

Universidade Federal de São Carlos  
Centro de Ciências Exatas e de Tecnologia  
Departamento de Estatística

**UMA FAMÍLIA DE MODELOS DE REGRESSÃO**  
**COM A DISTRIBUIÇÃO ORIGINAL DA VARIÁVEL RESPOSTA**

Marcelo de Paula  
Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de São Carlos PPGes / UFSCar, como parte dos requisitos necessários para obtenção do título de Doutor em Estatística.

UFSCar - São Carlos  
Abril/2013

**Ficha catalográfica elaborada pelo DePT da  
Biblioteca Comunitária/UFSCar**

P324fm

Paula, Marcelo de.

Uma família de modelos de regressão com a distribuição original da variável resposta / Marcelo de Paula. -- São Carlos : UFSCar, 2013.

106 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2013.

1. Análise de regressão. 2. Modelos lineares (estatística).  
3. Variável resposta de origem. I. Título.

CDD: 519.536 (20<sup>a</sup>)



**UNIVERSIDADE FEDERAL DE SÃO CARLOS**  
Centro de Ciências Exatas e de Tecnologia  
Programa de Pós-Graduação em Estatística  
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40  
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br  
13565-905 - SÃO CARLOS-SP - BRASIL

---

## FOLHA DE APROVAÇÃO

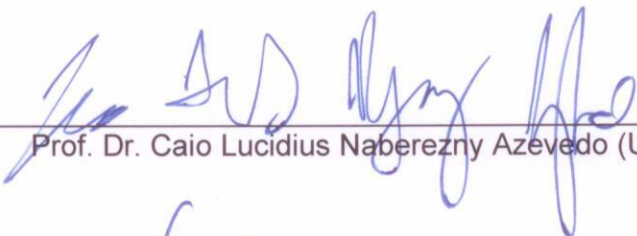
**Aluno(a) : Marcelo de Paula**

TESE DE DOUTORADO DEFENDIDA E APROVADA EM 05/04/2013 PELA  
COMISSÃO JULGADORA:

Presidente

  
Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar/Orientador)


1º Examinador

  
Prof. Dr. Caio Lucidius Naberezny Azevedo (UNICAMP)

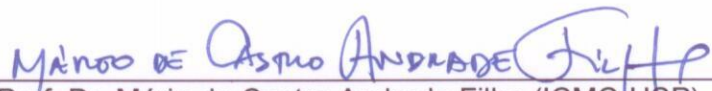
2º Examinador

  
Prof. Dr. Enrico Antônio Colosimo (UFMG)

3º Examinador

  
Prof. Dr. José Galvão Leite (DEs-UFSCar)

4º Examinador

  
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP)

---

# Agradecimentos

Agradeço

A Deus que me protege e me guia em direção ao bem.

Ao meu pai José de Aparecido de Paula que, apesar de pouca instrução, jamais foi de encontro as minhas aspirações.

A minha mãe Antônia Aparecida Pelicieri de Paula, cujo amor é incondicional.

Ao meu irmão Márcio de Paula que sempre acreditou em minha capacidade e jamais duvidou dos meus sonhos.

Ao meu irmão Luíz Carlos de Paula pela compreensão e amizade ao longo deste trabalho.

Ao meu primo Alberto Luis Gaspar, que sempre me inspirou com seu bom humor e otimismo perante a vida.

Ao meu amigo Nelinho, cuja amizade e companheirismo são especiais e eternos.

A minha amiga Elisabeth Regina de Toledo, a quem tenho como uma querida irmã.

Ao professor Dr. Carlos Alberto Ribeiro Diniz pela orientação e pelas ideias durante todo este trabalho.

Aos professores Dr. Caio Lucidius Naberezny Azevedo, Dr. José Galvão Leite, Dr. Mário de Castro, pelas ideias, sugestões e correções propostas no exame de qualificação e na defesa da tese.

Ao professor Dr. Enrico Colosimo pela importante contribuição na defesa da tese.

Aos colegas, professores e funcionários do Departamento de Estatística da UFSCar, pela grande amizade.

# Resumo

É sabido que a área de modelagem estatística por regressão sofreu um grande impulso desde o desenvolvimento dos modelos lineares generalizados (MLGs) no início da década de 70 do século XX, propostos por Nelder e Wedderburn (1972). A teoria dos MLGs pode ser interpretada como uma generalização do modelo de regressão linear tradicional, em que a variável resposta não precisa necessariamente assumir a distribuição normal, e sim, qualquer distribuição pertencente à família exponencial de distribuições.

