o modelo com fração de cura hP com função de sobrevivência dada em (3.3) com todas as covariáveis regressoras, ou seja,

$$\theta_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}), \quad i = 1, \ldots, 417.$$

De modo análogo ao feito por Koutras e Milienos (2017), a fim de determinar a grade da área de pesquisa de η para usar a técnica de verossimilhança perfilada, realizamos uma investigação numérica da função de verossimilhança observada, isto é, a função de verossimilhança foi computada para várias escolhas de η do espaço paramétrico, de modo que a área que produz os maiores valores de verossimilhança seja detectada. Aplicando este procedimento, verificamos que a busca da grade por η pode ser executada no intervalo $[8,9]$ (passo =0,05). Vale ressaltar

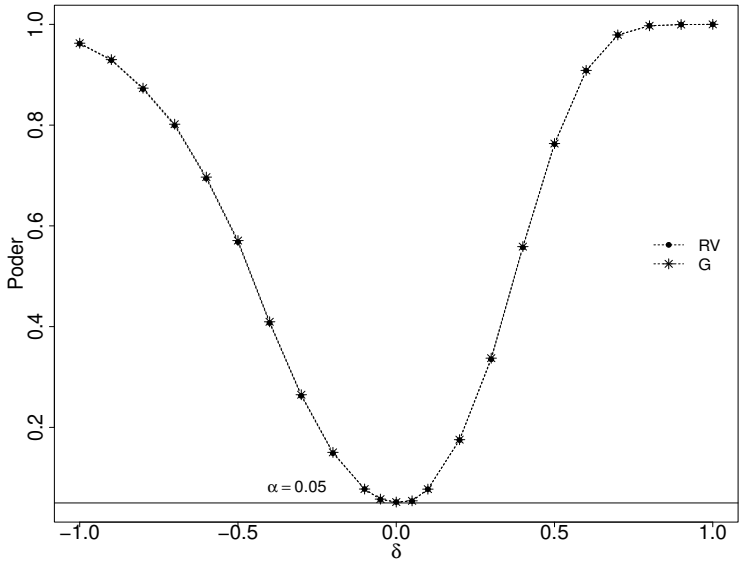


Figura 10 – Poder dos testes razão de verossimilhanças (RV) e gradiente (G) com $\alpha = 0,05$ no modelo de sobrevivência induzido por fragilidade discreta com $\eta = 2$ e $n = 400$.

Fonte: Elaborada pelo autor.

que várias outras escolhas foram também consideradas e os resultados foram bastante robustos. As estimativas obtidas juntamente com o critério AIC estão na Tabela 16.

Tabela 16 – Estimativa do modelo hP proposto para o conjunto de dados melanoma usando todas as covariáveis.

Parâmetro	Média	Erro padrão	AIC
γ_1	1,831	0,119	1036,559
γ_2	2,829	0,222	
$\beta_{\text{intercepto}}$	0,893	0,196	
$\beta_{\text{tratamento}}$	0,024	0,108	
β_{idade}	-0,050	0,108	
$\beta_{\text{nódulo}}$	0,285	0,050	
β_{sexo}	-0,147	0,113	
$\beta_{\text{capacidade}}$	0,109	0,144	
$\beta_{\text{espessura}}$	0,090	0,152	
η	8,850		

Fonte: Elaborada pelo autor.

A fim de determinar quais das covariáveis têm efeito significativo, utilizamos os testes da razão de verossimilhanças e gradiente para testar a seguinte hipótese: $H_0 : \beta_{\text{tratamento}} = \beta_{\text{idade}} = \beta_{\text{nódulo}} = \beta_{\text{sexo}} = \beta_{\text{capacidade}} = \beta_{\text{espessura}} = 0$ contra a hipótese alternativa H_a: Ao menos um coeficiente é diferente de 0. A Tabela 17 traz as estimativas das duas estatísticas e os respectivos valores- p, em que vemos que a hipótese nula é rejeita, ou seja, ao menos uma covariável

tem efeito significativo para o modelo. Dessa maneira, testamos os coeficientes um a um para determinar qual ou quais deles têm influência e os resultados estão dispostos na Tabela 18. Adotando um nível de significância de 5% vemos que apenas a covariável nódulo tem efeito significativo para o modelo.

Tabela 17 – Estimativas das estatísticas RV e gradiente e os respectivos valor-p ao testar a hipótese $H_0 : \beta_{\text{tratamento}} = \beta_{\text{idade}} = \beta_{\text{nódulo}} = \beta_{\text{sexo}} = \beta_{\text{capacidade}} = \beta_{\text{espessura}} = 0$ para o conjunto de dados melanoma

Estatísticas	Estimativas	Valor-p
RV	34,846	$< 0,0001$
Gradiente	32,865	$< 0,0001$

Fonte: Elaborada pelo autor.

Tabela 18 – Estimativas das estatísticas RV e gradiente e os respectivos valores-p ao testar as hipótese $H_0 : \beta_{\text{tratamento}} = 0$, $H_0 : \beta_{\text{idade}} = 0$, $H_0 : \beta_{\text{nódulo}} = 0$, $\beta_{\text{sexo}} = 0$, $\beta_{\text{capacidade}} = 0$ e $\beta_{\text{espessura}} = 0$ para o conjunto de dados melanoma.

	$H_0 : \beta_{\text{tratamento}} = 0$		$H_0 : \beta_{\text{idade}} = 0$	
Estatísticas	Estimativas	Valor-p	Estimativas	Valor-p
RV	0,0527	0,818	0,212	0,645
Gradiente	0,0525	0,819	0,213	0,6444
	$H_0 : \beta_{\text{nódulo}} = 0$		$H_0 : \beta_{\text{sexo}} = 0$	
Estatísticas	Estimativas	Valor-p	Estimativas	Valor-p
RV	31,446	$< 0,0001$	1,761	0,184
Gradiente	30,150	$< 0,0001$	1,784	0,182
	$H_0 : \beta_{\text{capacidade}} = 0$		$H_0 : \beta_{\text{espessura}} = 0$	
Estatísticas	Estimativas	Valor-p	Estimativas	Valor-p
RV	0,542	0,461	0,338	0,560
Gradiente	0,527	0,468	0,330	0,565

Fonte: Elaborada pelo autor.

Como apenas a covariável nódulo tem efeito significativo, conduzimos um estudo com o modelo de fração de cura hP simplificado, isto é, considerando apenas a categoria nódulo como covariável, que assume os valores 1, 2, 3 e 4 de acordo com a espessura do tumor. A Tabela 19 retrata as estimativas dos parâmetros, bem como os respectivos erro padrão e percentis, e traz ainda o critério AIC desse novo modelo ajustado. A Figura 11 (painel esquerdo) ilustra as curvas de Kaplan-Meier estratificadas pela categoria nódulo (1–4) e as estimativas das funções de sobrevivência, indicando que o modelo proposto fornece uma aproximação razoável para as estimativas de Kaplan-Meier da função de sobrevivência. A fim de avaliar a qualidade do ajuste do modelo, o painel direito da Figura 11 traz os resíduos normalizados ajustados para o modelo simplificado. Um bom ajuste é observado, uma vez que os pontos gerados seguem próximos

à diagonal, bem próximos da linearidade, não apresentando desvio significativo, confirmando assim que não há evidência de falta de ajuste do modelo proposto.

Tabela 19 – Estimativa do modelo hP proposto para o conjunto de dados melanoma usando apenas a categoria nódulo como covariável.

Parâmetro	Estimativa	Erro padrão	Intervalo de confiança 2,5%	Intervalo de confiança 97,5%
γ_1	1,827	0,347	1,147	2,508
γ_2	2,811	0,805	1,233	4,390
$\beta_{\text{intercepto}}$	0,881	0,145	0,596	1,165
$\beta_{\text{nódulo}}$	0,284	0,048	0,189	0,379
η	8,900			
AIC	1029,548			

Fonte: Elaborada pelo autor.

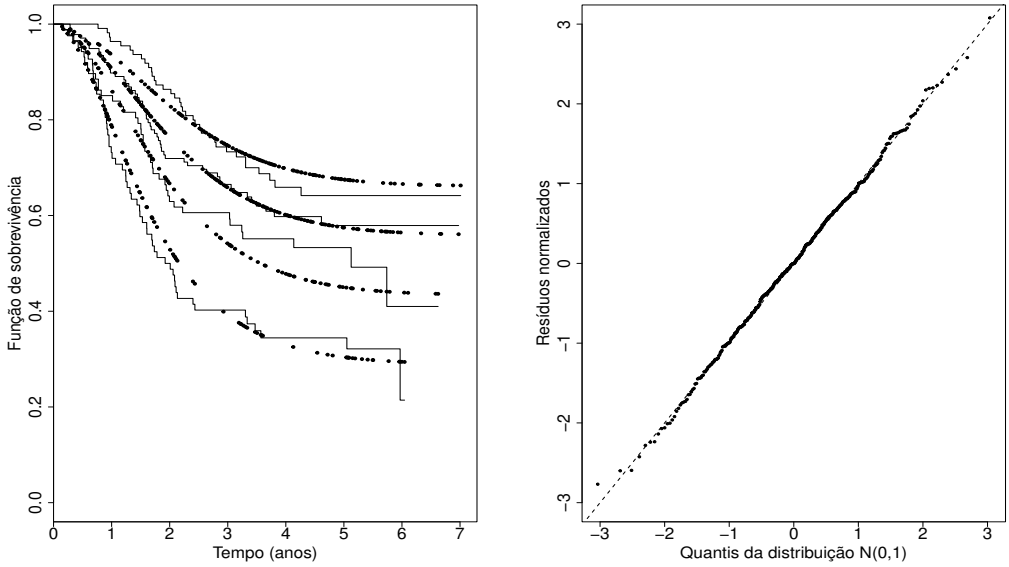


Figura 11 – Painel esquerdo: curvas de Kaplan-Meier estratificadas pela categoria nódulo (1–4) junto com as estimativas da função de sobrevivência hP. Painel direito: resíduos normalizados ajustados.

Fonte: Elaborada pelo autor.

Podemos ainda comparar o modelo de fração de cura hP ajustado aos dados melanoma na presença de todas as covariáveis com o modelo hP ajustado apenas com a categoria nódulo por meio dos critérios AIC presentes nas Tabelas 16 e 19, em que vemos claramente que o modelo simplificado fornece um melhor ajuste.

3.6 Alguns comentários

Neste capítulo propomos um modelo de sobrevivência induzido por fragilidade discreta para dados de sobrevivência univariados com fração de cura. Consideramos que a variável de fragilidade segue uma distribuição hiper-Poisson, que é muito útil para modelar dados de sobrevivência com fração de cura, uma vez que seu parâmetro adicional permite acomodar subdispersão e sobredispersão com respeito à distribuição de Poisson. Mostramos que o modelo é identificável.

O método inferencial abordado, baseado no algoritmo EM e combinado com uma abordagem de verossimilhança perfilada exibe uma alta acurácia. Os testes de hipóteses desenvolvidos para os coeficientes de regressão mostraram ótimas performances no que diz respeito ao controle da taxa de rejeição sob H_0 e em relação ao poder. Quando comparamos os testes da razão de verossimilhanças e gradiente neste modelo, vemos que apresentaram comportamentos similares em todos os cenários.

CAPÍTULO

4

UM MODELO DE SOBREVIVÊNCIA COM FRAGILIDADE HIPER-POISSON: ENFOQUE BAYESIANO

No Capítulo 3 foi apresenatdo o modelo de fração de cura induzido por fragilidade hiper-Poisson e então procedimentos inferenciais sob uma perspectiva clássica foram desenvolvidos. No presente capítulo trabalhamos com o mesmo modelo, porém sob o ponto de vista bayesiano. Critérios de comparação de modelos e medidas de diagnótico foram discutidos. Além disso, um estudado detalhado com dados simulados e uma aplicação a um conjunto de dados reais foram considerados.

4.1 Inferência bayesiana

De modo similar ao feito na Seção 3.3, vamos considerar novamente a situação em que o tempo de falha não é completamente observado e está sujeito à censura à direita. Seja C_i denotando o tempo de censura para o i-ésimo indivíduo e então a i-ésima observação é dada por $t_i = \min\{T_i, C_i\}$ com $\delta_i = \mathrm{I}(T_i \leq C_i)$, em que $\delta_i = 1$ se t_i é um tempo de falha e $\delta_i = 0$ se é uma observação censurada à direita, para $i = 1, \ldots, n$. Seja $\boldsymbol{\gamma}$ denotando o vetor de parâmetros da função de distribuição basal. A partir dos n pares de tempos e indicadores de censura, $(t_1, \delta_1), \ldots, (t_n, \delta_n)$, podemos construir a função de verossimilhança marginal sob censura não informativa dada pela seguinte expressão

$$L(\eta, \theta, \boldsymbol{\gamma}; \boldsymbol{t}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} f(t_i, \delta_i; \eta, \theta, \boldsymbol{\gamma}), \tag{4.1}$$

em que $\boldsymbol{t} = (t_1, \ldots, t_n)$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ e

$$f(t_i, \delta_i; \eta, \theta, \boldsymbol{\gamma}) = \sum_{z_i=0}^{\infty} \{S_B(t_i; \boldsymbol{\gamma})\}^{z_i - \delta_i} \{z_i f_B(t_i; \boldsymbol{\gamma})\}^{\delta_i} p(z_i; \eta, \theta).$$

A função de verossimilhança em (4.1) pode ser escrita como

$$L(\eta,\theta,\boldsymbol{\gamma};\boldsymbol{t},\boldsymbol{\delta}) \propto \prod_{i=1}^{n}\{f(t_i;\eta,\theta,\boldsymbol{\gamma})\}^{\delta_i}\{S(t_i;\eta,\theta,\boldsymbol{\gamma})\}^{1-\delta_i}. \tag{4.2}$$

Associamos novamente o parâmetro θ em (3.3) com as covariáveis $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^\top$ por meio da expressão $\log(\theta_i) = \boldsymbol{x}_i^\top\boldsymbol{\beta}$ ou $\theta_i = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$, em que $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$ denota o vetor de coeficientes de regressão. Tomando $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\top,\boldsymbol{\gamma}^\top,\eta)^\top$, substituindo (3.3) e (3.5) em (4.2), obtemos a função de verossimilhança marginal como

$$\begin{aligned} L(\boldsymbol{\vartheta};\mathscr{D}) \propto \prod_{i=1}^{n}\Bigg\{ & \frac{{}_1F_1(2,\eta+1,\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})S_B(t_i;\boldsymbol{\gamma}))\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{{}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})S_B(t_i;\boldsymbol{\gamma}))} \\ & \times\frac{f_B(t_i;\boldsymbol{\gamma})}{\eta}\Bigg\}^{\delta_i}\frac{{}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})S_B(t_i;\boldsymbol{\gamma}))}{{}_1F_1(1;\eta;\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}))}, \end{aligned} \tag{4.3}$$

em que $\mathscr{D} = (\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{x})$, $\boldsymbol{t} = (t_1,\ldots,t_n)^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^\top$ e $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_n)^\top$ tal que $S_B(.)$ e $f_B(.)$ são a função de sobrevivência basal e a função de distribuição basal, respectivamente. Vamos assumir de agora em diante uma distribuição Weibull para a função de distribuição basal em (4.3), com $S_B(t;\boldsymbol{\gamma}) = \exp\{(-t/\gamma_2)\}^{\gamma_1}$ e $f_B(t;\boldsymbol{\gamma}) = \dfrac{\gamma_1}{\gamma_2^{\gamma_1}}t^{\gamma_1-1}\exp\left\{-\left(\dfrac{t}{\gamma_2}\right)^{\gamma_1}\right\}$ para $t>0$ e $\gamma_1,\gamma_2>0$.

4.1.1 Distribuições a priori e a posteriori

Vamos agora analisar algumas ferramentas inferenciais sob o ponto de vista bayesiano. Assumimos $\boldsymbol{\beta}$, γ_1, γ_2 e η como *priori* independentes, ou seja,

$$\pi(\boldsymbol{\vartheta}) = \pi(\boldsymbol{\beta})\pi(\gamma_1)\pi(\gamma_2)\pi(\eta), \tag{4.4}$$

em que $\boldsymbol{\beta} \sim N_p(\mathbf{0},\Sigma)$, $\gamma_1 \sim \mathrm{G}(a_{\gamma 1},b_{\gamma_1})$, $\gamma_2 \sim N(0,\sigma^2_{\gamma_2})$, com $N(a,b)$ denotando a distribuição normal e $G(a,b)$ representando a distribuição gama com a como parâmetro de forma, b parâmetro de escala e média a/b.

A fim de se obter estabilidade numérica e motivados pelo trabalho de Yin e Ibrahim (2005), consideramos uma distribuição *a priori* uniforme discreta para η. Tomamos $\eta \in \mathbb{A} = \{\alpha_1,\alpha_2\ldots,\alpha_K\}$ com probabilidade $1/K$, $K \geq 1$. Adotamos uma grade $\mathbb{A}$ de valores para η, assim não ficamos restritos a um único modelo. A distribuição a *posteriori* de η transmite informação sobre a contribuição dos elementos na grade escolhida para o modelo ajustado (RODRIGUES *et al.*, 2016).

Combinando a função de verossimilhança (4.3) com a distribuição *a priori* em (4.4), a distribuição *a posteriori* conjunta para $\boldsymbol{\vartheta}$ é obtida por meio da expressão

$$\pi(\boldsymbol{\vartheta}|\mathscr{D}) \propto L(\boldsymbol{\vartheta};\mathscr{D})\pi(\boldsymbol{\beta})\pi(\gamma_1)\pi(\gamma_2)\pi(\eta). \tag{4.5}$$

A densidade *a posteriori* conjunta em (4.5) é analiticamente intratável, uma vez que a integração da densidade *a posteriori* conjunta não é fácil de ser obtida. Dessa forma, as

inferências são desenvolvidas baseadas nos métodos de simulação de Monte Carlo via cadeias de Markov (MCMC), tais como os algoritmos do amostrador de Gibbs e Metropolis-Hastings (GAMERMAN; LOPES, 2006). Seguindo nesta direção, escrevemos as distribuições condicionais completas de $\boldsymbol{\beta}$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ e η como $\pi(\boldsymbol{\beta}|\gamma_1, \gamma_2, \eta, \mathscr{D}) \propto L(\boldsymbol{\vartheta}; \mathscr{D})\pi(\boldsymbol{\beta})$, $\pi(\gamma_1|\boldsymbol{\beta}, \gamma_2, \eta, \mathscr{D}) \propto L(\boldsymbol{\vartheta}; \mathscr{D})\pi(\gamma_1)$, $\pi(\gamma_2|\boldsymbol{\beta}, \gamma_1, \eta, \mathscr{D}) \propto L(\boldsymbol{\vartheta}; \mathscr{D})\pi(\gamma_2)$, e

$$\pi(\eta|\boldsymbol{\beta}, \gamma_1, \gamma_2, \mathscr{D}) = \frac{L(\boldsymbol{\beta}, \gamma_1, \gamma_2, \eta; \mathscr{D})}{\sum_{\eta^* \in \mathbb{A}} L(\boldsymbol{\beta}, \gamma_1, \gamma_2, \eta^*; \mathscr{D})}, \ \eta \in \mathbb{A},$$

Aqui todos os hiperparâmetros são especificados. Então, o algoritmo de Metropolis- Hastings é utilizado para simular amostras de $\boldsymbol{\beta}$, γ_1,γ_2 e η.