Em algumas situações, porém, a distribuição da variável resposta é originalmente fruto de uma outra distribuição discreta ou contínua, ou seja, a variável resposta tem uma distribuição original que não é a usualmente considerada. Um exemplo desta situação é a dicotomização de uma variável discreta ou contínua por meio de um ponto de corte arbitrário. Além disso, a variável resposta pode estar relacionada, de alguma forma, com uma outra variável de interesse.

Nesse trabalho propomos uma família de modelos de regressão com a informação da variável resposta original, cuja distribuição de probabilidades ou função densidade de probabilidade pertence à família exponencial. O modelo de regressão logística com resposta normal e log-normal desenvolvido por Suissa e Blais (1995) é apresentado como caso particular dos modelos de regressão com resposta de origem. Para a resposta de origem normal consideramos os modelos logístico, exponencial, geométrico, Poisson e log-normal. Para a resposta de origem exponencial consideramos os modelos logístico, normal, geométrico, Poisson e log-normal. Em contribuição ao trabalho de Suissa e Blais atribuímos duas respostas discretas ao modelo logístico, geométrico e de Poisson, e também consideramos uma resposta contínua normal com estrutura heteroscedástica. Adicionalmente, propomos também o modelo logístico com resposta pertencente à classe de distribuições séries de potências inflacionadas considerando o caso particular da resposta geométrica zero inflacionada.

Realizamos vários estudos com dados artificiais comparando o modelo de regressão proposto com a informação da distribuição de origem e o modelo de regressão usual. Dois conjuntos de dados reais também são considerados. Assumindo uma distribuição corretamente especificada, o modelo produz estimativas de máxima verossimilhança mais eficientes e estimativas intervalares mais precisas para os coeficientes de regressão.

**Palavras-chave:** Modelos de regressão, Modelos lineares generalizados, Variável resposta de origem.

# Abstract

We know that statistic modeling by regression had a stronger impulse since generalized linear models (GLMs) development in 70 decade beginning of the XX century, proposed by Nelder e Wedderburn (1972). GLMs theory can be interpret like a traditional linear regression model generalization, where outcomes don't need necessary to assume a normal distribution, that is, any distribution belong to exponential distributions family.

In binary logistic regression case, however, in many practice situations the outcomes response is originally from a discrete or continuous distribution, that is, the outcomes response has an original distribution that is not Bernoulli distribution and, although, because some purpose this variable was later dicothomized by an arbitrary cut off point  $C$ .

In this work we propose a regression models family with original outcomes information, whose probability distribution or density function probability belong to exponential family. We present the models construction and development to each class, incorporating the original distribution outcomes response information. The proposed models are an extension of Suissa (1991) and Suissa and Blais (1995) works which present methods of estimating the risk of an event defined in a sample subspace of a continuous outcome variable. Simulation studies are presented in order to illustrate the performance of the developed methodology. For original normal outcomes we considered logistic, exponential, geometric, Poisson and lognormal models. For original exponential outcomes we considered logistic, normal, geometric, Poisson and lognormal models. In contribution to Suissa and Blais (1995) works we attribute two discrete outcomes for binary model, geometric and Poisson, and we also considered a normal distributions with multiplicative heteroscedastic structures continuous outcomes. In supplement we also propose the binary model with inflated power series distributions outcomes considering a sample subspace of a zero inflated geometric outcomes.

We do several artificial data studies comparing the model of original distribution information regression model with usual regression model. Simulation studies are presented in order to illustrate the performance of the developed methodology. A real data set is analyzed by using the proposed models. Assuming a correct specified distribution, the incorporation of this information about outcome response in the model produces more efficient likelihood estimates.

**Keywords:** Regression models, generalized linear models, original distribution.

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Objetivos do trabalho . . . . .	3
1.2	Organização do Trabalho . . . . .	4
<b>2</b>	<b>Uma família de modelos de regressão com a informação da distribuição original da variável resposta</b>	<b>5</b>
2.1	Introdução . . . . .	5
2.2	Informação da distribuição da variável resposta . . . . .	5
2.3	Modelos da família exponencial com resposta de origem normal . . . . .	6
2.3.1	O caso particular de Suissa e Blais (1995): Modelo de regressão logística com resposta normal . . . . .	7
2.3.2	Modelo exponencial com resposta normal . . . . .	8
2.3.3	Modelo geométrico com resposta normal . . . . .	9
2.3.4	Modelo Poisson com resposta normal . . . . .	10
2.3.5	Modelo log-normal com resposta normal . . . . .	11
2.3.6	Função de verossimilhança . . . . .	11
2.4	Modelos da família exponencial com resposta exponencial . . . . .	12
2.4.1	Modelo logístico com resposta exponencial . . . . .	13
2.4.2	Modelo normal com resposta exponencial . . . . .	14
2.4.3	Modelo geométrico com resposta exponencial . . . . .	15
2.4.4	Modelo de Poisson com resposta exponencial . . . . .	16
2.4.5	Modelo log-normal com resposta exponencial . . . . .	18
2.4.6	Função de verossimilhança para os modelos propostos com resposta exponencial . . . . .	19
2.5	O modelo de regressão logística com distribuições diversas para a variável resposta . . . . .	19
2.5.1	Modelo de regressão logística com resposta geométrica . . . . .	19
2.5.2	Escores para o modelo logístico com resposta geométrica . . . . .	21
2.5.3	Modelo de regressão logística com resposta Poisson . . . . .	22
2.5.4	Escores para o modelo de regressão logística com resposta Poisson . . . . .	24
2.5.5	Modelo de regressão logística com resposta normal em uma estrutura heteroscedástica . . . . .	25
2.5.6	Escores do modelo logístico com resposta normal considerando uma estrutura heteroscedástica . . . . .	28
2.6	Diagnósticos para o modelo com resposta de origem . . . . .	29
2.6.1	Resíduos padronizados para a variável de interesse . . . . .	30
2.6.2	Pontos influentes . . . . .	31
2.6.3	Teste de hipóteses para os coeficientes de regressão dos modelos com resposta de origem normal . . . . .	32
<b>3</b>	<b>Ilustração com dados artificiais</b>	<b>33</b>
3.1	Introdução . . . . .	33
3.2	Uma revisão sobre o estudo de simulação de Suissa e Blais (1995) . . . . .	33

3.3	Estudo de simulação para o modelo de regressão logística segundo a família de regressão proposta . . . . .	33
3.4	Modelo de regressão logística com resposta normal . . . . .	34
3.5	Modelo de regressão logística com resposta log-normal . . . . .	36
3.6	Modelo de regressão logística com resposta normal em uma estrutura de heteroscedasticidade . . . . .	38
3.7	Modelo de regressão logística com resposta exponencial . . . . .	40
3.8	Modelo de regressão logística com resposta geométrica . . . . .	41
3.9	Modelo de regressão logística com resposta Poisson . . . . .	43
3.10	Modelo geométrico com resposta Normal . . . . .	44
3.11	Modelo geométrico com resposta exponencial . . . . .	46
3.12	Análise de resíduos da variável de interesse binária . . . . .	47
3.13	Análise de influência considerando os dados artificiais . . . . .	48
3.14	Análise de influência considerando uma amostra particular . . . . .	49
<b>4</b>	<b>Modelo de regressão logística com resposta original pertencente à classe de distribuições série de potências inflacionadas</b>	<b>52</b>
4.1	Introdução . . . . .	52
4.2	Classe de distribuições série de potências . . . . .	52
4.3	Função de verossimilhança . . . . .	52
4.4	Modelo de regressão logística com resposta original pertencente à classe de distribuições série de potências . . . . .	53
4.5	Classe de distribuições série de potências inflacionadas . . . . .	54
4.5.1	Introdução . . . . .	54
4.5.2	Excesso de valores em dados de contagens . . . . .	54
4.5.3	Excesso de zero em dados de contagem . . . . .	54
4.5.4	O modelo estatístico inflacionado no ponto $s$ . . . . .	55
4.6	Modelo de regressão logística com resposta pertencente à família de distribuições série de potências inflacionadas . . . . .	55
4.7	Caso particular: o modelo geométrico para dados inflacionados de zeros . . .	56
4.8	Modelo de regressão logística com resposta geométrica inflacionado de zeros .	57
<b>5</b>	<b>Abordagem bayesiana</b>	<b>59</b>
5.1	Introdução . . . . .	59
5.2	Função de verossimilhança para os modelos de regressão com resposta normal	59
5.3	Modelo de regressão logística bayesiano com a informação da variável resposta normal . . . . .	60
5.4	Função de verossimilhança para os modelos de regressão com resposta exponencial . . . . .	61
5.5	Modelo de regressão logística bayesiano com a informação da variável resposta exponencial . . . . .	61
5.6	Modelo logístico bayesiano com a informação da variável resposta normal heteroscedástica multiplicativa . . . . .	62