4.1.2 Critérios de comparações de modelo

Existem várias metodologias para comparar diversos modelos concorrentes para um determinado conjunto de dados e então selecionar o melhor modelo que se ajusta aos dados. Um dos mais utilizados na abordagem bayesiana aplicada é o *Deviance Information Criterion* (DIC), proposto por Spiegelhalter *et al.* (2002). Este critério é baseado na média a *posteriori* da *deviance*, $D(\boldsymbol{\vartheta})$, que pode ser aproximado a partir dos resultados do MCMC por

$$\overline{D} = \sum_{q=1}^{Q} D(\boldsymbol{\vartheta}_q)/Q,$$

em que o índice q indica a q-ésima realização de um total de Q realizações e

$$D(\boldsymbol{\vartheta}) = -2\sum_{i=1}^{n} \log\left[g(t_i; \boldsymbol{\vartheta})\right],$$

com $g(\cdot)$ representando a função densidade de probabilidade correspondente ao modelo. Para os dados observados, temos que $g(t_i; \boldsymbol{\vartheta})$ é o i-ésimo componente da função de verossimilhança. O critério DIC pode ser estimado considerando amostras do MCMC, por $\widehat{DIC} = 2\overline{D} - \widehat{D}$, no qual

$$\widehat{D} = D\left(\frac{1}{Q}\sum_{q=1}^{Q} \boldsymbol{\beta}^{(q)}, \frac{1}{Q}\sum_{q=1}^{Q} \boldsymbol{\gamma}^{(q)}, \frac{1}{Q}\sum_{q=1}^{Q} \eta^{(q)}\right).$$

O modelo que melhor se ajusta aos dados especificados é aquele corresponde ao menor valor do DIC.

Outro critério, que também é bastante popular, é derivado da ordenada preditiva condicional (CPO). Para uma detalhada discussão sobre a CPO e suas aplicações para seleção de modelos, sugerimos os trabalhos de Gelfand, Dey e Chang (1992) e Geisser e Eddy (1979). Seja $\mathscr{D}$ denotando os dados completos e $\mathscr{D}^{(-i)}$ a i-ésima observação deletada. Representamos a densidade a *posteriori* de $\boldsymbol{\vartheta}$, dado $\mathscr{D}^{(-i)}$, por $\pi(\boldsymbol{\vartheta}|\mathscr{D}^{(-i)})$, $i = 1, \ldots, n$. Para a i-ésima observação, a CPO_i pode ser escrita como

$$CPO_i = \int_{\vartheta \in \Theta} g(t_i; \boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\mathscr{D}^{(-i)})d\boldsymbol{\vartheta} = \left\{\int_{\vartheta} \frac{\pi(\boldsymbol{\vartheta}|\mathscr{D})}{g(t_i; \boldsymbol{\vartheta})}d\boldsymbol{\vartheta}\right\}^{-1}. \tag{4.6}$$

Para o modelo proposto, uma forma fechada da expressão CPO não é disponível. No entanto, uma estimativa de Monte Carlo da CPO$_i$ pode ser obtida usando uma única amostra MCMC da distribuição a *posteriori* $\pi(\boldsymbol{\vartheta}|\mathscr{D})$. Para esse propósito, seja $\boldsymbol{\vartheta}^{(1)},\ldots,\boldsymbol{\vartheta}^{(Q)}$ uma amostra de tamanho Q obtida de $\pi(\boldsymbol{\vartheta}|\mathscr{D})$ depois dos processos de *burn in*. Uma aproximação de Monte Carlo da CPO_i (IBRAHIM; CHEN; SINHA, 2001b) é então dada por

$$\widehat{CPO}_i = \left\{ \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{g(t_i;\boldsymbol{\vartheta}^{(q)})} \right\}^{-1}.$$

Para a comparação de modelos, usamos a *log pseudo marginal likelihood* (LPML) definida por $\text{LPML} = \sum_{i=1}^{n} \log(\widehat{CPO}_i)$. Dessa maneira, o maior valor do LPLM indica o melhor modelo ajustado.

4.1.3 Análise de diagnóstico

Usualmente em modelos de regressão, após a modelagem, um estudo de sensibilidade é realizada com o intuito de encontrar possíveis observações influentes que podem causar distorções nos resultados da análise. Uma das propostas mais inovadoras em análise de sensibilidade foi apresentada por Cook (1986), que sugeriu que se pode depositar maior confiança em um modelo que é relativamente estável sob pequenas modificações. Um dos melhores esquemas de pertubação conhecidos é baseado na exclusão de caso (COOK; WEISBERG, 1982), em que os efeitos são estudados removendo completamente os casos da análise. Essa ideia é essencial para nossa metodologia de influência global a fim de determinar quais observações podem ser altamente influentes.

Seja $D_\psi(P,P_{(-i)})$ a divergência ψ entre P e $P_{(-i)}$, em que P denota a distribuição a *posteriori* de $\boldsymbol{\vartheta}$ para o conjunto de dados completo, e $P_{(-i)}$ a distribuição a *posteriori* de $\boldsymbol{\vartheta}$ sem a $i-$ésima oboservação. Especificamente,

$$D_\psi(P,P_{(-i)}) = \int \psi\left(\frac{\pi(\boldsymbol{\vartheta}|\mathscr{D}^{(-i)})}{\pi(\boldsymbol{\vartheta}|\mathscr{D})}\right)\pi(\boldsymbol{\vartheta}|\mathscr{D})\,d\boldsymbol{\vartheta},$$

em que ψ é uma função convexa com $\psi(1)=0$. Escolhas diferentes de ψ são dadas em Dey e Birmiwal (1994) e Pardo (2006). Por exemplo, $\psi(z)=-\log(z)$ define a divergência Kullback-Leibler (K-L), $\psi(z)=(z-1)\log(z)$ retrata a distância J (ou a versão simétrica da divergência K-L), $\psi(z)=0,5|z-1|$ representa a distância variacional ou norma L_1 e $\psi(z)=z(1/z-1)^2$ define a divergência χ^2-quadrada.

A relação entre a CP0 em (4.6) e a medida de divergência ψ é dada por

$$D_\psi(P,P_{(-i)}) = E_{\boldsymbol{\vartheta}|\mathscr{D}}\left[\psi\left(\frac{CPO_i}{g(y_i;\boldsymbol{\vartheta})}\right)\right], \tag{4.7}$$

em que o valor esperado é tomado com respeito a distribuição a *posteriori* conjunta $\pi(\boldsymbol{\vartheta};|\mathscr{D})$.

Em particular, a divergência K-L pode ser expressa como

$$\begin{aligned} D_{\text{K-L}}(P, P_{(-i)}) &= -E_{\boldsymbol{\vartheta}|\mathscr{D}}\{\log(CPO_i)\} + E_{\boldsymbol{\vartheta}|\mathscr{D}}\{\log[g(y_i;\boldsymbol{\vartheta})]\} \\ &= -\log(CPO_i) + E_{\boldsymbol{\vartheta}|\mathscr{D}}\{\log[g(y_i;\boldsymbol{\vartheta})]\}. \end{aligned}$$

Da expressão (4.7), podemos calcular $D_\psi(P, P_{(-i)})$ pela amostragem da distribuição a *posteriori* de $\boldsymbol{\vartheta}$ via métodos MCMC. Seja $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(Q)}$ uma amostra de tamanho Q de $\pi(\boldsymbol{\vartheta}|\mathscr{D})$. Então, uma estimativa de Monte Carlo de $K(P, P_{(-i)})$ é dada por

$$\widehat{D_\psi}(P, P_{(-i)}) = \frac{1}{Q}\sum_{q=1}^{Q} \psi\left(\frac{\widehat{CPO_i}}{g(y_i;\boldsymbol{\vartheta}^{(q)})}\right). \tag{4.8}$$

Em particular, de (4.8), uma estimativa de Monte Carlo da divergência K-L $D_{K-L}(P, P_{(-i)})$ é obtida como

$$\widehat{D_{\text{K-L}}}(P, P_{(-i)}) = -\log(\widehat{CPO_i}) + \frac{1}{Q}\sum_{q=1}^{Q} \log\left[g(y_i;\boldsymbol{\vartheta}^{(q)})\right].$$

Valores de $D_\psi(P, P_{(-i)})$ podem ser interpretados como a divergência ψ do efeito da exclusão do i-ésimo caso dos dados completos na distribuição a *posteriori* conjunta de $\boldsymbol{\vartheta}$. Como apontado por Peng e Dey (1995) e Weiss (1996) (ver também Cancho, Ortega e Paula (2010), Cancho *et al.* (2011)) pode ser difícil para um profissional avaliar o ponto de corte da medida de divergência para determinar se um pequeno subconjunto de observações dos dados completos é influente ou não. Neste contexto, utilizamos a proposta de calibração dada por Peng e Dey (1995) e Weiss (1996), considerando uma moeda viciada com probabilidade de sucesso p. A divergência ψ entre as moedas viciadas e não viciadas é

$$D_\psi(f_0, f_1) = \int \psi\left(\frac{f_0(x)}{f_1(x)}\right) f_1(x)dx,$$

em que $f_0(x) = p^x(1-p)^{1-x}$ e $f_1(x) = 0{,}5$, $x = 0; 1$. Se $D_\psi(f_0, f_1) = d_\psi(p)$, pode ser facilmente verificado que d_ψ satisfaz a seguinte equação

$$d_\psi(p) = \frac{\psi(2p) + \psi(2(1-p))}{2}.$$

Não é difícil ver que d_ψ aumenta conforme p se afasta de 0,5. Além disso, $d_\psi(p)$ é simétrico em $p = 0{,}5$ e d_ψ atinge seu mínimo com $p = 0{,}5$. Neste ponto, $d_\psi(0{,}5) = 0$ e $f_0 = f_1$. Se considerarmos $p > 0{,}80$ (ou $p < 0{,}20$) como um alto vício em uma moeda, então $d_{L_1}(0{,}80) = 0{,}30$. Esta equação implica que o $i-$ésimo caso pode ser considerado influente quando $d_{L_1} > 0.30$. Então, se utilizarmos a divergência K-L podemos considerar uma observação influente quando $d_{K-L} > 0{,}223$. De modo análogo, se usarmos a distância J e a divergência χ^2-quadrada, uma observação pode ser tomada como influente quando $d_J > 0{,}416$ e $d_{\chi^2} > 0{,}562$, respectivamente.

4.2 Estudo de Simulação

Um estudo de simulação foi realizado com dois propósitos principais, o primeiro é avaliar a performance das estimativas bayesianas dos parâmetros do modelo de fração de cura hP e verificar se podemos distinguir entre os modelos de fração de cura hP e o modelo de cura tempo de promoção baseados nos critérios descritos na Seção 4.1; o segundo objetivo é analisar as medidas de diagnóstico.

Nesse estudo, consideramos a distribuição de Weibull como a distribuição basal com parâmetros $\gamma_1 = 2$ e $\gamma_2 = 2$. Para cada indivíduo i, $i = 1,\ldots,n$, a variável de fragilidade foi gerada de uma distribuição hP com parâmetros $\eta = 2$, $\theta_i = \exp(\beta_0 + \beta_1 x_i)$, em que $\beta_0 = -0,5$, $\beta_1 = 1$ e as covariáveis x_i foram geradas de uma distribuição Bernoulli com parâmetro $0,5$. Os tempos de censura C_i foram amostrados de uma distribuição uniforme no intervalo $(0,\tau)$, em que τ é o conjunto que controla a proporção de observações censuradas, que neste caso é aproximadamente igual a 55%. O estudo de simulação foi conduzido primeiramente com o intuito de examinar o comportamento das estimativas bayesianas baseadas na média frequentista da raiz do erro quadrático médio, a média frequentista, o desvio padrão e a probabilidade de cobertura. Consideramos as seguintes distribuições a *priori* independentes para a performance do algorítimo de Metropolis-Hasting: $\beta_k \sim N(0,10^2)$, $k = 0,1$, $\gamma_l \sim G(1,0.01)$,$\ell = 1,2$ e $\phi(\eta) = 1/10$, para $\eta \in \{0,5;1;1,5;2;\ldots,10\}$. Os tamanhos de amostras foram tomados como $n = 200,400$ e 600. Para cada conjunto de dados gerados foram obtidos o sumário a *posteriori* e o intervalo HPD (*highest posterior density*) com 95% de credibilidade dos parâmetros do modelo.

Geramos amostras de Gibbs através do algoritmo de Gibbs com Metropolis-Hasting da seguinte maneira: consideramos uma cadeia com 15.000 iterações, das quais 5.000 foram eliminadas a fim de eliminar os efeitos dos valores iniciais, obtendo assim uma amostra de tamanho 10.000. A autocorrelação desses valores amostrais foram reduzidos tomando um espaçamento de tamanho 10, resultando então em uma amostra final de tamanho 1.000. Para cada configuração, conduzimos 500 replicações e então determinamos as estimativas de cada parâmetro, a média, o desvio padrão (DP), a raiz do erro quadrático médio (REQM) e a probabilidade de cobertura (PC). Os resultados obtidos estão organizados na Tabela 20, dos quais observamos que o REQM e o DP decrescem à medida que o tamanho da amostra aumenta. Além disso, notamos que a diferença entre a média dos parâmetros estimados e os verdadeiros valores dos parâmetro é muito pequena, indicando que as estimativas possuem muito pouco viés. Observamos ainda que probabilidade de cobertura é muito próxima do nível nominal especificado para os parâmetros γ_1 e γ_2 quando o tamanho da amostra é 400. Para o parâmetro η, a PC assumiu valor igual a 1 em todos os casos, pois a probabilidade a *posteriori* de η é altamente concentrada em $\eta = 2$, como pode ser visto na Figura 12.

A Tabela 21 apresenta a porcentagem das amostras em que a distribuição das quais os dados foram gerados foram selecionados como o melhor de acordo com os critérios DIC e LPML (ambos medidas quase sempre concordam entre si). Assim, podemos observar claramente que

Tabela 20 – Médias, desvio padrão (DP), raiz do erro quadrático médio (REQM) e probabilidade de cobertura (PC) a *p*osteriori para os parâmetros do modelo de fração de cura hP.

		Parâmetro				
n		η	γ_1	γ_2	β_0	β_1
200	Média	2,051	1,986	2,097	-0,534	1,037
	DP	0,018	0,205	0,271	0,238	0,268
	REQM	0,054	0,205	0,287	0,240	0,270
	PC	1,000	0,942	0,948	0,930	0,936
400	Média	2,030	1,995	2,030	-0,516	1,020
	DP	0,011	0,139	0,134	0,155	0,179
	REQM	0,032	0,139	0,137	0,156	0,180
	PC	1,000	0,950	0,952	0,936	0,936
600	Média	2,023	2,004	2,030	-0,512	1,013
	DP	0,008	0,105	0,105	0,131	0,142
	REQM	0,024	0,105	0,109	0,132	0,142
	PC	1,000	0,962	0,938	0,932	0,956

Fonte: Elaborada pelo autor.

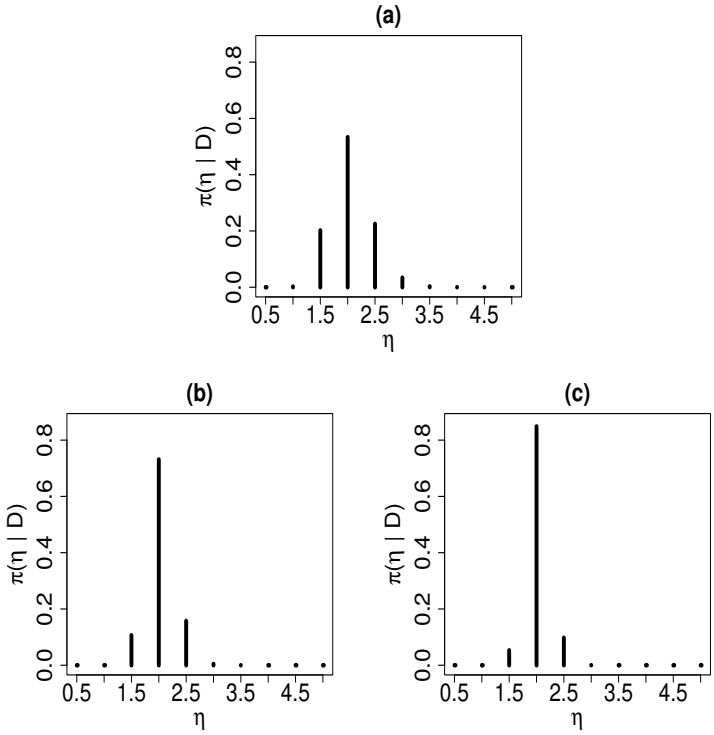


Figura 12 – Probabilidade a *posteriori* de η para (a) n = 200, (b) n = 400 e (c) n = 600.

Fonte: Elaborada pelo autor.

o verdadeiro modelo do qual a amostra foi gerada tem maior taxa de seleção, e essa taxa vai aumentando conforme o tamanho da amostra cresce.

Tabela 21 – Porcentagens das amostras em que o modelo ajustado foi indicado como o melhor de acordo com os critérios DIC e LPLM.

	Modelo ajustado	
n	Fração de cura hP	Tempo de promoção
200	77,3	22,7
400	89,8	10,2
600	92,5	7,5

Fonte: Elaborada pelo autor.

A fim de examinar o desempenho das medidas de diagnósticos propostas, consideramos conjuntos de dados simulados com um ou mais casos perturbados gerados. Para produzir isto, tomamos uma amostra de tamanho 200 gerada pelo modelo de cura hP. Nos dados simulados, y_i variam de 0,07824 a 7,870 com mediana igual a 2,27, média igual a 4,2 e desvio padrão igual a 3,5. Foram selecionados os casos 41 e 188 para pertubação. Para criar observações influentes no conjunto de dados, escolhemos um entre os dois casos selecionados e a variável resposta foi perturbada como $\tilde{y}_i = y_i + 3S_y$, $i = 41$ e $i = 188$, em que S_y é o desvio padrão de y_i. Os estudos computacionais MCMC foram feitos de maneira similar aos detalhados no início desta seção e para monitorar o convergência das amostras de Gibbs, os métodos recomendados por Cowles e Carlin (1996) foram implementados.

A Tabela 22 revela que a inferência a *posteriori* dos parâmetros do modelo de fração de cura hP são sensíveis à pertubação dos casos selecionados, com a exceção do parâmetro η. Na Tabela 22, o conjunto de dados (a) denota o conjunto de dados originais simulados sem nenhuma pertubação e os conjunto de dados (b)-(d) representam os casos com pertubação.

Tabela 22 – Média e desvio padrão (DP) dos parâmetros para cada conjunto de dados para o modelo de fração de cura hP.

		η		γ_1		γ_2		β_0		β_1	
Conjunto de dados	Caso perturbado	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
a	Nenhum	2,058	0,393	2,049	0,231	2,020	0,184	-0,500	0,213	0,866	0,254
b	41	2,053	0,407	1,671	0,182	2,526	0,521	-0,410	0,242	0,855	0,264
c	188	2,081	0,409	1,612	0,176	2,653	0,576	-0,387	0,237	0,882	0,251
d	{41,188}	2,059	0,504	1,459	0,190	6,528	14,982	-0,151	0,542	0,866	0,252

Fonte: Elaborada pelo autor.

Para encerrar, consideramos as amostras das distribuições dos parâmetros do modelo de fração de cura hP para calcular as medidas de divergência ψ descritas na Seção 15. Os resultados expostos na Tabela 23 indicam mais uma vez que antes das pertubações (conjunto de dados

(a)), que todos os casos selecionados não são influenciados de acordo com todas as medidas de divergência ψ. No entanto, depois das pertubações (conjunto de dados (b)-(d)), as médias crescem substancialmente indicando que os casos perturbados são de fato influentes.

Tabela 23 – Medidas de divergência ψ para os dados simulados.

Conjunto de dados	Casos perturbados	$d_{\text{K-L}}$	d_{J}	d_{L_1}	d_{χ^2}
a	Nenhum	0,011	0,024	0,061	0,024
		0,029	0,050	0,089	0,050
b	41	2.177	4,767	0,733	19,323
c	188	1,429	2,070	0,497	8,319
d	41	0,383	0,863	0,362	0,962
	188	0,682	1,439	0,453	2,056

Fonte: Elaborada pelo autor.

Nas Figuras 13 e 14 estão retratadas as quatro medias de divergência ψ para os casos (a) e (d), respectivamente. Podemos ver nitidamente que todas as medidas tiveram um bom desempenho na identificação dos casos influentes ao fornecer grandes medidas de divergência em relação aos outros casos.

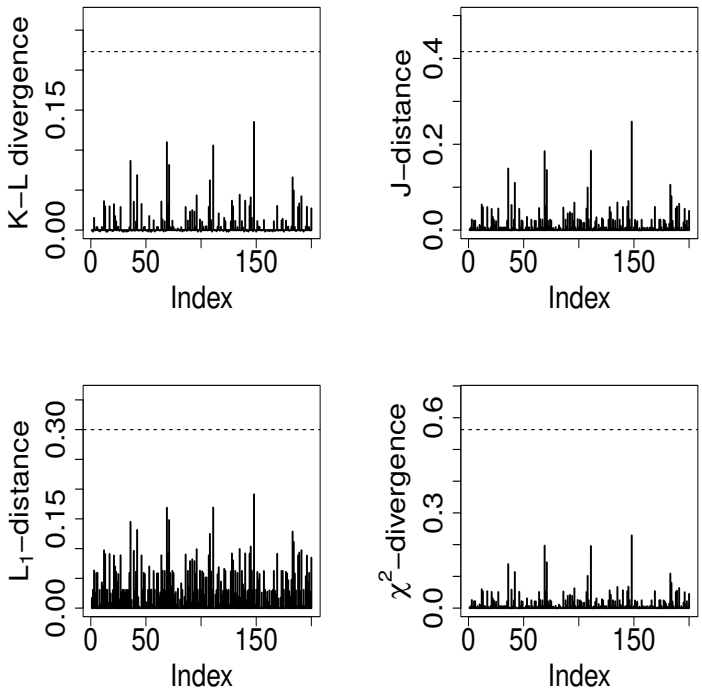


Figura 13 – Medidas de divergência ψ do conjunto de dados (a).

Fonte: Elaborada pelo autor.

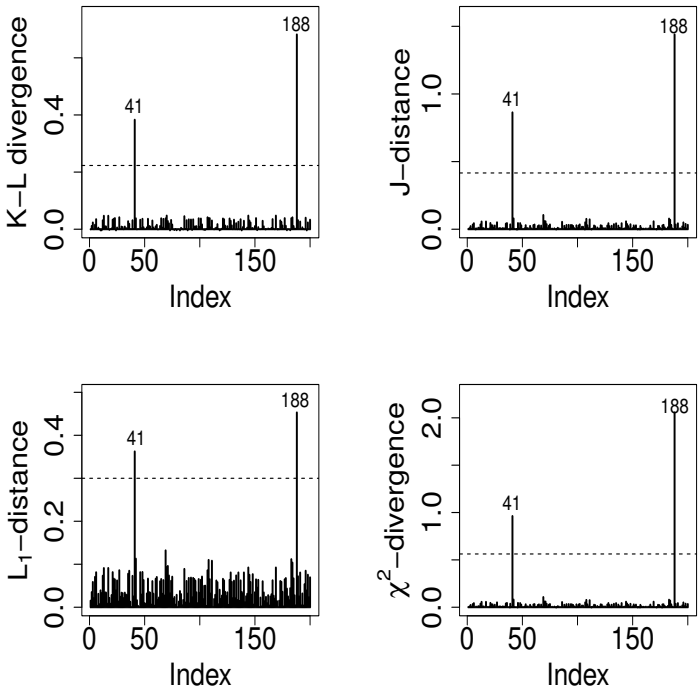


Figura 14 – Medidas de divergência ψ do conjunto de dados (d).

Fonte: Elaborada pelo autor.

4.3 Aplicação a dados reais

Ilustramos nesta seção uma aplicação do modelo proposto para o conjunto de dados do ensaio clínico do melanoma cutâneo já descrito na Seção 3.5.

Assim como feito anteriormente, consideramos primeiramente o modelo com fração de cura hP com função de sobrevivência dada em (3.3) com todas as covariáveis regressoras,

$$\theta_i = \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}\}, \quad i = 1, \ldots, 417.$$

Para a análise bayesiana consideramos as seguintes prioris independentes para a performance do algorítimo Metropolis-Hasting: $\beta_k \sim N(0, 10^2), k = 0, 1, \ldots, 6; \gamma_l \sim G(1, 10)$, $l = 1, 2$ e $\pi(\eta) = 1/20$, para $\eta \in \{0, 5, 1, \ldots, 10\}$. Os estudos computacionais MCMC foram realizados de modo similar ao apresentado na Seção 4.2. Geramos uma cadeia com 15.000 iterações, com *burn in* de 5.000 e saltos de tamanho 10, resultando em uma amostra de Gibbs de tamanho 1.000. O sumário a *posteriori* dos parâmetros do modelo de fração de cura hP ajustado aos dados melanoma com todas as covariáveis está exibido na Tabela 24, em que podemos notar claramente por meio dos percentis que a variável nódulo (x_3) tem um efeito significante no modelo uma vez que seu intervalo não contém o valor zero.

A fim de detectar observações influentes na distribuição a *posteriori* dos parâmetros,

Tabela 24 – Média, desvio padrão e percenil a *posteriori* do modelo de fração de cura hP ajustado nos dados melanoma com todas as covariáveis.

Parâmetro	Média	Desvio padrão	Percentil 2,5%	Percentil 97,5%
η	6,032	0,532	5,000	7,000
γ_1	1,797	0,120	1,573	2,042
γ_2	2,860	0,242	2,470	3,414
$\beta_{\text{intercepto}}$	0,506	0,199	0,147	0,902
$\beta_{\text{tratamento}}$	0,033	0,110	-0,181	0,237
β_{idade}	-0,054	0,108	-0,263	0,160
$\beta_{\text{nódulo}}$	0,289	0,0523	0,183	0,386
β_{sexo}	-0,161	0,115	-0,392	0,050
$\beta_{\text{capacidade}}$	0,100	0,165	-0,248	0,396
$\beta_{\text{espessura}}$	0,083	0,160	-0,253	0,384

Fonte: Elaborada pelo autor.

obtivemos as estimativas das medidas de divergência ψ por meio das amostras a *posteriori* dos parâmetros. A Figura 15 mostra que não existe nenhuma observação influente no conjunto de dados.

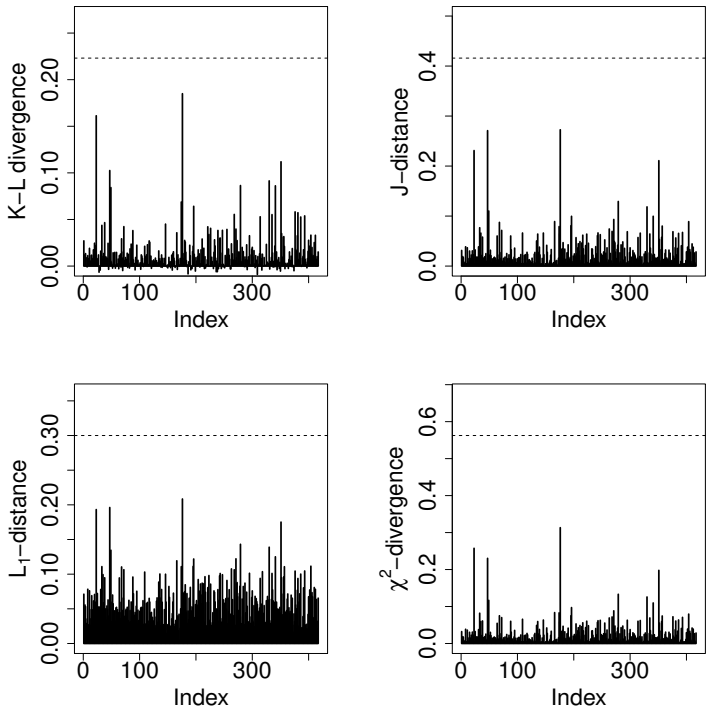


Figura 15 – Medidas de divergência ψ dos dados melanoma.

Fonte: Elaborada pelo autor.

Como visto na Tabela 24, apenas a covariável nódulo é significante. Diante disto, condu-

zimos um estudo com o modelo de fração de cura hP simplificado, ou seja, tomando apenas a categoria nódulo (x_3) como covariável. O sumário a *posteriori* dos parâmetros está exposto na Tabela 25. Os gráficos das densidades das distribuições a *posteriori* marginais de $\eta, \gamma_1, \gamma_2, \beta_0$ e β_3 estão exibidos na Figura 16. A Figura 17 apresenta os valores simulados das variáveis indicando a convergência das cadeias.

Tabela 25 – Sumário a *posteriori* do modelo de fração de cura HP com a apenas a categoria nódulo como covariável ajustado ao conjunto de dados melanoma.

Parâmetro	Média	Desvio padrão	Percentil 2,5%	Percentil 97,5%
η	6,014	0,541	5,000	7,000
γ_1	1,783	0,121	1,549	2,022
γ_2	2,867	0,274	2,477	3,549
$\beta_{\text{intercepto}}$	0,485	0,169	0,130	0,808
$\beta_{\text{nódulo}}$	0,293	0,053	0,189	0,400

Fonte: Elaborada pelo autor.

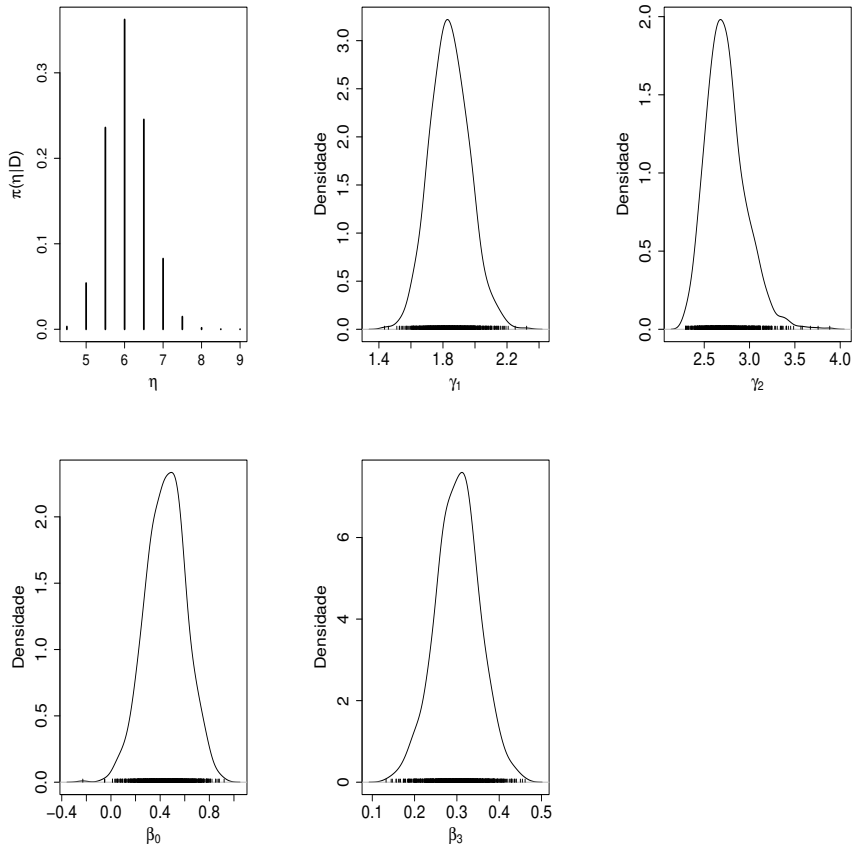


Figura 16 – Densidades a *posteriori* marginais de $\eta, \gamma_1, \gamma_2, \beta_0$ e β_3.

Fonte: Elaborada pelo autor.

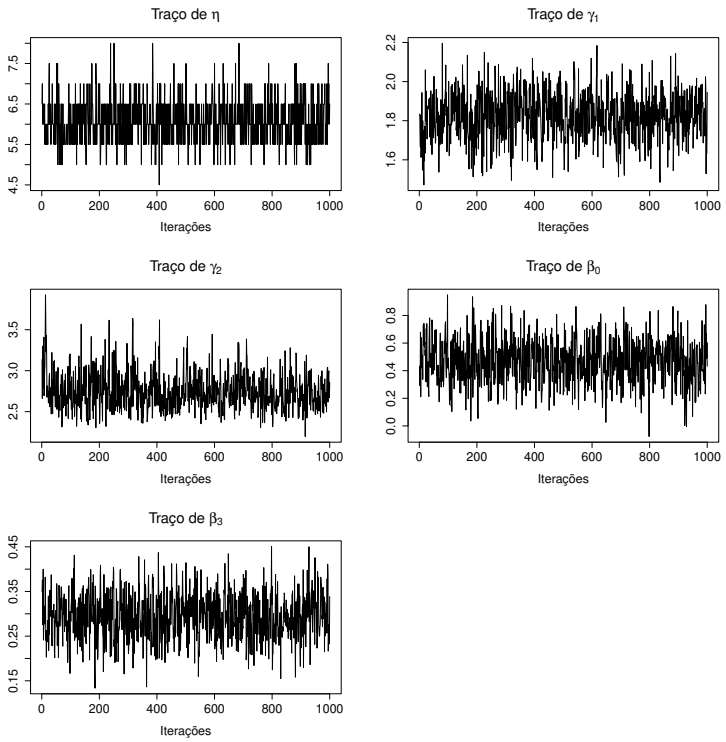


Figura 17 – Gráficos dos traços das cadeias de $\eta, \gamma_1, \gamma_2, \beta_0$ e β_3.

Fonte: Elaborada pelo autor.

Para comparar o modelo de fração de cura hP na presença de todas as covariáveis com o modelo hP simplificado, obtivemos ainda os valores dos critérios DIC e LPLM, que podem ser vistos na Tabela 26. Baseados nestes critérios, vemos que o modelo simplificado fornece um melhor ajuste em relação ao modelo com todas as covariáveis. Pelo trabalho de Cancho, Castro e Rodrigues (2012) podemos notar que os valores dos critérios DIC e LPLM do modelo de fração de cura de promoção ajustado aos dados melanoma considerando apenas a categoria nódulo como covariável são 1034,11 e -517,54, respectivamente. Ao comparar estes valores como os correspondentes do modelo de fração de cura hP, constatamos que o modelo hP supera o modelo de fração de cura de promoção, mas não supera o modelo geométrico. Isto ocorre aqui talvez devido ao uso da distribuição a *priori* uniforme para o parâmetro η que é menos informativa que a distribuição a *priori* escolhida em Cancho, Castro e Rodrigues (2012) para o correspondente parâmetro de dispersão da distribuição Poisson Conway-Maxwell.

Tabela 26 – Critérios bayesianos para o modelo completo e para o modelo reduzido com apenas a categoria nódulo como covariável.

Critério	Modelo	
	Com todas as covariáveis	Com apenas a categoria nódulo
DIC	1036,91	1030,91
LPLM	-519,15	-515,66

A Figura 18 (a) ilustra as curvas de Kaplan-Meier estratificadas pela categoria nódulo (1-4) e as estimativas bayesianas da funções de sobrevivência, mostrando que o modelo proposto fornece uma boa aproximação para as estimativas de Kaplan-Meier da função de sobrevivência. Na figura 18 (b) está exposta a função de sobrevivência estratificada pela categoria nódulo para indivíduos sob risco, da qual notamos que os pacientes sob risco qualquer categoria de nódulo tem uma baixa chance de sobreviver depois de sete anos, e isso diminui mais rapidamente para a categoria de nódulo 4.

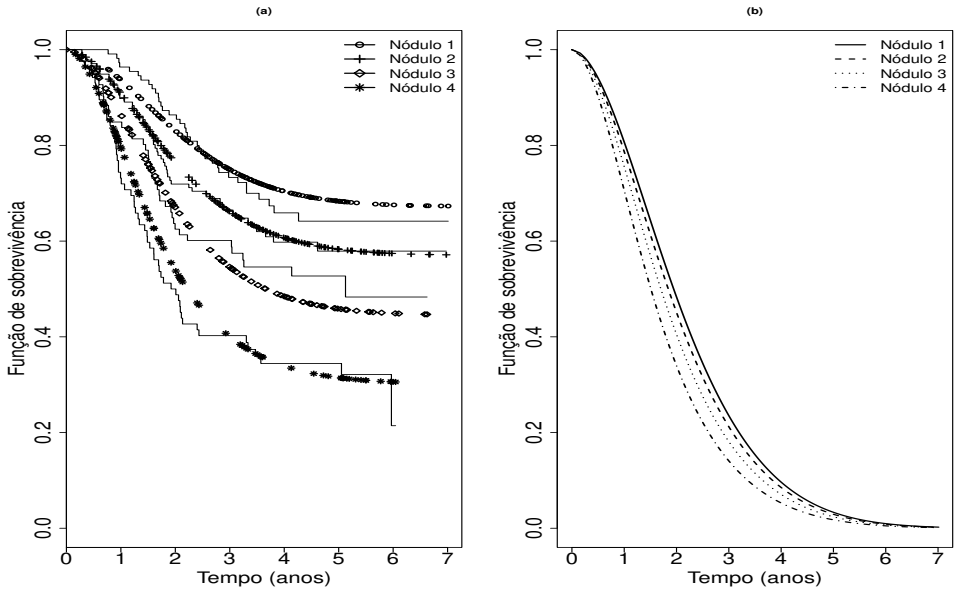


Figura 18 – (a): Curvas de Kaplan-Meier estratificadas pela categoria nódulo (1-4) junto com as estimativas bayesianas da função de sobrevivência hP, (b): Função de sobrevivência estratificada pela categoria nódulo para indivíduos sob risco.

Fonte: Elaborada pelo autor.

Finalizamos nossas análises com os dados melanoma considerando as estimativas das taxas de curados (p_0) empregando o modelo de fração de cura hP. Para tal, os paciente foram estratificados de acordo com a categoria nódulo. Da Seção 3.2, temos que $p_{0_j} = 1/{}_1F_1(1,\eta,\theta_j)$, com $\theta_j = \exp(\beta_0 + j\beta_3)$ para $j = 1,\ldots,4$. Na Tabela 27, o sumário a *posteriori* das 1.000 amostras extraídas do modelo de fração de cura hP revela que p_{0_j} é decrescente com respeito a j-ésima categoria de nódulo, para $j = 1,2,3,4$. A Figura 19 ilustra as densidades a *posteriori* marginais aproximadas de p_{0_j}.

Tabela 27 – Sumário a *posteriori* para a fração de cura p_0 para o modelo hP de acordo com a categoria nódulo.

Fração de cura	Média	Desvio padrão	Percentil 2.5%	Percentil 97.5%
p_{0_1}	0,654	0,040	0,594	0,748
p_{0_2}	0,552	0,037	0,497	0,633
p_{0_3}	0,428	0,039	0,349	0,491
p_{0_4}	0,289	0,051	0,183	0,393

Fonte: Elaborada pelo autor.