5.7	Modelo de regressão logística bayesiano com a informação da variável resposta geométrica . . . . .	64
5.8	Modelo de regressão logística bayesiano com informação da variável resposta Poisson . . . . .	65
5.9	Modelo logístico bayesiano com resposta geométrica inflacionada em zero . .	65
5.10	Simulação para os modelos propostos . . . . .	67
5.11	Resultados obtidos para o modelo logístico bayesiano com resposta normal .	67
5.12	Resultados obtidos para o modelo logístico bayesiano com resposta log-normal	69
5.13	Resultados obtidos para o modelo logístico bayesiano com resposta exponencial	70
5.14	Resultados obtidos com o modelo logístico bayesiano com resposta geométrica	71
5.15	Resultados obtidos para o modelo logístico bayesiano com resposta Poisson .	73
5.16	Resultados obtidos para o modelo geométrico bayesiano com resposta normal	74
5.17	Resultados obtidos para o modelo geométrico bayesiano com resposta exponencial . . . . .	75
<b>6</b>	<b>Aplicações a dados reais</b>	<b>77</b>
6.1	Introdução . . . . .	77
6.2	Aplicação do modelo de regressão logística com resposta exponencial . . . .	77
6.3	Aplicação do modelo logístico com resposta geométrica inflacionada de zeros	81
<b>7</b>	<b>Discussões e considerações finais</b>	<b>86</b>
7.1	Sobre a metodologia proposta e ganhos inferenciais . . . . .	86
7.2	Sobre as probabilidades de cobertura dos intervalos de confiança dos parâmetros	86
7.3	Sobre a análise de resíduos nas probabilidades de sucesso . . . . .	86
7.4	Sobre a análise de influência . . . . .	87
7.5	Sobre o enfoque bayesiano dos modelos propostos . . . . .	87
7.6	Perspectivas do trabalho e possíveis desdobramentos . . . . .	87
<b>8</b>	<b>Apêndice A: Resultados usados para as funções score do modelo logístico com respostas diversas</b>	<b>88</b>
8.1	Resultados usados para o modelo logístico com resposta normal . . . . .	88
8.1.1	Derivadas de primeira ordem . . . . .	88
8.1.2	Derivadas de segunda ordem . . . . .	88
8.2	Resultados usados para o modelo logístico com resposta exponencial . . . . .	89
8.2.1	Derivadas de primeira ordem . . . . .	89
8.2.2	Funções derivadas de segunda ordem . . . . .	89
8.3	Resultados usados para o modelo logístico com resposta geométrica . . . . .	90
8.3.1	Derivadas de primeira ordem . . . . .	90
8.3.2	Derivadas de segunda ordem . . . . .	90
8.4	Resultados usados para o modelo logístico com resposta Poisson . . . . .	91
8.4.1	Derivadas de primeira ordem . . . . .	91
8.4.2	Derivadas de segunda ordem . . . . .	91
8.5	Resultados usados para o modelo logístico com resposta Normal em uma estrutura de heteroscedasticidade multiplicativa . . . . .	91
8.5.1	Derivadas de primeira ordem . . . . .	91

8.5.2	Derivadas de segunda ordem . . . . .	92
<b>9</b>	<b>Apêndice B: Histogramas das estimativas bayesianas</b>	<b>94</b>
9.1	Modelo logístico com resposta normal . . . . .	94
9.2	Modelo logístico com resposta lognormal . . . . .	95
9.3	Modelo logístico com resposta exponencial . . . . .	96
9.4	Modelo logístico com resposta geométrica . . . . .	97
9.5	Modelo logístico com resposta Poisson . . . . .	99
9.6	Modelo geométrico com resposta normal . . . . .	100
9.7	Modelo geométrico com resposta exponencial . . . . .	101
<b>10</b>	<b>Apêndice C: Programas desenvolvidos</b>	<b>103</b>
<b>11</b>	<b>Referências Bibliográficas</b>	<b>105</b>

# 1 Introdução

Os modelos de regressão são normalmente utilizados para estudar e estabelecer uma relação entre uma variável de interesse, denominada variável resposta ou variável dependente, e um conjunto de fatores ou atributos, chamados de variáveis de entrada, variáveis explicativas, variáveis preditoras, variáveis explanatórias ou covariáveis. Além disso, a área de modelagem estatística por regressão sofreu um grande impulso desde o desenvolvimento dos modelos lineares generalizados no início da década de 70 do século XX, propostos por Nelder e Wedderburn (1972). A teoria dos MLGs pode ser interpretada como uma generalização do modelo de regressão linear tradicional, em que a variável resposta não precisa necessariamente assumir a distribuição normal, e sim, qualquer distribuição pertencente à família exponencial de distribuições.