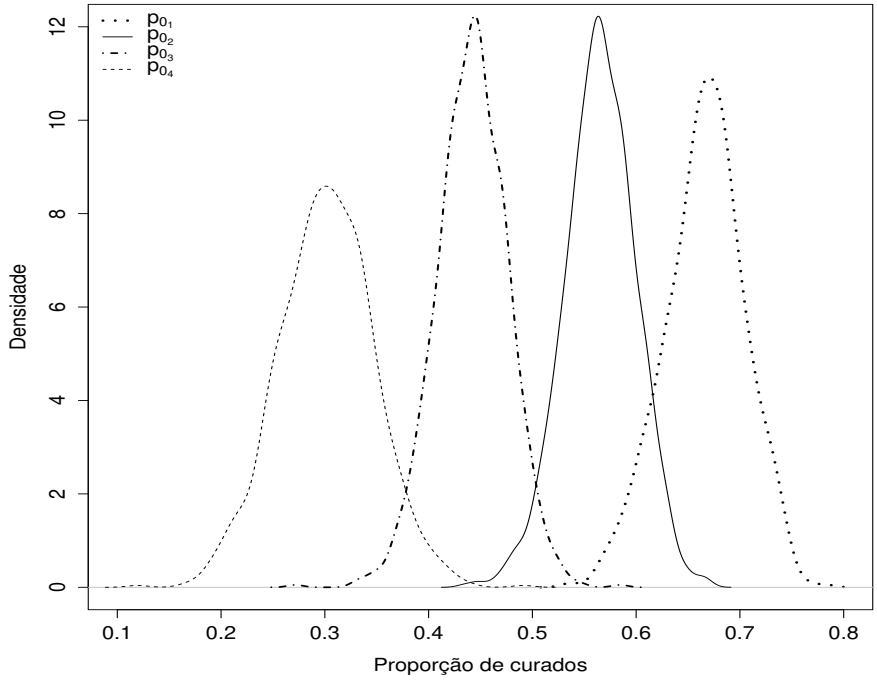


Figura 19 – Densidades a *posteriori* marginais aproximadas da proporção de curados (p_0) para o modelo de fração de cura hP de acordo com a categoria nódulo (1-4).

Fonte: Elaborada pelo autor.

CAPÍTULO

5

CONCLUSÕES E PROPOSTAS FUTURAS

Na primeira parte desta tese foi trabalhada a distribuição hiper-Poisson como uma variável de contagem observada. Dotada de um parâmetro a mais em relação a distribuição Poisson, um parâmetro de dispersão, a distribuição hiper-Poisson modela dados com sobredispersão e subdispersão. Como a distribuição hiper-Poisson apresenta forma fechada para a média, a distribuição hiper-Poisson foi reparametrizada e covariáveis foram ligadas diretamente na média. Além disso, foram adicionadas covariáveis em um dos parâmetros do modelo, formando assim dois modelos de regressão distintos, que foram averiguados por meio dos estudos de simulação e aplicados a conjunto de dados reais. Procedimentos inferenciais baseados na abordagem da máxima verossimilhança foram desenvolvidos para a estimação dos parâmetros do modelo hiper-Poisson e para os modelos de regressão hiper-Poisson. Não foram necessários grande esforços computacionais, o procedimento de estimação mostrou-se eficaz e as inferências sobre os parâmetros foram bastante adequadas. Interessados em testar o parâmetro de dispersão do modelo hiper-Poisson, os testes de hipóteses da razão de verossimilhanças e gradiente foram formulados para o modelo hiper-Poisson e para os dois modelos de regressão. Com os estudos de simulação foram avaliados o desempenho desses testes no que diz respeito ao controle do erro do tipo I e em relação ao poder. Os resultados mostraram que a estatística gradiente é bastante competitiva em comparação à estatística clássica da razão de verossimilhanças. Com a constatação do bom desempenho das metodologias adotadas, foram realizadas duas aplicações a conjunto de dados reais provenientes da literatura.

Um novo modelo para dados de sobrevivência com fração de cura foi proposto na segunda parte desta tese. O novo modelo foi obtido a partir de um modelo de fragilidade com distribuição discreta hiper-Poisson. Dessa vez a distribuição hiper-Poisson atua como uma variável não observável (latente) que devido ao seu parâmetro de dispersão permite a expressão da dispersão dos fatores de riscos. Na inferência clássica, utilizamos o algoritmo EM juntamente com a técnica da verossimilhança perfilada para a estimação dos parâmetros do modelo. A fim de analisar as propriedades assintóticas dos estimadores de máxima verossimilhança, estudos de simulação

foram realizados e os resultados obtidos foram muito bons. Além disso, os testes de hipóteses da razão de verossimilhanças e gradiente foram desenvolvidos, porém agora com o objetivo de testar os coeficientes de regressão do novo modelo. Os testes mostraram ótimo desempenho e resultados muito parecidos entre si em relação ao poder e às taxas de rejeição da hipótese nula. O novo modelo e a metologia escolhida foi avaliada em uma aplicação a um conjunto de dados muito utilizado na literatura sobre o estudo de pacientes com melanoma.

Alternativamente à metodologia da inferência clássica, foram desenvolvidos procedimentos inferenciais sob uma perspectiva bayesiana para o modelo com fração de cura hP proposto. A função de verossimilhança foi formulada sob o ponto de vista bayesiano e critérios de comparação de modelos e medidas de diagnóstico foram estudados. O algoritmo Metropolis–Hasting foi implementado e as simulações de Monte Carlo via Cadeia de Markov foram realizadas para a estimação dos parâmetros. Os resultados das simulações foram satisfatórios. O conjunto de dados sobre os pacientes com melanoma foi novamente aplicado a fim de validar a metodologia utilizada.

Por fim, existem diversas pesquisas que podem ser realizadas como continuação da desenvolvida neste trabalho. Dentre estas, propomos os seguintes tópicos:

(i) Estender os estudos feitos no modelo hiper-Poisson para o modelo COM-Poisson estendido em problemas de contagem considerando seus casos particulares e comparando-os entre si;

(ii) Desenvolver o teste de hipótese gradiente para o modelo COM-Poisson e fazer um estudo comparativo com o modelo hiper-Poisson;

(iii) Estudar as propriedades estruturais do modelo induzido por fragilidade proposto assumindo que a função de risco de base $h_0(t)$ segue outras distribuições como a distribuição exponencial, distribuição Gompertz, entre outras;

(iv) Estender o modelo de sobrevivência proposto para dados com censura intervalar (HASHIMOTO *et al.*, 2010);

(y) Propor um modelo de sobrevivência com fração de cura com erro de medida nas linhas de Mizoi (2004) e Mizoi (2004);

(vi) Propor um modelo semiparamétrico (IBRAHIM; CHEN; SINHA, 2001a) baseado no modelo proposto nesta tese e desenvolver procedimentos inferenciais sob as perspectivas clássica e bayesiana;

(vii) Utilizar o modelo com fração de cura hiper-Poisson proposto para dados com eventos recorrentes, estendendo assim o trabalho de Macera *et al.* (2015). Além disso, investigar o uso de distribuições de fragilidade discretas em modelos semiparamétricos para eventos recorrentes;

(viii) Propor um modelo dinâmico baseado no modelo de Kim *et al.* (2007) e desenvolver o procedimento de inferência bayesiana para o modelo proposto;

(ix) Realizar os procedimentos de diagnóstico em uma perspectiva clássica para o modelo de sobrevivência induzido por fragilidade hP proposto aqui.

(x) Estender o modelo proposto para dados de sobrevivência bivariados.

REFERÊNCIAS

ANCARANI, L.; GASANEO, G. Derivatives of any order of the confluent hypergeometric function. **Journal of Mathematical Physics**, v. 49, 2008. Citado na página 33.

ATA, N.; ÖZEL, G. Survival functions for the frailty models based on the discrete compound Poisson process. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 83, p. 2105–2116, 2013. Citado na página 54.

BALAKRISHNAN, N.; PENG, Y. Generalized gamma frailty model. **Statistics in Medicine**, v. 25, p. 2797–2816, 2006. Citado na página 54.

BARDWELL, G. E.; CROW, E. L. A two-parameter family of hyper-Poisson distributions. **Journal of the American Statistical Association**, Taylor & Francis, v. 59, p. 133–141, 1964. Citado nas páginas 21, 26 e 27.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, v. 47, p. 501–515, 1952. Citado nas páginas 52 e 57.

BICKEL, P. J.; DOKSUM, K. A. **Mathematical statistics: basic ideas and selected topics**. [S.l.]: CRC Press, Boca Raton, 2015. Citado na página 30.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **Journal of the Royal Statistical Society B**, v. 11, p. 15–53, 1949. Citado nas páginas 52 e 57.

BRESLOW, N. E. Extra-Poisson variation in log-linear models. **Applied Statistics**, JSTOR, v. 33, p. 38–44, 1984. Citado na página 26.

CAMERON, A. C.; JOHANSSON, P. *et al.* Count data regression using series expansions: with applications. **Journal of Applied Econometrics**, JSTOR, v. 12, p. 203–223, 1997. Citado na página 46.

CANCHO, V.; DEY, D.; LACHOS, V.; ANDRADE, M. Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. **Computational Statistics & Data Analysis**, Elsevier, v. 55, p. 588–602, 2011. Citado na página 75.

CANCHO, V.; ORTEGA, E.; PAULA, G. On estimation and influence diagnostics for log-Birnbaum-Saunders Student-t regression models: Full Bayesian analysis. **Journal of Statistical Planning and Inference**, Elsevier, v. 140, p. 2486–2496, 2010. Citado na página 75.

CANCHO, V.; RODRIGUES, J.; CASTRO, M. de. A flexible model for survival data with a cure rate: a Bayesian approach. **Journal of Applied Statistics**, Taylor and Francis Journals, v. 38, p. 57–70, 2011. Citado nas páginas 52 e 54.

CANCHO, V. G.; CASTRO, M. de; RODRIGUES, J. A Bayesian analysis of the Conway–Maxwell–Poisson cure rate model. **Statistical Papers**, Springer, v. 53, p. 165–176, 2012. Citado nas páginas 52 e 83.

CARONI, C.; CROWDER, M.; KIMBER, A. Proportional hazards models with discrete frailty. **Lifetime Data Analysis**, Springer, v. 16, p. 374–384, 2010. Citado nas páginas 22, 54 e 55.

CASTILLO, J. del; PÉREZ-CASANY, M. Overdispersed and underdispersed Poisson generalizations. **Journal of Statistical Planning and Inference**, Elsevier, v. 134, p. 486–500, 2005. Citado na página 26.

CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. A new Bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 94, p. 909–919, 1999. Citado nas páginas 52 e 55.

CLAYTON, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, Biometrika Trust, v. 65, p. 141–151, 1978. Citado na página 53.

CONSUL, P.; FAMOYE, F. Generalized Poisson regression model. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 21, p. 89–109, 1992. Citado na página 26.

CONWAY, R. W.; MAXWELL, W. L. A queuing model with state dependent service rates. **Journal of Industrial Engineering**, v. 12, p. 132–136, 1962. Citado nas páginas 21 e 26.

COOK, R. D. Assessment of local influence. **Journal of the Royal Statistical Society, Series B**, v. 48, p. 133–169, 1986. Citado na página 74.

COOK, R. D.; WEISBERG, S. **Residuals and Influence in Regression**. [S.l.]: Chapman & Hall, New York, 1982. Citado na página 74.

COONER, F.; BANERJEE, S.; CARLIN, B. P.; SINHA, D. Flexible cure rate modeling under latent activation schemes. **Journal of the American Statistical Association**, v. 102, p. 560–572, 2007. Citado na página 52.

CORDEIRO, G. M.; CANCHO, V. G.; ORTEGA, E. M.; BARRIGA, G. D. A model with long-term survivors: Negative binomial Birnbaum-Saunders. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 45, p. 1370–1387, 2016. Citado na página 52.

COWLES, M. K.; CARLIN, B. P. Markov chain Monte Carlo convergence diagnostics: A comparative review. **Journal of the American Statistical Association**, v. 91, p. 883–904, 1996. Citado na página 78.

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. London: Chapman and Hall, 1974. xii+511 p. Citado na página 60.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 39, p. 1–22, 1977. Citado na página 59.

DEY, D. K.; BIRMIWAL, L. R. Robust Bayesian analysis using divergence measures. **Statistics & Probability Letters**, Elsevier, v. 20, p. 287–294, 1994. Citado na página 74.

EFRON, B. Double exponential families and their use in generalized linear regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 81, p. 709–721, 1986. Citado na página 26.

FAMOYE, F.; WANG, W. Censored generalized poisson regression model. **Computational Statistics & Data Analysis**, Elsevier, v. 46, p. 547–560, 2004. Citado na página 26.

FRANCIS, R. A.; GEEDIPALLY, S. R.; GUIKEMA, S. D.; DHAVALA, S. S.; LORD, D.; LAROCCA, S. Characterizing the performance of the Conway-Maxwell Poisson generalized linear model. **Risk Analysis**, Wiley Online Library, v. 32, p. 167–183, 2012. Citado nas páginas 21 e 26.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference**. 2nd. ed. Boca Raton: Chapman & Hall/CRC, 2006. Citado na página 73.

GEISSER, S.; EDDY, W. A predictive approach to model selection. **Journal of the American Statistical Association**, JSTOR, v. 74, p. 153–160, 1979. Citado na página 73.

GELFAND, A.; DEY, D.; CHANG, H. Model determination using predictive distributions with implementation via sampling based methods (with discussion). **Bayesian Statistics 4. Eds: J. Bernardo et al**, v. 1, p. 147–167, 1992. Citado na página 73.

GSCHLÖSSL, S.; CZADO, C. Modelling count data with overdispersion and spatial effects. **Statistical papers**, Springer, v. 49, p. 531–552, 2008. Citado na página 26.

HASHIMOTO, E. M.; ORTEGA, E. M.; CANCHO, V. G.; CORDEIRO, G. M. The log-exponentiated weibull regression model for interval-censored data. **Computational Statistics & Data Analysis**, Elsevier, v. 54, n. 4, p. 1017–1035, 2010. Citado na página 88.

HOUGAARD, P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. **Biometrika**, Biometrika Trust, v. 71, p. 75–83, 1984. Citado na página 54.

____. A class of multivanate failure time distributions. **Biometrika**, Biometrika Trust, v. 73, p. 671–678, 1986. Citado na página 54.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. Bayesian semiparametric models for survival data with a cure fraction. **Biometrics**, Wiley Online Library, v. 57, n. 2, p. 383–388, 2001. Citado na página 88.

____. **Bayesian Survival Analysis**. New York: Springer, 2001. Citado nas páginas 52, 64 e 74.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. **Univariate Discrete Distributions**. [S.l.]: John Wiley, New Jersey, 2005. Citado na página 27.

KADANE, J. B.; SHMUELI, G.; MINKA, T. P.; BORLE, S.; BOATWRIGHT, P. Conjugate analysis of the Conway-Maxwell-Poisson distribution. **Bayesian Analysis**, v. 1, p. 363–374, 2006. Citado na página 26.

KIM, S.; CHEN, M.-H.; DEY, D. K.; GAMERMAN, D. Bayesian dynamic models for survival data with a cure fraction. **Lifetime Data Analysis**, Springer, v. 13, n. 1, p. 17–35, 2007. Citado na página 89.

KIRKWOOD, J. M.; IBRAHIM, J. G.; SONDAK, V. K.; RICHARDS, J.; FLAHERTY, L. E.; ERNSTOFF, M. S.; SMITH, T. J.; RAO, U.; STEELE, M.; BLUM, R. H. High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of Intergroup Trial E1690/S9111/C9190. **Journal of Clinical Oncology**, v. 18, p. 2444–2458, 2000. Citado na página 62.

KOUTRAS, M.; MILIENOS, F. A flexible family of transformation cure rate models. **Statistics in medicine**, Wiley Online Library, v. 36, p. 2559–2575, 2017. Citado nas páginas 59 e 65.

LANCASTER, T. Econometric methods for the duration of unemployment. **Econometrica**, JSTOR, v. 47, p. 939–956, 1979. Citado na página 53.

LANGE, K. A gradient algorithm locally equivalent to the em algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 57, p. 425–437, 1995. Citado na página 59.

LAWLESS, J. F. Negative binomial and mixed poisson regression. **Canadian Journal of Statistics**, Wiley Online Library, v. 15, p. 209–225, 1987. Citado na página 26.

LEE, D.-J.; DURBÁN, M. Smooth-CAR mixed models for spatial count data. **Computational Statistics & Data Analysis**, Elsevier, v. 53, p. 2968–2979, 2009. Citado na página 26.

LEMONTE, A. **The Gradient Test: Another Likelihood-Based Test**. [S.l.]: Academic Press, London, 2016. Citado nas páginas 29 e 31.

LEMONTE, A. J. On the gradient statistic under model misspecification. **Statistics & Probability Letters**, Elsevier, v. 83, p. 390–398, 2013. Citado na página 32.

LEMONTE, A. J.; FERRARI, S. L. Testing hypotheses in the birnbaum–saunders distribution under type-ii censored samples. **Computational Statistics & Data Analysis**, Elsevier, v. 55, n. 7, p. 2388–2399, 2011. Citado nas páginas 32 e 50.

____. Local power and size properties of the lr, wald, score and gradient tests in dispersion models. **Statistical Methodology**, Elsevier, v. 9, p. 537–554, 2012. Citado na página 32.

____. The local power of the gradient test. **Annals of the Institute of Statistical Mathematics**, Springer, v. 64, p. 373–381, 2012. Citado na página 32.

LINDSEY, J. **Modelling Frequency and Count Data**. [S.l.]: Oxford University Press, New York, 1995. Citado na página 26.

MACERA, M. A.; LOUZADA, F.; CANCHO, V. G.; FONTES, C. J. The exponential-poisson model for recurrent event data: An application to a set of data on malaria in brazil. **Biometrical Journal**, Wiley Online Library, v. 57, n. 2, p. 201–214, 2015. Citado na página 88.

MALLER, R. A.; ZHOU, X. **Survival Analysis with Long-Term Survivors**. New York: Wiley, 1996. Citado na página 52.

____. **Survival analysis with long-term survivors**. [S.l.]: John Wiley & Sons, 1996. Citado na página 59.

MIZOI, M. F. **Influência local em modelos de sobrevivência com fração de cura**. Tese (Doutorado) — Instituto de Matemática e Estatística da Universidade de São Paulo, 12/11/2004., 2004. Citado na página 88.

OAKES, D. A model for association in bivariate survival data. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 44, p. 414–422, 1982. Citado na página 53.

PARDO, L. **Statistical Inference Based on Divergence Measures**. Boca Raton: Chapman & Hall/CRC, 2006. Citado na página 74.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME–USP, São Paulo, 2004. Citado na página 49.

PENG, F.; DEY, D. Bayesian analysis of outlier problems using divergence measures. **Canadian Journal of Statistics**, Wiley Online Library, v. 23, p. 199–213, 1995. Citado na página 75.

PLAN, E. L. Modeling and simulation of count data. **CPT: pharmacometrics & systems pharmacology**, Wiley Online Library, v. 3, n. 8, p. 1–12, 2014. Citado na página 25.

PRESS, W. H.; TEUKOLSKY, S. A. ; VETTERLING, W. T.; FLANNERY, B. P. **Numerical Recipes: The Art of Scientific Computing 3rd ed.** [S.l.]: Cambridge University Press, New York, 2007. Citado na página 33.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>. Citado nas páginas 35, 37 e 59.

RIDOUT, M. S.; BESBEAS, P. An empirical model for underdispersed count data. **Statistical Modelling**, SAGE Publications, v. 4, p. 77–89, 2004. Citado na página 26.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de; LOUZADA-NETO, F. On the unification of the long-term survival models. **Statistics & Probability Letters**, v. 79, p. 753–759, 2009. Citado na página 52.

RODRIGUES, J.; CASTRO, M. de; CANCHO, V. G.; BALAKRISHNAN, N. COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. **Journal of Statistical Planning Inference**, v. 139, p. 3605–3611, 2009. Citado nas páginas 54 e 55.

RODRIGUES, J.; CORDEIRO, G. M.; CANCHO, V. G.; BALAKRISHNAN, N. Relaxed poisson cure rate models. **Biometrical Journal**, v. 58, p. 397–415, 2016. Citado na página 72.

RODRIGUES, J.; PRADO, S.; BALAKRISHNAN, N. Flexible M/G/1 queueing system with state dependent service rate. **Operations Research Letters**, Elsevier, v. 44, p. 383–389, 2016. Citado nas páginas 27 e 28.

SÁEZ-CASTILLO, A.; CONDE-SÁNCHEZ, A. A hyper-Poisson regression model for over-dispersed and underdispersed count data. **Computational Statistics & Data Analysis**, Elsevier, v. 61, p. 148–157, 2013. Citado nas páginas 27, 35, 36 e 46.

SANTOS, D.; CANCHO, V.; RODRIGUES, J. Hypothesis testing for the dispersion parameter of the hyper–poisson regression model. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 89, p. 763–775, 2019. Citado na página 23.

SANTOS, D. M. dos; DAVIES, R. B.; FRANCIS, B. Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer. **Journal of Statistical Planning and Inference**, Elsevier, v. 47, p. 111–127, 1995. Citado na página 54.

SELLERS, K. F.; BORLE, S.; SHMUELI, G. The COM–Poisson model for count data: a survey of methods and applications. **Applied Stochastic Models in Business and Industry**, Wiley Online Library, v. 28, p. 104–116, 2012. Citado na página 26.

SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. **The Annals of Applied Statistics**, v. 4, p. 943–961, 2010. Citado na página 26.

SEN, P. K.; SINGER, J. M. **Large sample methods in statistics: an introduction with applications**. [S.l.]: Chapman & Hall, New York, 1993. Citado na página 30.

SHMUELI, G.; MINKA, T. P.; KADANE, J. B.; BORLE, S.; BOATWRIGHT, P. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. **Journal of the Royal Statistical Society: Series C**, Wiley Online Library, v. 54, p. 127–142, 2005. Citado na página 26.

SOUZA, D. de; CANCHO, V. G.; RODRIGUES, J.; BALAKRISHNAN, N. Bayesian cure rate models induced by frailty in survival analysis. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 26, p. 2011–2028, 2017. Citado na página 24.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society B**, v. 64, p. 583–639, 2002. Citado na página 73.

TERRELL, G. R. The gradient statistic. **Computing Science and Statistics**, v. 34, p. 206–215, 2002. Citado nas páginas 21, 29, 30 e 31.

TSODIKOV, A. D.; IBRAHIM, J. G.; YAKOVLEV, A. Y. Estimating cure rates from survival data: An alternative to two-component mixture models. **Journal of the American Statistical Association**, v. 98, p. 1063–1078, 2003. Citado nas páginas 52, 54 e 55.

VARGAS, T. M.; FERRARI, S. L.; LEMONTE, A. J. *et al.* Gradient statistic: Higher-order asymptotics and Bartlett-type correction. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 7, p. 43–61, 2013. Citado na página 32.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, Springer, v. 16, p. 439–454, 1979. Citado nas páginas 53 e 54.

WANG, W.; FAMOYE, F. Modeling household fertility decisions with generalized Poisson regression. **Journal of Population Economics**, Springer, v. 10, p. 273–283, 1997. Citado na página 26.

WEISS, R. An approach to Bayesian sensitivity analysis. **Journal of the Royal Statistical Society, Series B**, v. 58, p. 739–750, 1996. Citado na página 75.

WIENKE, A. **Frailty Models in Survival Analysis**. Boca Raton: CRC Press, 2010. Citado nas páginas 22 e 54.

YAKOVLEV, A. Y.; TSODIKOV, A. D. **Stochastic Models of Tumor Latency and Their Biostatistical Applications**. Singapore: World Scientific, 1996. Citado nas páginas 52 e 55.

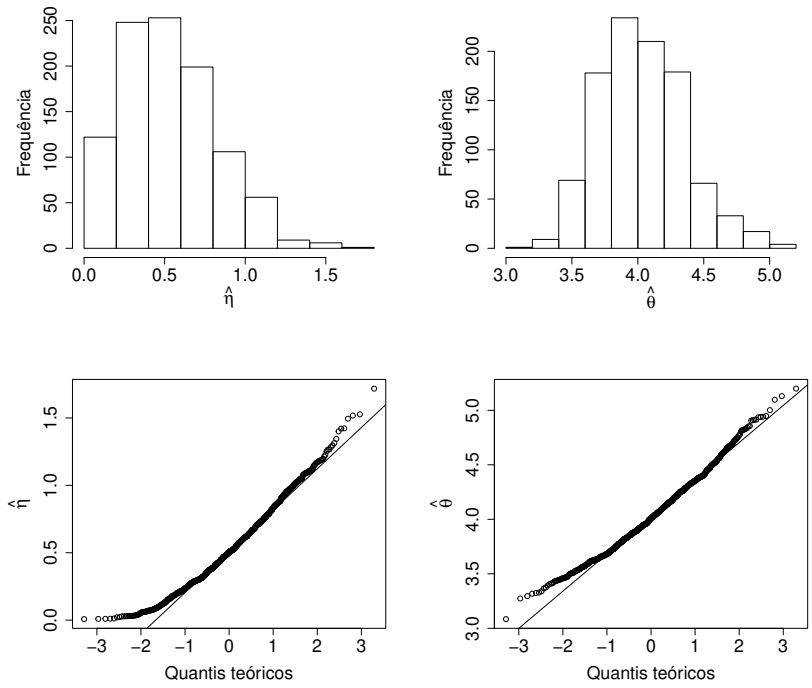
YIN, G.; IBRAHIM, J. G. Cure rate models: a unified approach. **The Canadian Journal of Statistics**, v. 33, p. 559–570, 2005. Citado nas páginas 52 e 72.

APÊNDICE

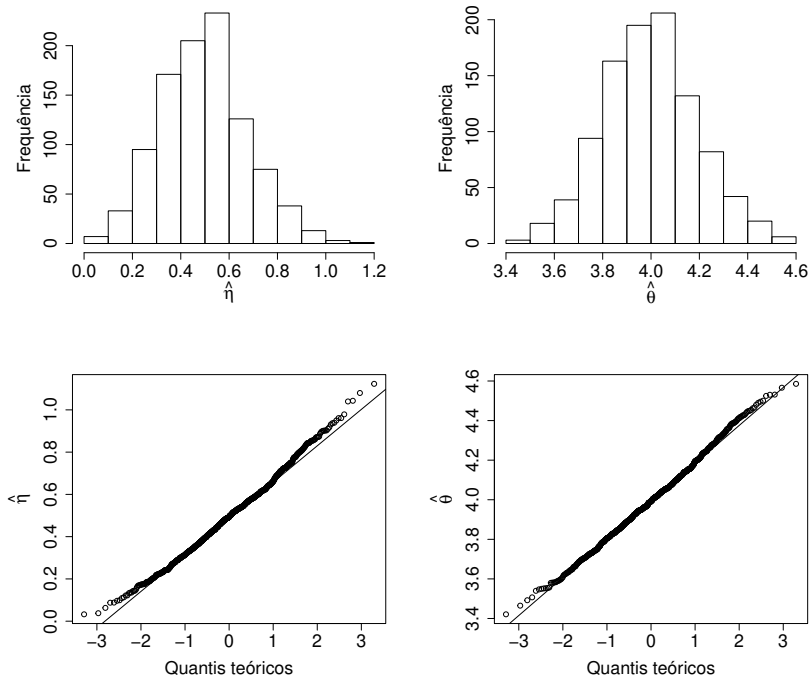
A

HISTOGRAMAS E GRÁFICOS QQ-NORMAL DOS PARÂMETROS ESTIMADOS PARA O MODELO HIPER-POISSON

Nesta seção apresentamos alguns dos histogramas e dos gráficos QQ-normal baseados nas 1.000 replicações de Monte Carlos para o modelo hiper-Poisson considerado na Seção 2.6.1.



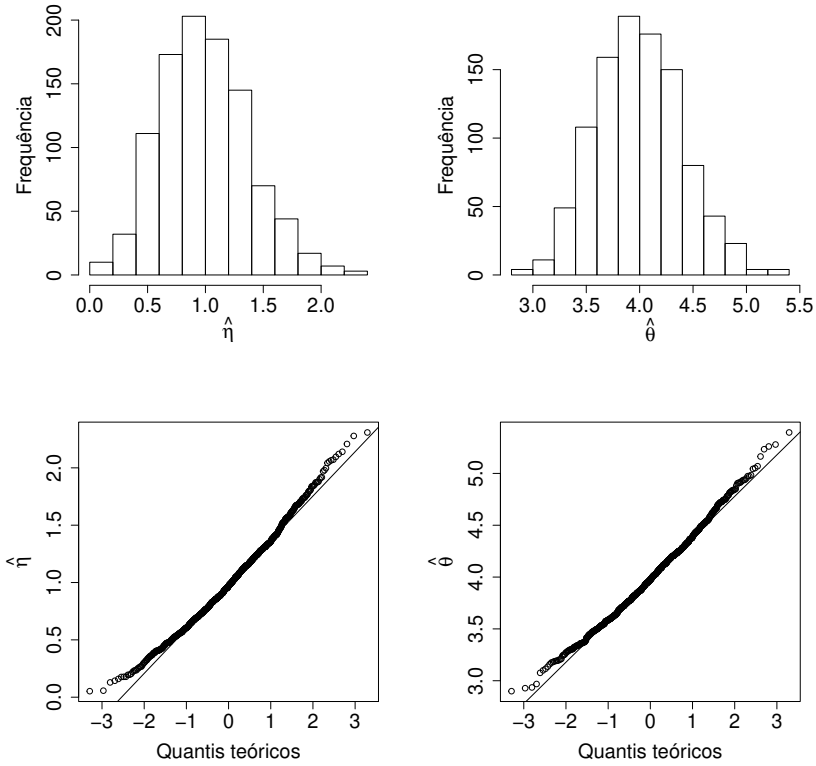
(a) $n = 200$



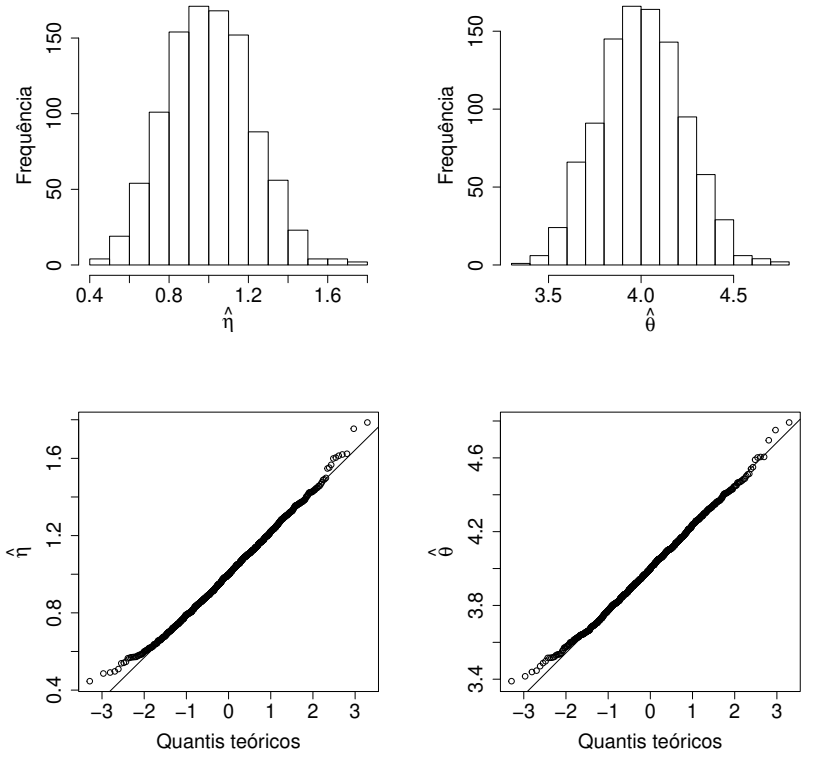
(b) $n = 600$

Figura 20 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo hP com $\eta = 0,5$ e $\theta = 4$.

Fonte: Elaborada pelo autor.



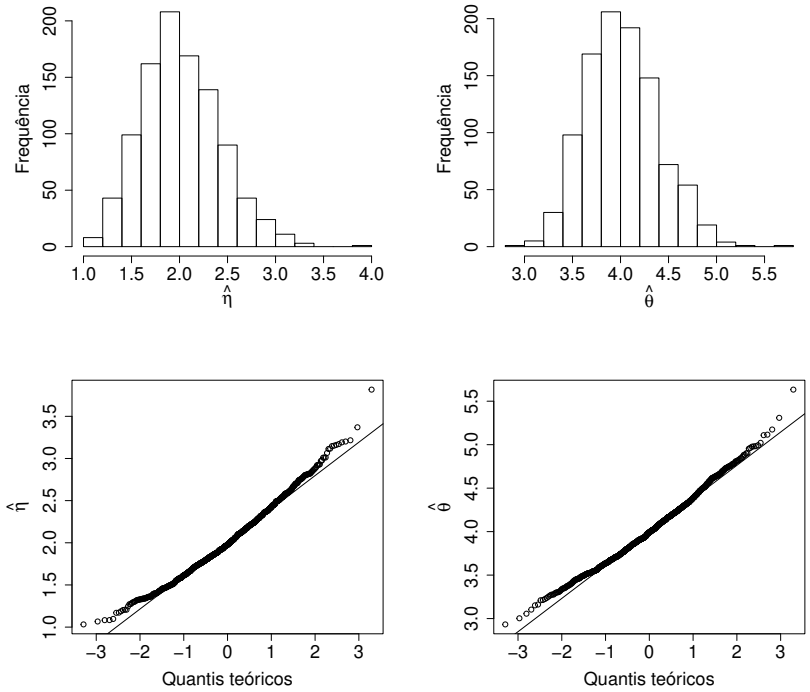
(a) $n = 200$



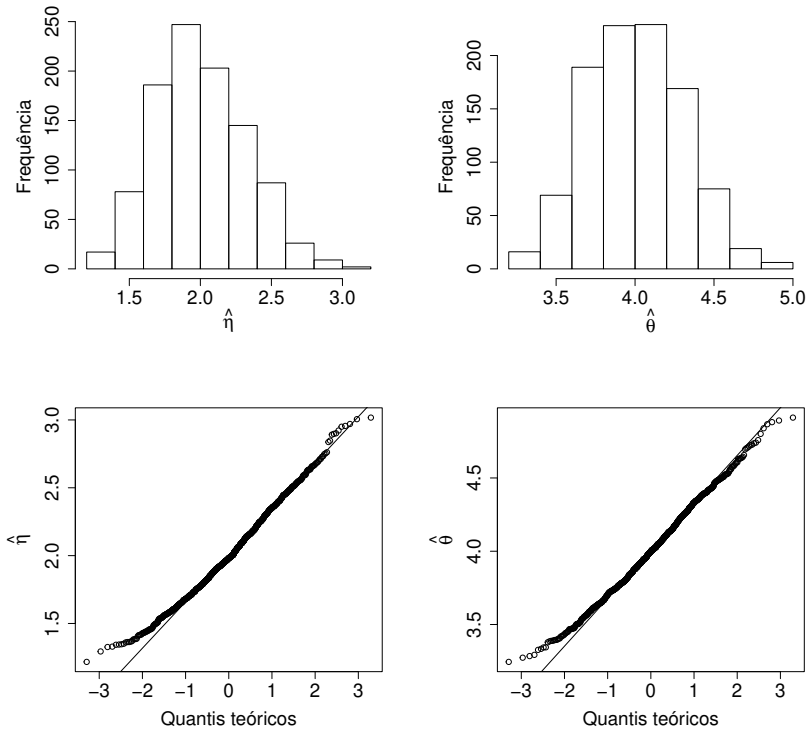
(b) $n = 600$

Figura 21 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo hP com $\eta = 1$ e $\theta = 4$.

Fonte: Elaborada pelo autor.



(a) $n = 200$



(b) $n = 600$

Figura 22 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo hP com $\eta = 2$ e $\theta = 4$.

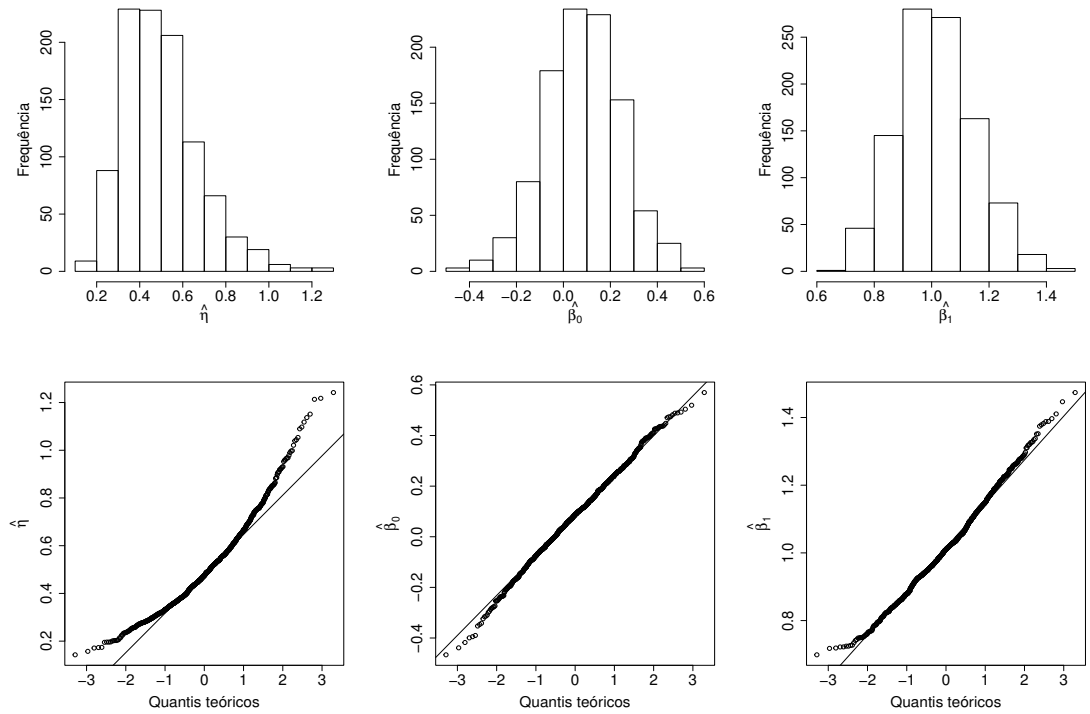
Fonte: Elaborada pelo autor.