Por exemplo, o modelo de regressão logística binária, que pertence à classe dos MLGs, é indicada quando a variável resposta de interesse é dicotômica, isto é, quando a variável resposta assume apenas dois valores possíveis e a regressão logística politômica quando a variável resposta assume diversas categorias possíveis. Essa relação determina a probabilidade de ocorrência de um evento em presença de um conjunto de variáveis explicativas, formando um modelo preditivo indutivo. Para o caso da regressão logística com variável dependente dicotômica, modelamos a probabilidade de resposta de uma das duas categorias, isto é, a probabilidade de sucesso, em função das variáveis explanatórias.

Embora a regressão logística seja conhecida desde os anos 50 do século XX, ela tornou-se mais usual por meio de Cox (1989) e de Hosmer e Lemeshow (2000). Diversos aspectos teóricos da regressão logística são amplamente discutidos na literatura, destacando-se Kleinbaum (1994), Agresti (1990), Hosmer e Lemeshow (2000), Cox e Snell (1989) e Kleinbaum e Klein (2002).

Porém, em muitas situações práticas pode ocorrer que a variável resposta binária tenha uma distribuição original pertencente a alguma classe de distribuições, sejam elas discretas ou contínuas. Em outras palavras, a variável resposta original  $R_i$  tem uma distribuição que não é a de Bernoulli e, por algum motivo, tal variável foi posteriormente dicotomizada, considerando um ponto de corte  $C_R$ , pertencente ao suporte da distribuição de  $R_i$ . Dessa forma, a variável resposta original  $R_i$  se configura em dois eventos distintos:  $R_i > C_R$  e  $R_i \leq C_R$ ,  $i = 1, 2, \dots, n$ .

Por exemplo, a hipertensão tem sido definida em indivíduos com uma pressão sanguínea diastólica maior que 90 *mmHg*. Uma função renal anormal tem sido definida em indivíduos com creatinina maior do que 1,4 *mg/dl*. Em análise de dados epidemiológicos a variável resposta contínua é frequentemente dicotomizada a fim de permitir a estimação do risco da doença e, conseqüentemente, métodos estatísticos padrão para dados binários são usados. O índice de massa corporal (IMC), que se trata de uma variável aleatória contínua, é utilizado para definir obesidade. O indivíduo é considerado portador de obesidade mórbida caso ele apresente um IMC acima de 30 e, portanto, para fins de diagnóstico tal variável é dicotomizada por meio da constante  $C_R = 30$ .

Na área de finanças o cliente tem um prazo de 60 dias para realizar o pagamento mínimo da fatura do cartão de crédito. Caso este prazo seja ultrapassado, o cliente já é considerado inadimplente. Assim, o número de dias entre o pagamento da fatura do cartão

de crédito e seu vencimento é uma variável aleatória discreta que foi dicotomizada por meio da constante  $C_R = 60$  dias.

No setor industrial o número de defeitos de um determinado produto, ou lote fabricado, em uma linha de produção, segue uma variável aleatória discreta. Em geral, basta que o item ou lote apresente um único defeito para que o mesmo seja considerado impróprio para o consumo. Ou seja, uma variável aleatória discreta que foi dicotomizada com  $C_R = 0$  defeito.

Nesse sentido, pode ser de interesse incorporar a informação sobre a distribuição original da variável resposta no ajuste do modelo logístico usual.

Suissa (1991) apresentou um método paramétrico baseado na distribuição normal para a variável resposta contínua em que são aplicadas algumas técnicas para a estimação pontual e intervalar para o risco (ou prevalência) de uma determinada doença. Tal risco é uma medida muito importante usada em epidemiologia e é dada pela probabilidade da variável resposta ser maior do que um valor de corte  $C$ . Usando dados reais de epidemiologia, Suissa (1991) discutiu a eficiência do método em relação ao método usual para dados binários, apresentando estimativas para o risco e sua variância, no caso de uma amostra, e estimativas para a diferença de riscos e o risco relativo no caso de duas amostras. O autor destacou três vantagens: primeiro, o método de estimação do risco, segundo o modelo gaussiano, é mais eficiente do que o método baseado nos dados binários genuínos sob o modelo binomial. Para um mesmo tamanho de amostra, a variância do estimador da prevalência da doença segundo o modelo normal representa  $2/3$  da variância do modelo binomial quando o risco está entre 10% e 90%. Quando o risco assume um valor extremo (abaixo de 10% ou acima de 90%) tal eficiência é particularmente acentuada e o risco relativo decresce rapidamente a zero. Segundo, a estimação do risco e de sua variância, particularmente no caso de duas amostras, segundo o modelo contínuo, não está sujeita a problemas encontrados com o modelo binomial quando não há ocorrência do evento de interesse na amostra, pois, segundo Suissa (1991), o numerador no modelo contínuo nunca assume o valor zero. Terceiro, o método usual baseado nos dados binários é frequentemente subjetivo no sentido em que, muitas vezes, se discretiza medidas contínuas.