APÊNDICE

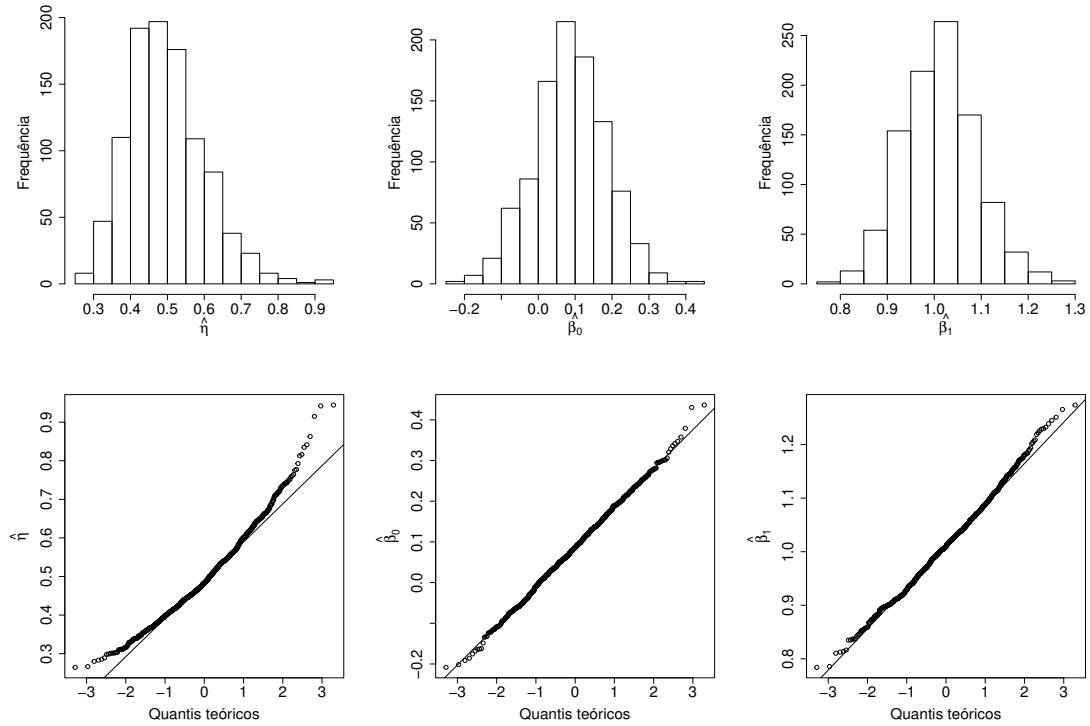
B

HISTOGRAMAS E GRÁFICOS QQ-NORMAL DOS PARÂMETROS ESTIMADOS PARA OS MODELOS DE REGRESSÃO HIPER-POISSON

Nesta seção apresentamos alguns dos histogramas e dos gráficos QQ-normal baseados nas 1.000 replicações de Monte Carlos para os dois modelos de regressão hiper-Poisson (com covariáveis no parâmetro θ e com covariáveis na média μ) considerados na Seção 2.6.2.



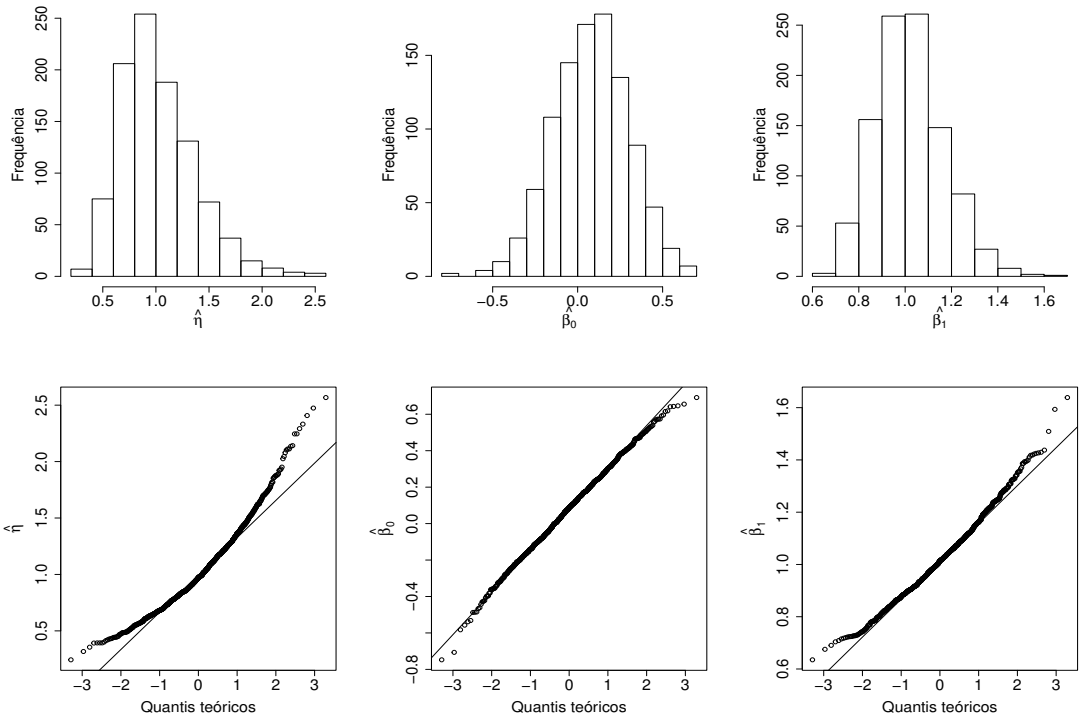
(a) $n = 200$



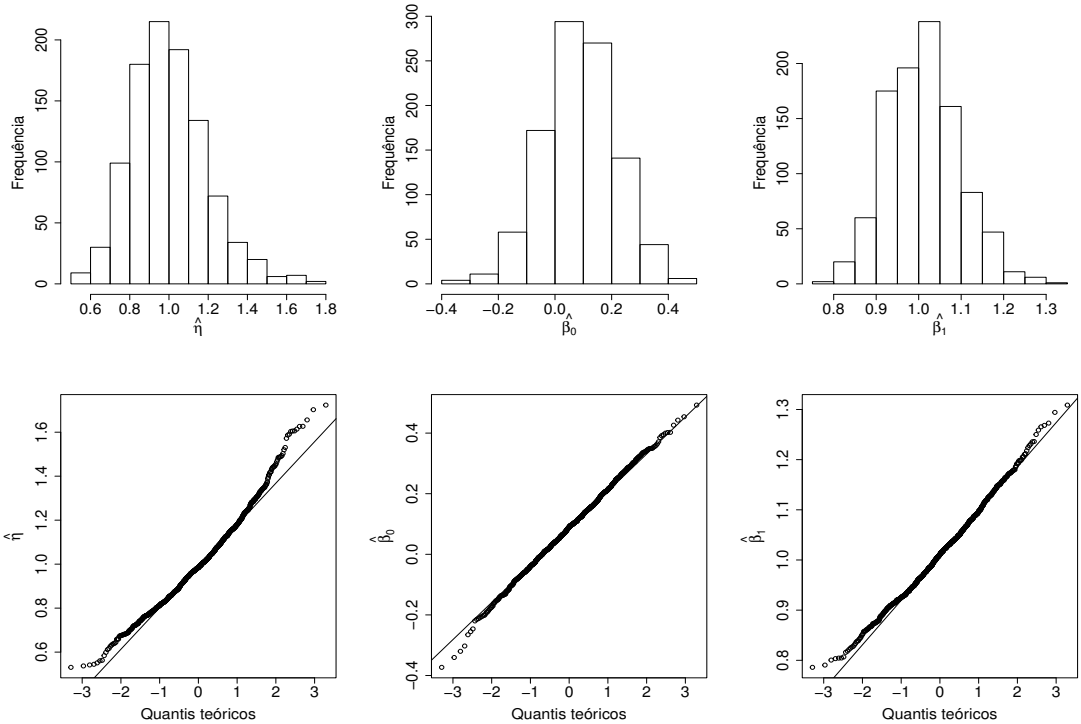
(b) $n = 600$

Figura 23 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo de regressão hP com regressores em θ e $\eta = 0,5$.

Fonte: Elaborada pelo autor.



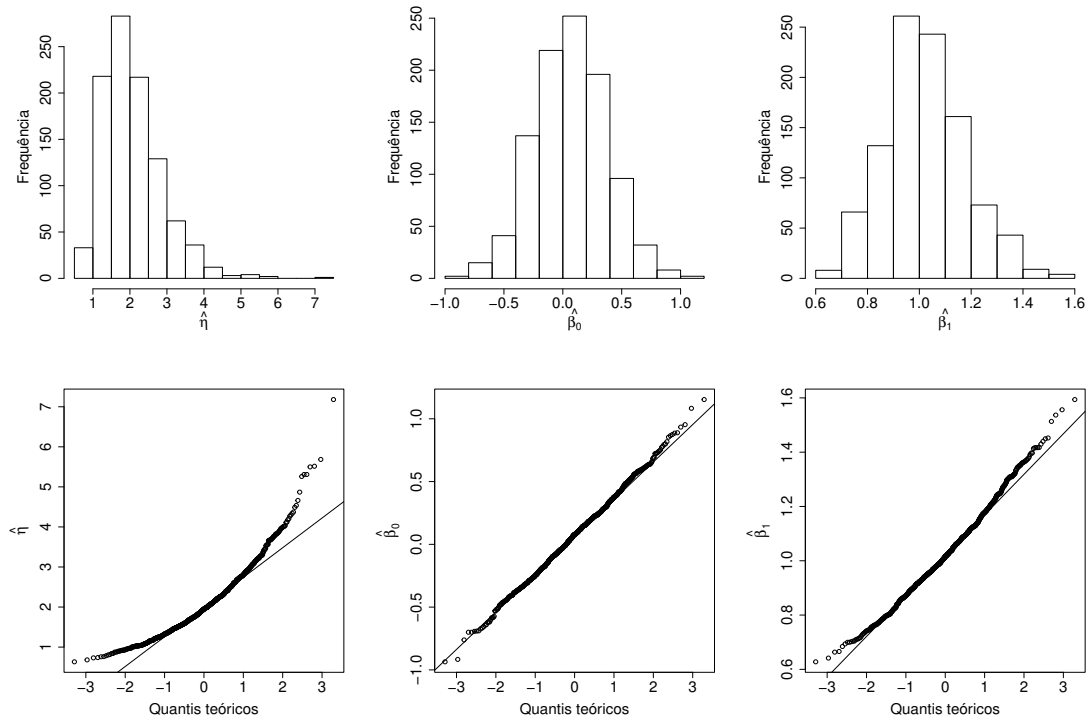
(a) $n = 200$



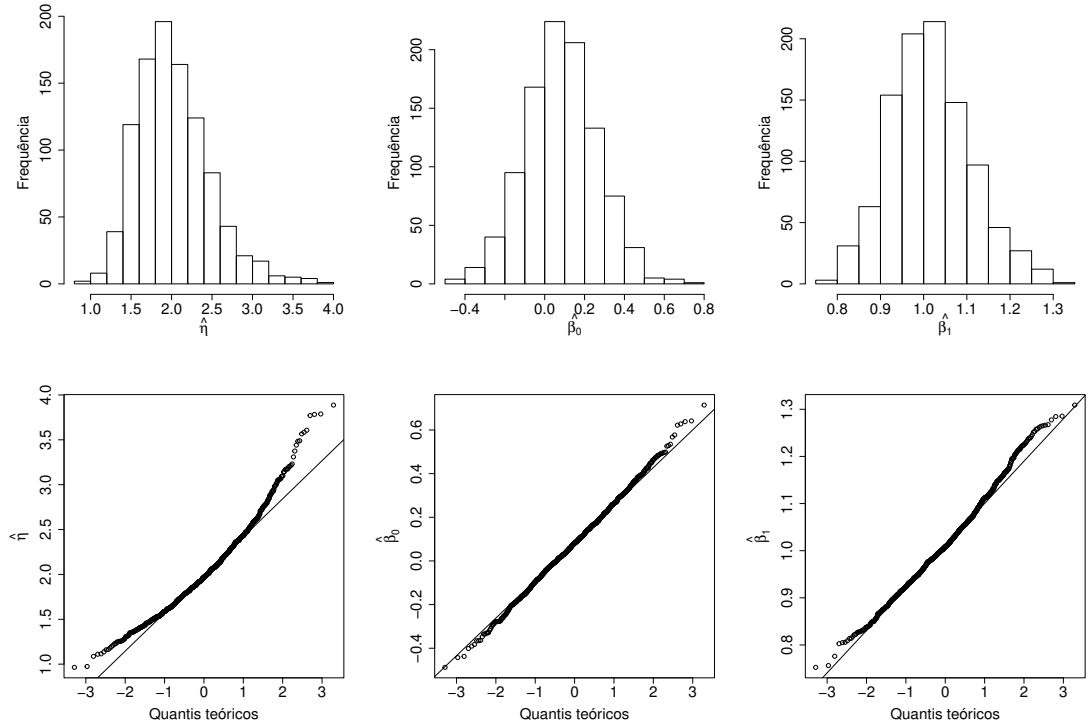
(b) $n = 600$

Figura 24 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo de regressão hP com regressores em θ e $\eta = 1$.

Fonte: Elaborada pelo autor.



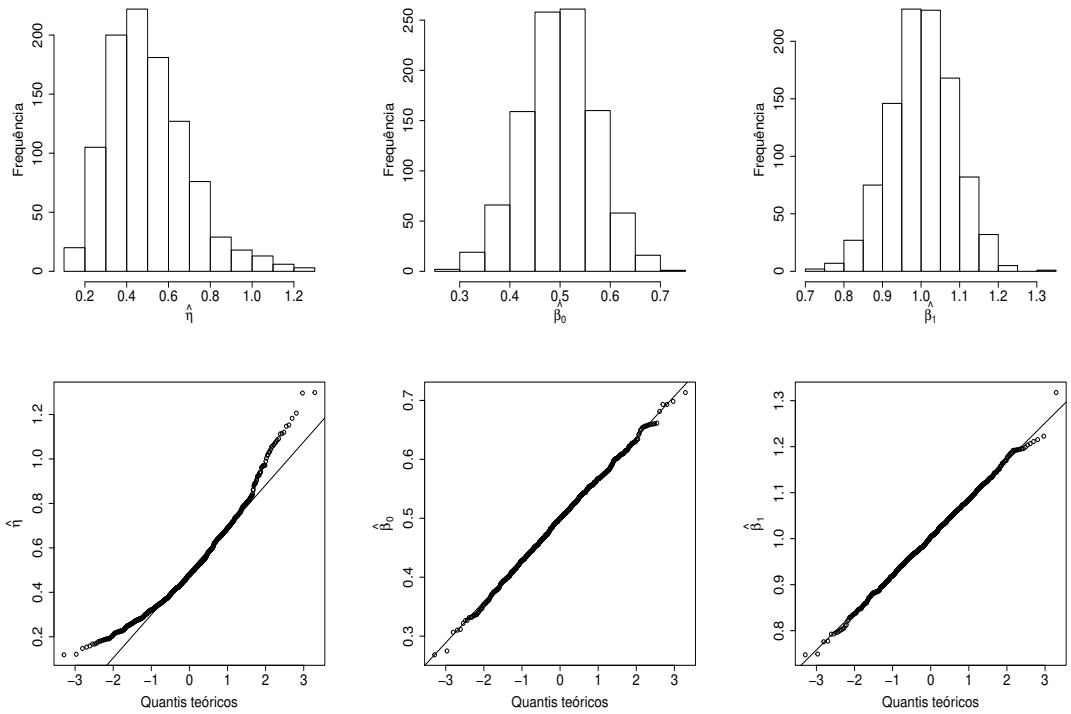
(a) $n = 200$



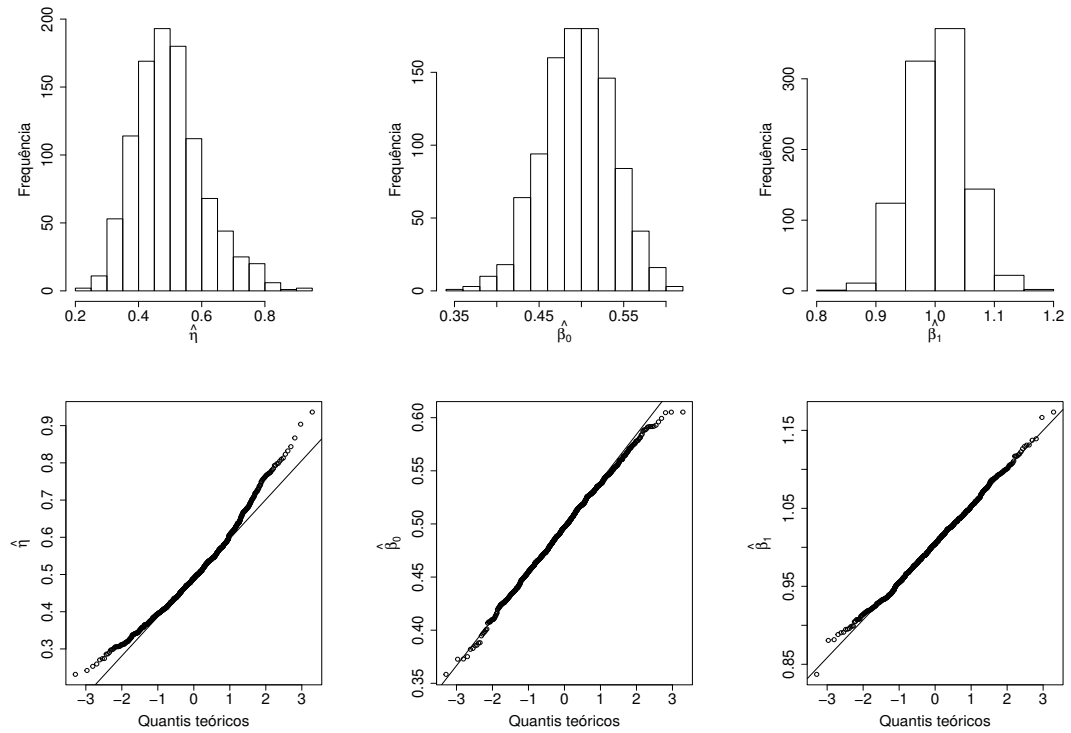
(b) $n = 600$

Figura 25 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo de regressão hP com regressores em θ e $\eta = 1$.

Fonte: Elaborada pelo autor.



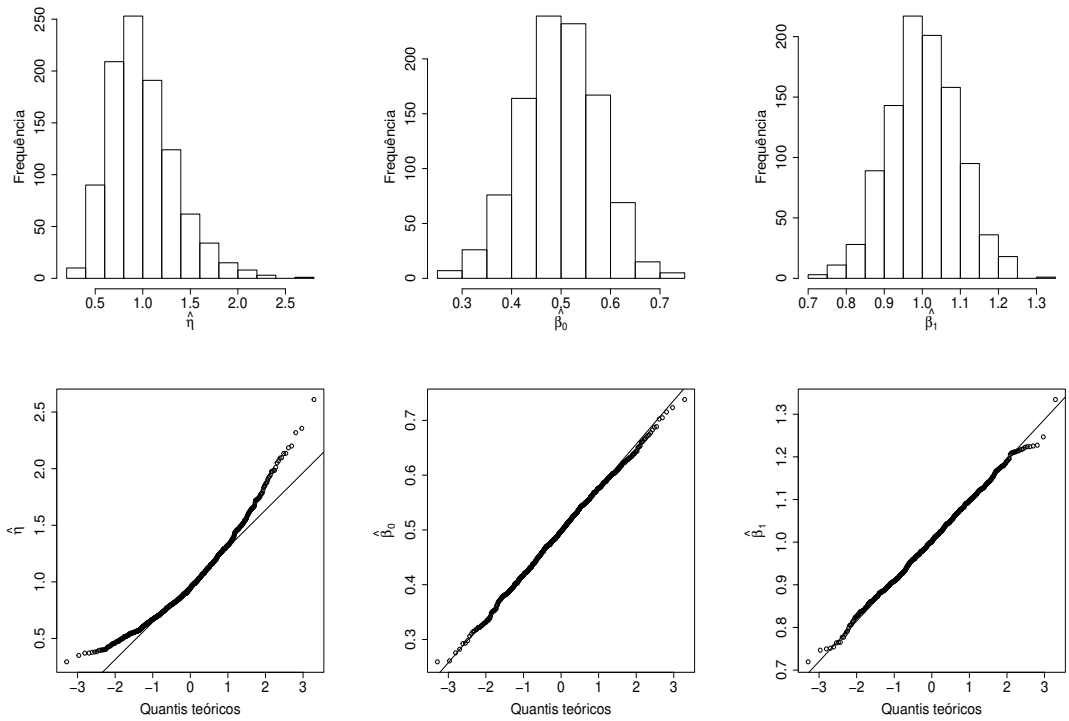
(a) $n = 200$



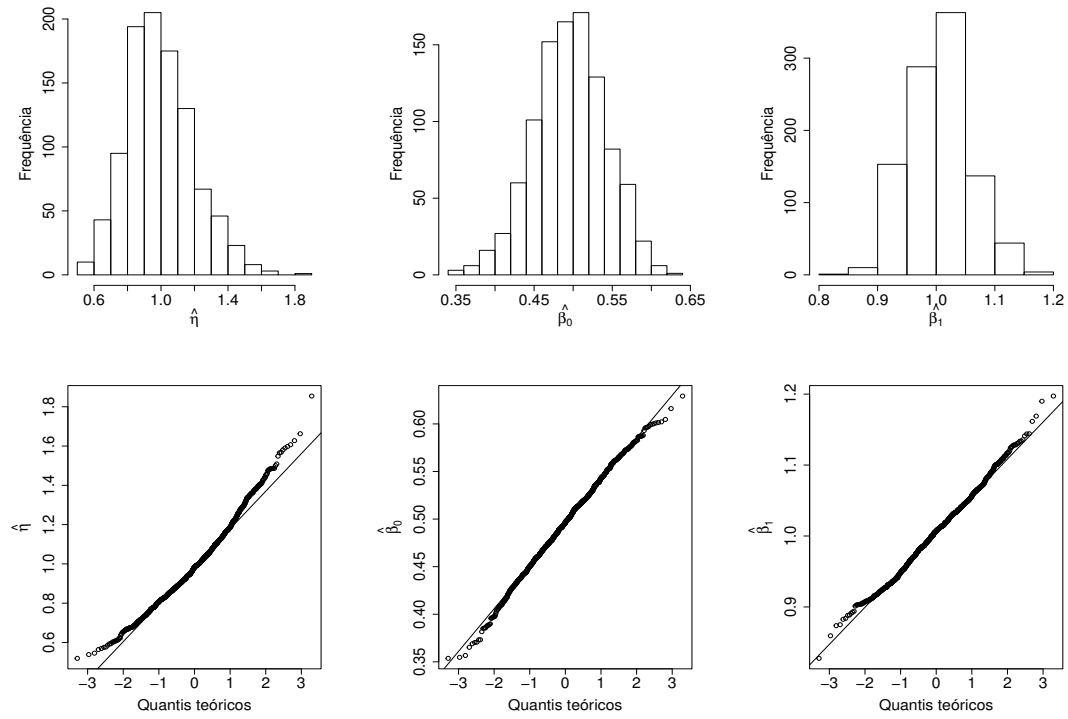
(b) $n = 600$

Figura 26 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo de regressão hP com regressores na média μ e $\eta = 0,5$.