A desvantagem do método é que a má especificação da distribuição original da variável resposta compromete as estimativas obtidas, o que pode causar estimativas viesadas.

Suissa e Blais (1995) estenderam os resultados apresentados por Suissa (1991) em uma estrutura de modelos lineares generalizados com função de ligação composta para ajustar modelos de regressão logística baseados na variável resposta contínua original. Primeiramente os autores assumem uma distribuição normal para a resposta contínua e, em seguida, generalizam tal suposição usando a distribuição log-normal para dados reais de estudos clínicos e também para dados simulados. Para o caso de dados simulados, as estimativas de máxima verossimilhança dos parâmetros do modelo são de 25% a 85% mais eficientes do que os estimadores de máxima verossimilhança dos parâmetros do modelo logístico usual. A medida de eficiência usada por Suissa e Blais (1995) trata-se do quociente entre o erro quadrático médio das estimativas dos parâmetros do modelo com resposta de origem e o erro quadrático médio das estimativas dos parâmetros do modelo logístico usual.

Araújo (2002) reproduziu parte do estudo de simulação feito por Suissa e Blais (1995) e aplicou o modelo log-normal para dados reais de poluição considerando o risco de

concentração de  $NO_2$  na cidade de São Paulo. Araújo (2002) descreveu um procedimento de ajuste pelo método de máxima verossimilhança para os casos em que a variável resposta contínua tem distribuição normal e log-normal.

## 1.1 Objetivos do trabalho

Nesse trabalho apresentamos uma família de modelos de regressão com a informação da variável resposta original, cuja distribuição de probabilidades ou função densidade de probabilidade pertence à família exponencial. Incorporamos a informação da distribuição da variável resposta normal e exponencial no ajuste do modelo de regressão para diversas distribuições para a variável resposta de interesse. O modelo de regressão logística com resposta normal e log-normal desenvolvido por Suissa e Blais (1995) é apresentado como caso particular dos modelos de regressão com informação da distribuição da resposta de origem. Para a resposta de origem normal consideramos os modelos logístico, exponencial, geométrico, Poisson e log-normal. Para a resposta de origem exponencial consideramos os modelos logístico, normal, geométrico, Poisson e log-normal. Em contribuição ao trabalho de Suissa e Blais atribuímos duas respostas discretas ao modelo logístico, geométrico e de Poisson, e também consideramos uma resposta contínua normal com estrutura heteroscedástica. Adicionalmente, propomos também o modelo logístico com resposta pertencente à classe de distribuições séries de potências inflacionadas considerando o caso particular da resposta geométrica zero inflacionada.

Apresentamos a construção e o desenvolvimento dos modelos para cada distribuição, incorporando a informação sobre a distribuição original da variável resposta. Vários estudos com dados artificiais são utilizados para o caso particular do modelo de regressão logística com a informação da distribuição de origem e o modelo de regressão logística usual, bem como o modelo geométrico com resposta normal e exponencial. Assumindo uma distribuição corretamente especificada, a incorporação desta informação sobre a variável resposta no modelo produz estimativas de máxima verossimilhança mais eficientes e estimativas intervalares mais precisas.

As probabilidades de cobertura dos coeficientes de regressão são monitoradas e comparadas. Desenvolvemos a análise de resíduos para os dados simulados segundo o modelo logístico com respostas diversas e para os dois conjuntos de dados reais. Discutimos o ajuste dos modelos com e sem os pontos influentes para os dados artificiais. Como parte complementar, apresentamos o enfoque bayesiano do modelo de regressão logística com respostas diversas, discutindo os aspectos de convergência e autocorrelação das cadeias. Por fim, apresentamos as próximas fases do trabalho bem como as perspectivas e desdobramentos futuros. Abordaremos ao longo deste trabalho as vantagens dessa metodologia, ou seja, as melhorias obtidas ao se incorporar a informação da distribuição original dos dados no ajuste do modelo usual.

A importância dessa metodologia é especialmente verificada nas situações em que há excessos de zeros ou uns no vetor da variável resposta binária, isto é, nas situações extremas, em que há problemas no ajuste do modelo logístico usual. Não é raro encontrarmos conjuntos reais de dados, de diversas áreas do conhecimento, com o vetor da variável resposta binária composto por excessos de zeros ou excesso de uns. Por exemplo, a incidência da obesidade

mórbida na população ou ainda incidência de inadimplência sobre o pagamento mínimo da fatura do cartão de crédito, bem como fraudes, em geral, apresentam excessos de zeros. Nestes e em outros exemplos de eventos raros, a regressão logística, com o uso da informação da distribuição original da variável resposta, é ajustável, inclusive para os casos extremos, apresentando melhores estimativas dos coeficientes de regressão quando comparada com a regressão logística usual, desde que a distribuição de origem seja corretamente especificada.

## 1.2 Organização do Trabalho

No Capítulo 2 apresentamos uma família de modelos de regressão com resposta de origem usando diferentes distribuições pertencentes à família exponencial no contexto dos modelos lineares generalizados. Como caso particular, abordamos o modelo de regressão logística com resposta normal e log-normal desenvolvido por Suissa e Blais (1995). Estendemos a teoria para outros modelos de regressão.

No Capítulo 3 apresentamos um estudo de simulação para alguns modelos da classe de distribuições consideradas no Capítulo 2 a fim de comparar as estimativas obtidas segundo o método proposto e o modelo de regressão usual. Diversas métricas foram comparadas, tais como, o vício, erro padrão, erro quadrático médio e eficiência (quociente entre o erro quadrático médio das estimativas dos parâmetros do modelo com resposta de origem e o erro quadrático médio das estimativas dos parâmetros do modelo logístico usual). Além disso, a probabilidade de cobertura dos intervalos de confiança foi monitorada e discutida considerando os casos nominais 90%, 95% e 99%. Realizamos também uma simulação para o modelo geométrico com resposta de origem normal e exponencial. Análises de resíduos e diagnósticos são também apresentados, e discutimos o ajuste dos modelos com e sem os pontos influentes.

No Capítulo 4 apresentamos o modelo de regressão logística com resposta de origem usando distribuições pertencentes à classe de distribuições série de potências inflacionadas.

No Capítulo 5 abordamos o enfoque bayesiano dos modelos propostos e apresentamos os resultados considerando distribuições *a priori* vagas. Realizamos um estudo de simulação considerando diferentes modelos de regressão com diferentes variáveis respostas.

No Capítulo 6 apresentamos a aplicação da metodologia proposta em dois conjuntos de dados reais.

No Capítulo 7 apresentamos as discussões e considerações finais a respeito dos resultados obtidos para cada Capítulo desse trabalho, bem como suas perspectivas e seus possíveis desdobramentos.

Para a obtenção das estimativas dos parâmetros dos modelos expostos nesse trabalho, considerando o enfoque da inferência clássica, utilizamos o método da maximização direta por meio do algoritmo BFGS (Press, 1992).

Considerando o enfoque bayesiano utilizamos o algoritmo do amostrador de Gibbs com passos de Metropolis-Hastings. A convergência das cadeias geradas foram diagnosticadas pelo critério de diagnóstico de Gelman-Rubin (1992).

## 2 Uma família de modelos de regressão com a informação da distribuição original da variável resposta

### 2.1 Introdução

Nesse capítulo apresentamos uma família de modelos de regressão com a informação da variável resposta original, cuja distribuição de probabilidades ou função densidade de probabilidade pertence à família exponencial. Incorporamos a informação da distribuição da variável resposta original normal e exponencial ao ajuste de diferentes modelos de regressão. O modelo de regressão logística com respostas normal e log-normal, desenvolvido por Suissa e Blais (1995), é apresentado como caso particular dos modelos de regressão com resposta de origem. Para a resposta de origem normal consideramos os modelos logístico, exponencial, geométrico, Poisson e log-normal. Para a resposta de origem exponencial consideramos os modelos logístico, normal, geométrico, Poisson e log-normal.

### 2.2 Informação da distribuição da variável resposta

Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  com  $E(R_i) = \mu_{R_i}$ ,  $i = 1, 2, \dots, n$ , e um conjunto de  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  com  $E(Y_i) = \mu_{Y_i}$ ,  $i = 1, 2, \dots, n$ , com distribuições de probabilidades ou funções densidade de probabilidades pertencentes à família exponencial, ou seja,

$$f(r_i | \theta_{ri}, \phi_r) = \exp \left\{ \frac{r_i \theta_{ri} - b_r(\theta_{ri})}{a_r(\phi_r)} + c_r(r_i, \phi_r) \right\}$$

e

$$f(y_i | \theta_{yi}, \phi_y) = \exp \left\{ \frac{y_i \theta_{yi} - b_y(\theta_{yi})}{a_y(\phi_y)} + c_y(y_i, \phi_y) \right\},$$

em que  $a_r(\cdot)$ ,  $b_r(\cdot)$  e  $c_r(\cdot)$  são funções específicas da distribuição de probabilidades de  $R_i$  e  $a_y(\cdot)$ ,  $b_y(\cdot)$  e  $c_y(\cdot)$  são funções específicas da distribuição de probabilidades de  $Y_i$ ,  $\theta_{ri}$  e  $\theta_{yi}$  são os parâmetros canônicos das densidades de  $R_i$  e  $Y_i$ , respectivamente, na forma exponencial, e  $\phi_r$  e  $\phi_y$  são os parâmetros de dispersão de  $R_i$  e  $Y_i$ .