Fonte: Elaborada pelo autor.



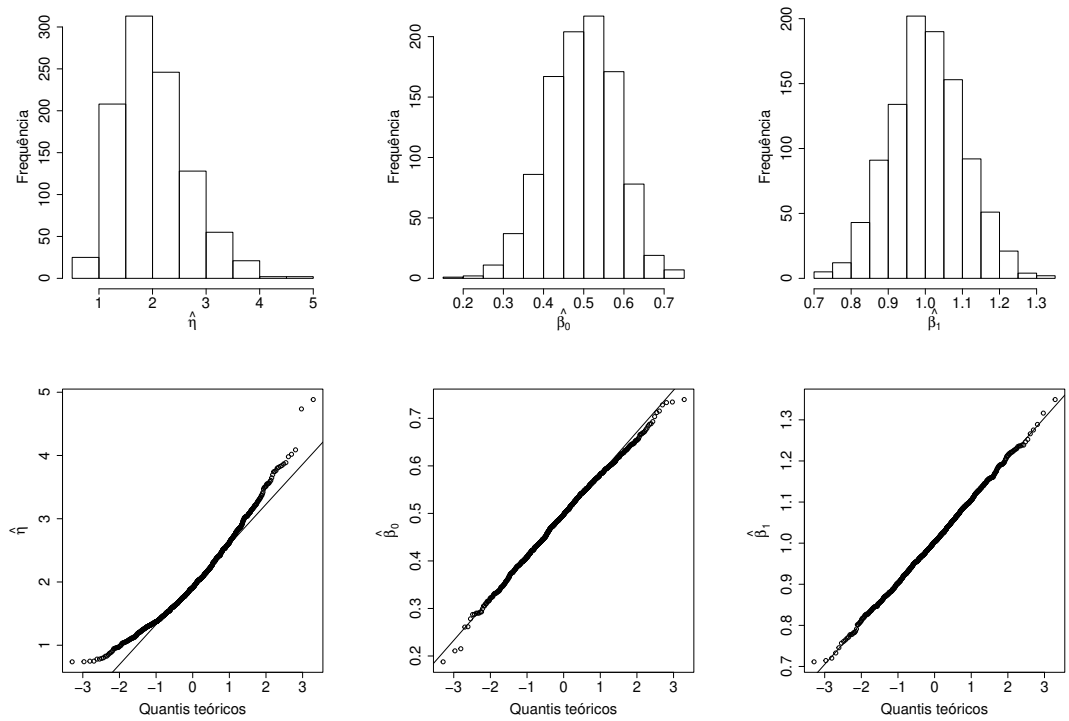
(a) $n = 200$



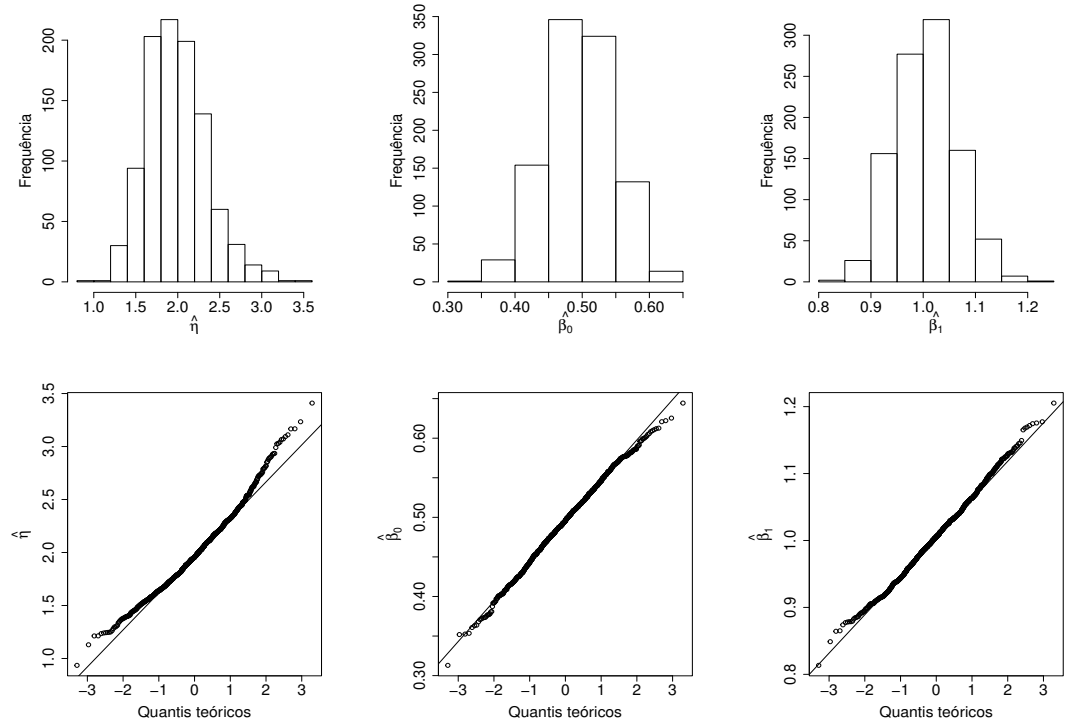
(b) $n = 600$

Figura 27 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo de regressão hP com regressores na média μ e $\eta = 1$.

Fonte: Elaborada pelo autor.



(a) $n = 200$



(b) $n = 600$

Figura 28 – Histogramas e gráficos QQ-normal dos EMVs dos parâmetros do modelo de regressão hP com regressores na média μ e $\eta = 2$.

Fonte: Elaborada pelo autor.

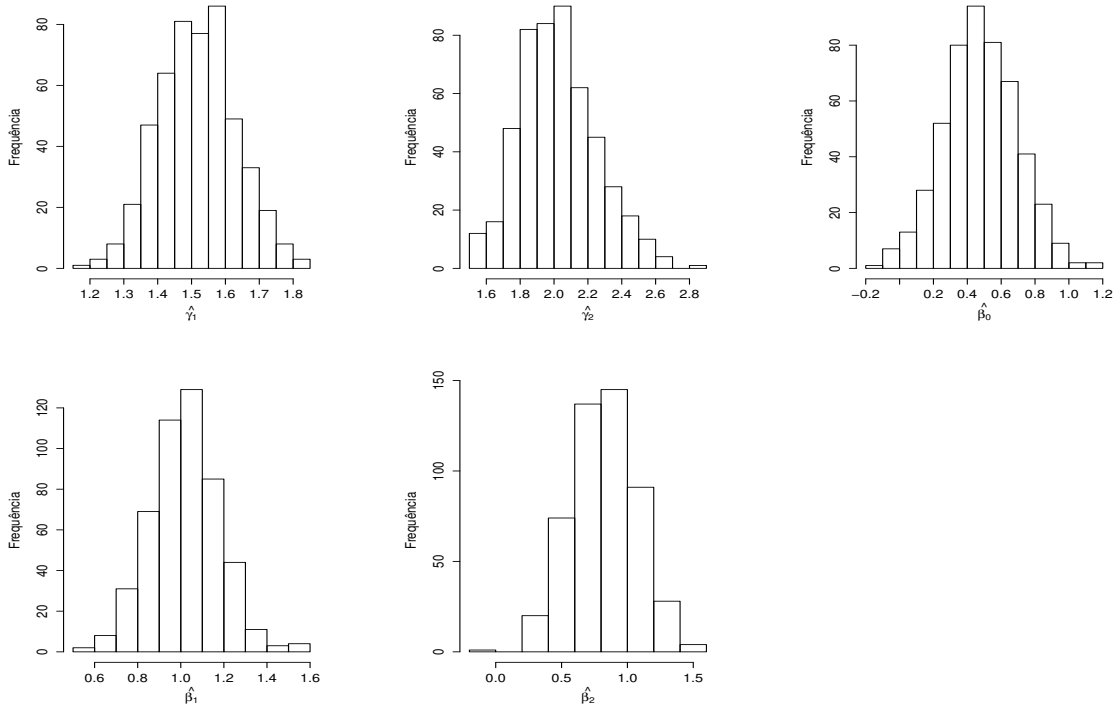
APÊNDICE

C

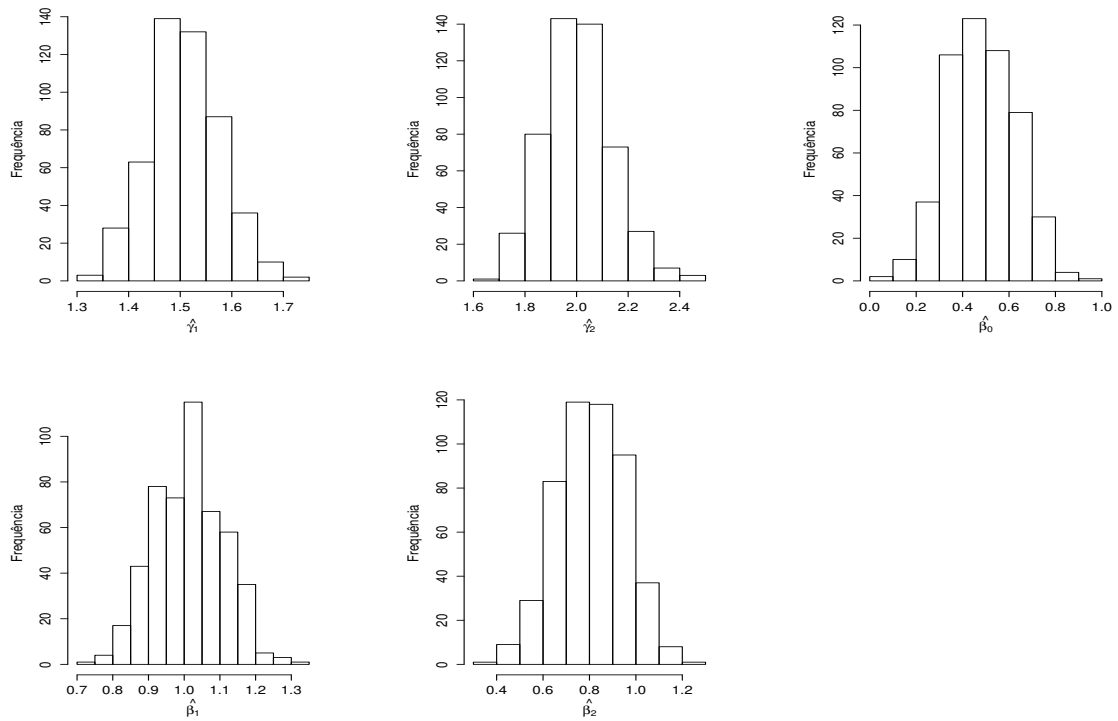
HISTOGRAMAS E GRÁFICOS QQ-NORMAL DOS PARÂMETROS ESTIMADOS PARA O MODELO DE FRAÇÃO DE CURA HIPER-POISSON

Nesta seção apresentamos primeiramente alguns dos histogramas e dos gráficos QQ-normal dos parâmetros estimados com base nas 500 replicações de Monte Carlos para o modelo proposto no Capítulo 3 quando a técnica da verossimilhança perfilada foi empregada.

Na sequência apresentamos os histogramas e os gráficos QQ-normal dos parâmetros para o modelo com fração de cura hP com o parâmetro de dispersão proveniente da variável de fragilidade fixado.



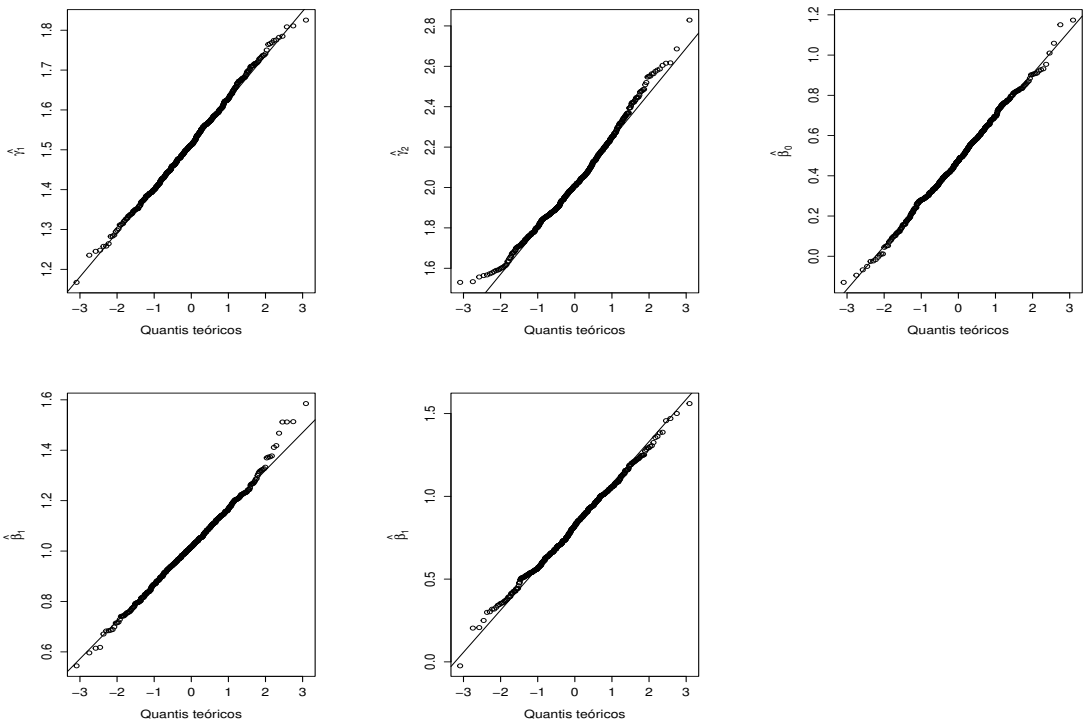
(a) $n = 200$



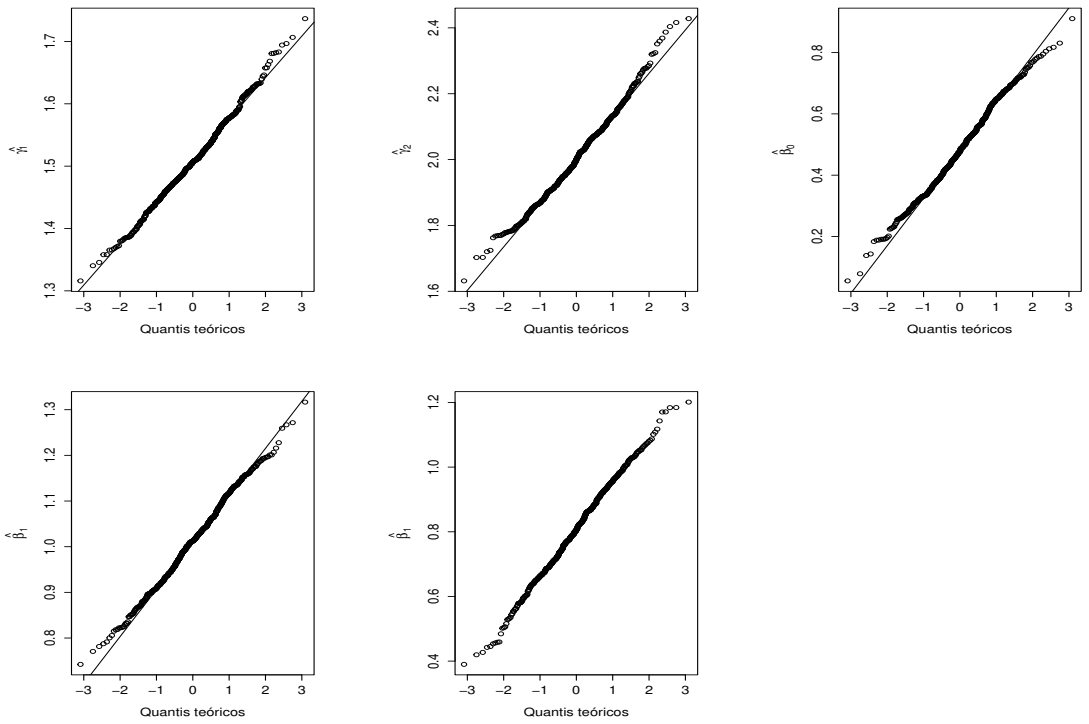
(b) $n = 600$

Figura 29 – Histogramas das estimativas dos parâmetros do modelo de fração de cura hP baseados em 500 replicações e η estimado pela verossimilhança perfilada.

Fonte: Elaborada pelo autor.



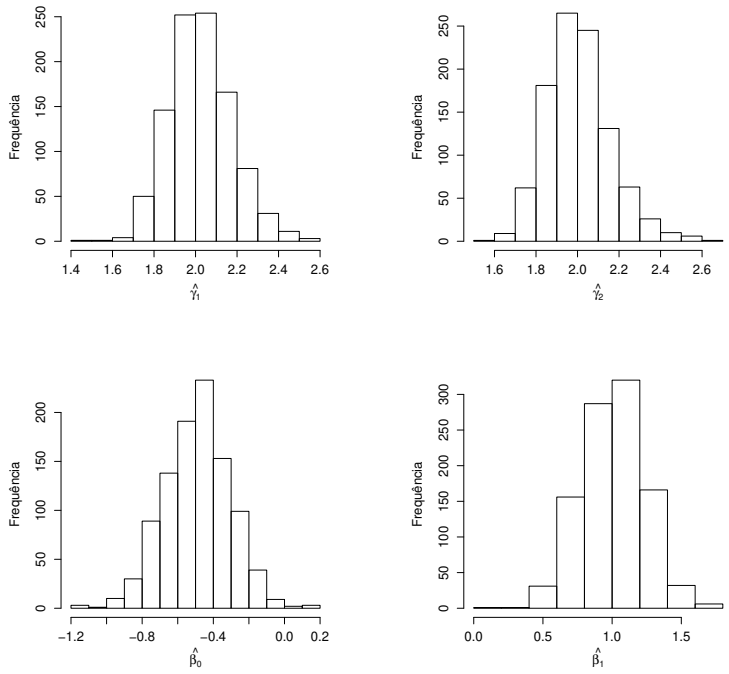
(a) $n = 200$



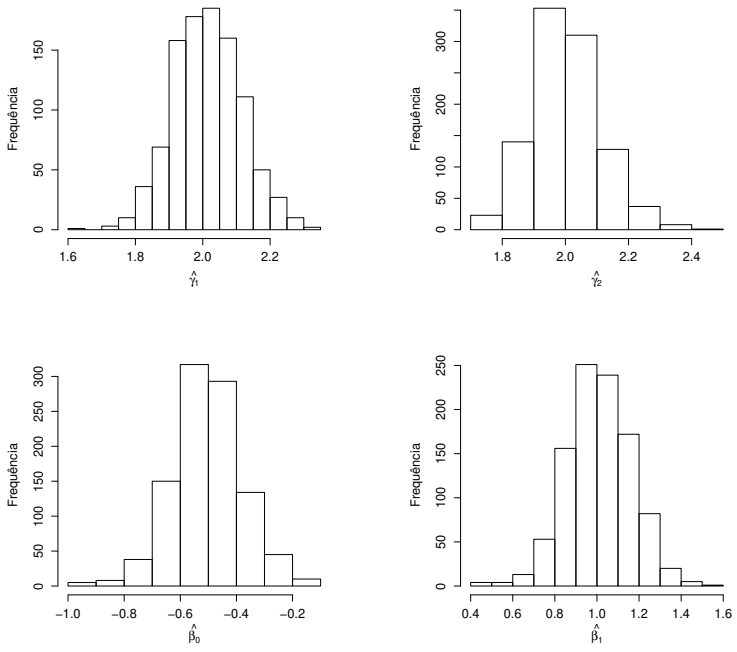
(b) $n = 600$

Figura 30 – Gráficos QQ-normal das estimativas dos parâmetros do modelo de fração de cura hP baseados em 500 replicações e η estimado pela verossimilhança perfilada.

Fonte: Elaborada pelo autor.



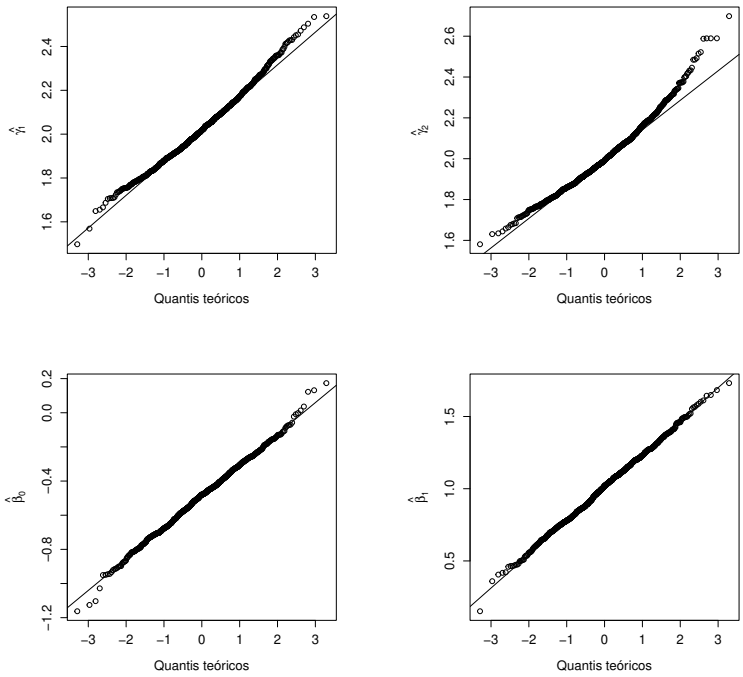
(a) $n = 200$



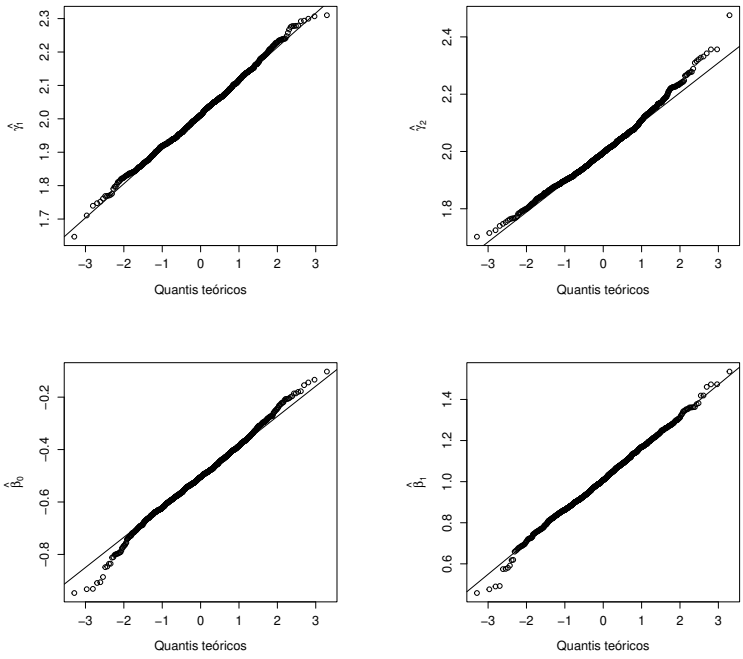
(b) $n = 400$

Figura 31 – Histogramas das estimativas dos parâmetros do modelo de fração de cura hP com $\eta = 0,5$ baseados em 1000 replicações.

Fonte: Elaborada pelo autor.



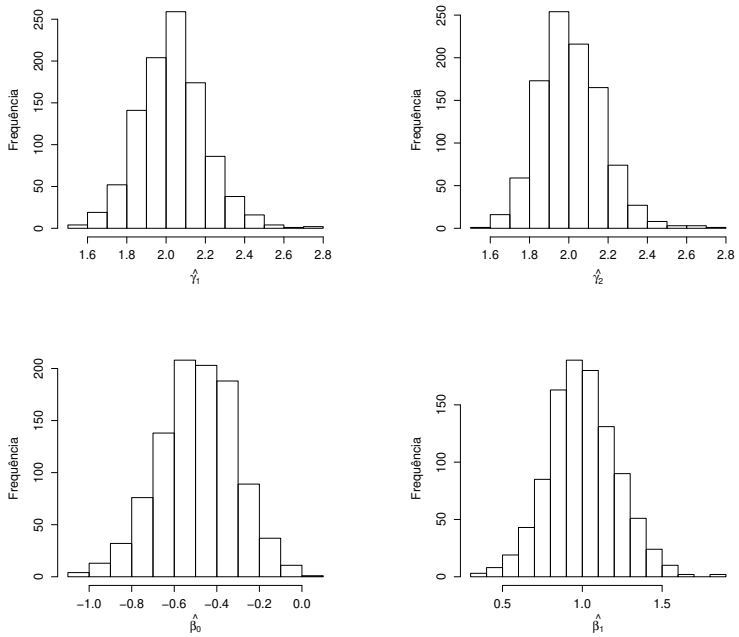
(a) $n = 200$



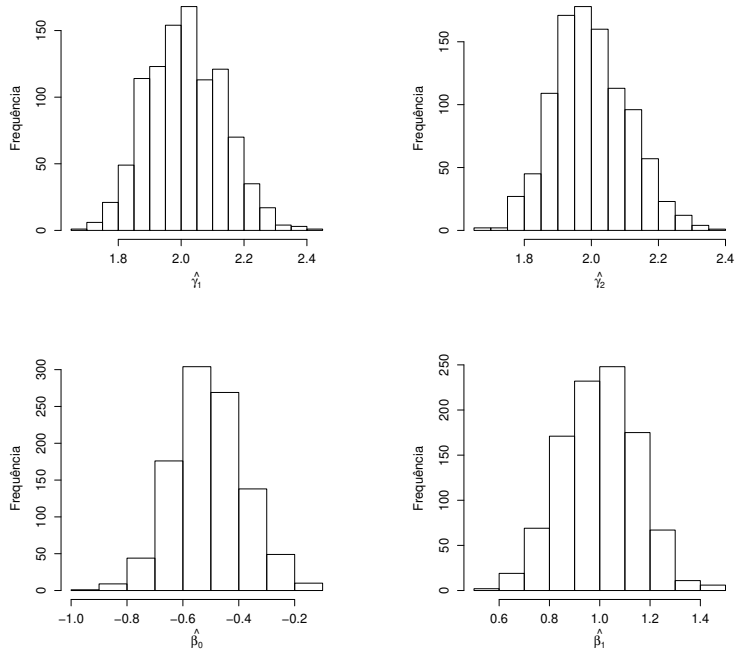
(b) $n = 400$

Figura 32 – Gráficos QQ-normal das estimativas dos parâmetros do modelo de fração de cura hP com $\eta = 0,5$.

Fonte: Elaborada pelo autor.



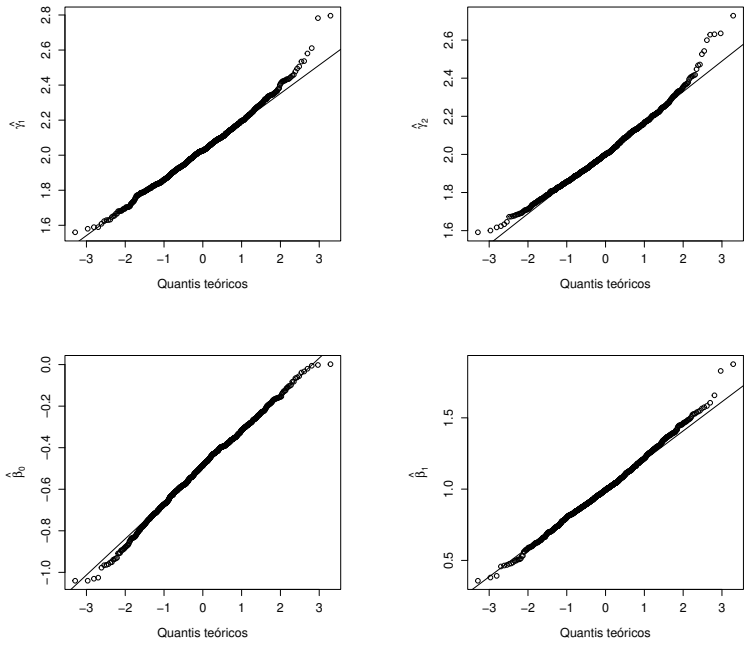
(a) $n = 200$



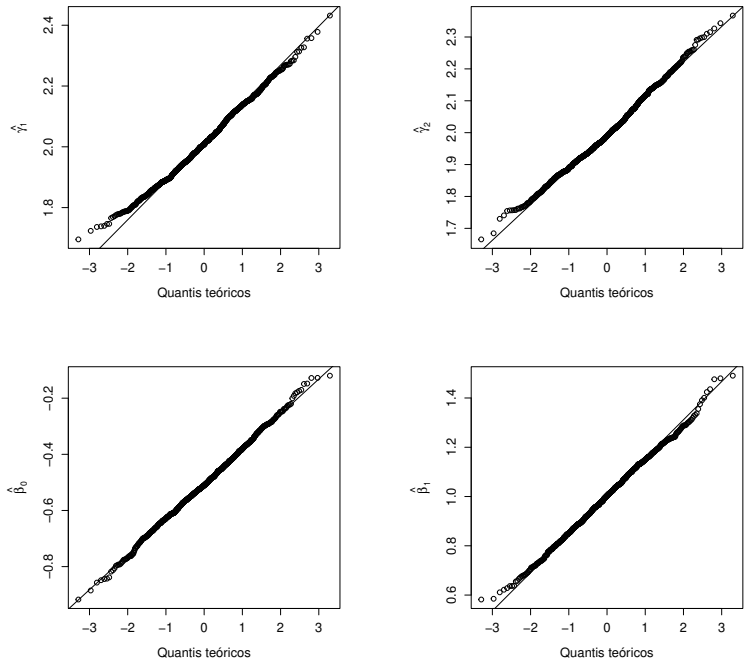
(b) $n = 400$

Figura 33 – Histogramas das estimativas dos parâmetros do modelo de fração de cura hP com $\eta = 1$ baseados em 1000 replicações.

Fonte: Elaborada pelo autor.



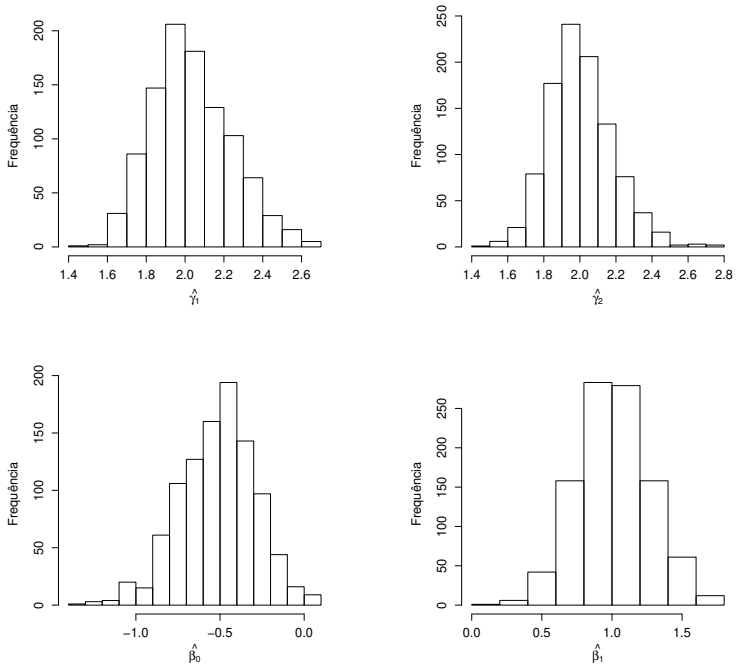
(a) $n = 200$



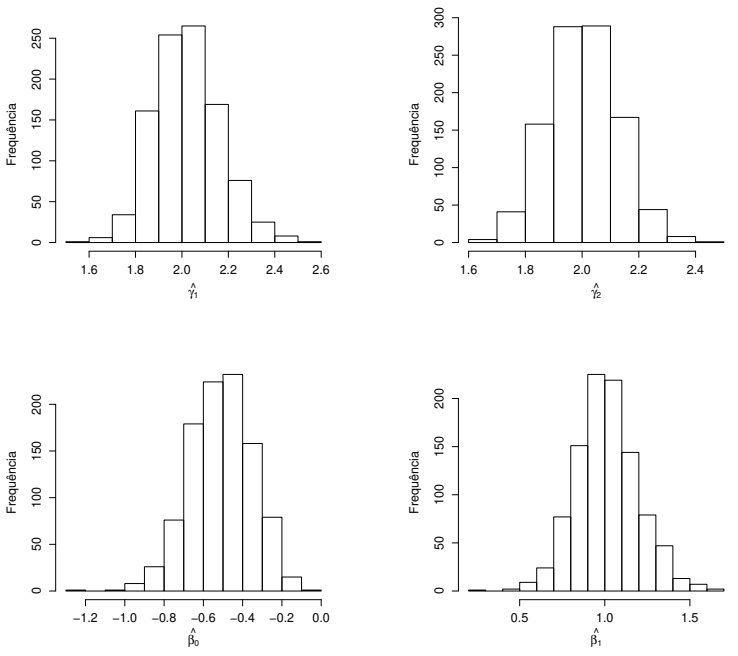
(b) $n = 400$

Figura 34 – Gráficos QQ-normal das estimativas dos parâmetros do modelo de fração de cura hP com $\eta = 1$.

Fonte: Elaborada pelo autor.



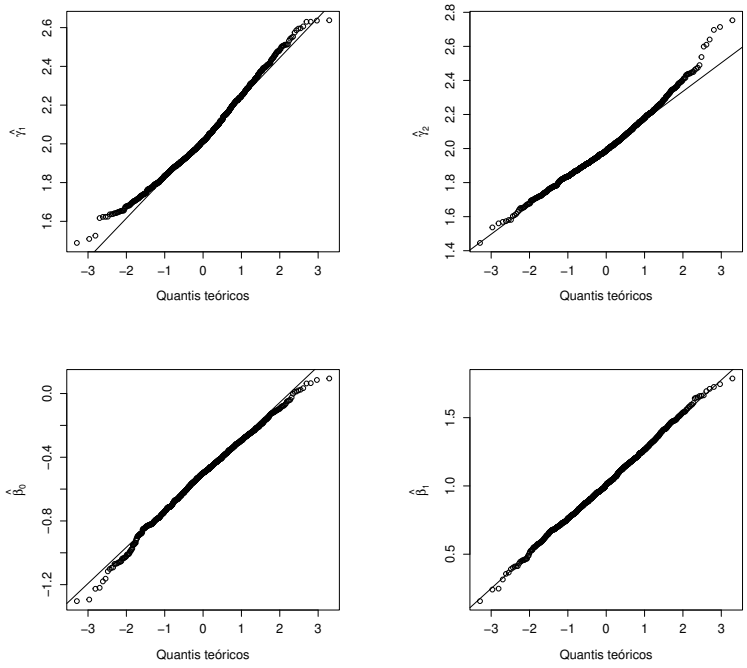
(a) $n = 200$



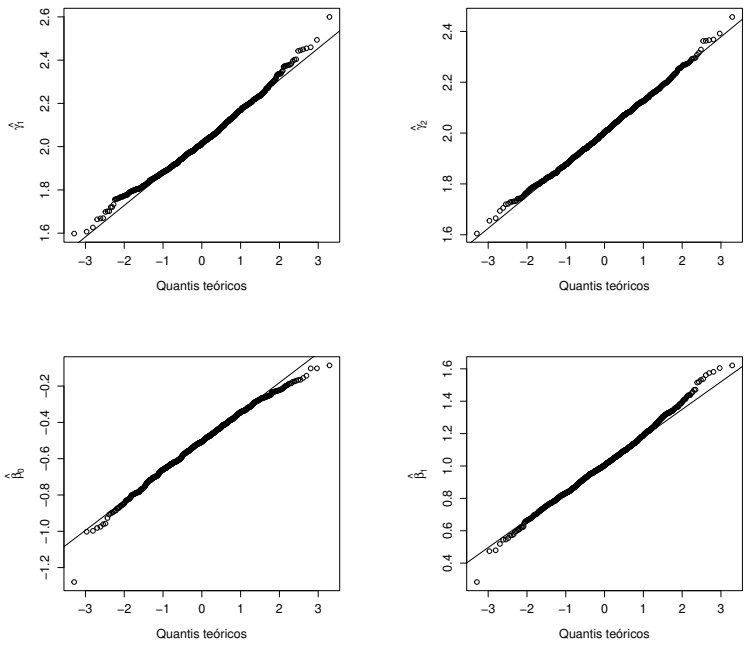
(b) $n = 400$

Figura 35 – Histogramas das estimativas dos parâmetros do modelo de fração de cura hP com $\eta = 2$ baseados em 1000 replicações.

Fonte: Elaborada pelo autor.



(a) $n = 200$



(b) $n = 400$

Figura 36 – Gráficos QQ-normal das estimativas dos parâmetros do modelo de fração de cura hP com $\eta = 2$.

Fonte: Elaborada pelo autor.


ICMC
USP
SÃO CARLOS