Se as variáveis  $R_i$  e  $Y_i$  são contínuas, então existem uma constante  $C_R$  no suporte da densidade de  $R_i$  e uma constante  $C_Y$  no suporte da densidade de  $Y_i$  tais que vale a seguinte relação:

$$P(R_i > C_R) = P(Y_i > C_Y), \quad i = 1, 2, \dots, n, \quad (2.2.1)$$

ou, equivalentemente,

$$F_{R_i}(C_R) = F_{Y_i}(C_Y), \quad i = 1, 2, \dots, n,$$

em que  $F_{R_i}(C_R)$  é a função distribuição da variável aleatória  $R_i$  no ponto  $C_R$  e  $F_{Y_i}(C_Y)$  é a função distribuição da variável aleatória  $Y_i$  no ponto  $C_Y$ .

Se as variáveis  $R_i$  são contínuas e  $Y_i$  são discretas, dado o seu ponto de corte  $C_Y$ , é possível determinar uma constante  $C_R$ , tal que a relação (2.2.1) seja verdadeira. Neste trabalho, a informação da distribuição da variável resposta original  $R_i$  é incorporada ao ajuste de um modelo para a variável  $Y_i$  por meio da relação (2.2.1).

## 2.3 Modelos da família exponencial com resposta de origem normal

Nessa seção apresentamos uma classe de modelos pertencentes à família exponencial com resposta de origem normal. Vamos supor que a variável resposta segue originalmente uma distribuição normal, isto é,  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição  $N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , isto é

$$\begin{aligned} f(r_i | \mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(r_i - \mu_i)^2\right\} \\ &= \exp\left\{\frac{1}{\sigma^2}\left(r_i\mu_i - \frac{\mu_i^2}{2}\right) - \frac{1}{2}\left[\frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]\right\}, \quad i = 1, 2, \dots, n, \end{aligned} \quad (2.3.1)$$

com  $\mu_i \in \mathbb{R}$ ,  $r_i \in \mathbb{R}$  e  $\sigma^2 > 0$ , fornecendo as seguintes funções específicas, com  $\theta_i = \mu_i$  e  $\phi = \sigma^2$ , da forma

$$b_r(\theta_i) = \theta_i^2/2, \quad a_r(\phi) = \sigma^2 \quad \text{e} \quad c_r(r_i, \phi) = -\frac{1}{2}\left[\frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\right].$$

**Proposição 1** *Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  tais que  $R_i \sim N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , e considere outro conjunto de  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  com distribuições de probabilidade ou funções densidade de probabilidade pertencentes à família exponencial. Incorporando a informação da variável resposta normal, segundo a relação dada em (2.2.1), no ajuste de  $Y_i$  temos que a média de  $R_i$  é  $\mu_i = \sigma\Phi^{-1}[1 - F_{Y_i}(C_Y)] + C_R$  e a distribuição de probabilidades de  $R_i$  é tal que*

$$R_i \sim N\{\sigma\Phi^{-1}[1 - F_{Y_i}(C_Y)] + C_R, \sigma^2\}, \quad i = 1, 2, \dots, n,$$

em que  $\Phi^{-1}$  é a inversa da função de distribuição da distribuição da normal padrão no ponto  $[1 - F_{Y_i}(C_Y)]$  e  $F_{Y_i}(C_Y)$  é a função de distribuição da variável aleatória  $Y_i$  no ponto  $C_Y$ .

**Prova.** Como  $R_i \sim N(\mu_i, \sigma^2)$ , segue que

$$P(R_i > C_R) = P\left[Z_i > \frac{C_R - \mu_i}{\sigma}\right] = P\left[Z_i < \frac{\mu_i - C_R}{\sigma}\right] = \Phi\left(\frac{\mu_i - C_R}{\sigma}\right),$$

em que  $Z_i$  é uma variável aleatória com distribuição normal padrão. Usando a relação dada em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \\ \text{e} \quad \Phi\left(\frac{\mu_i - C_R}{\sigma}\right) &= 1 - F_{Y_i}(C_Y), \end{aligned}$$

o que implica

$$\mu_i = E(R_i) = \sigma\Phi^{-1}[1 - F_{Y_i}(C_Y)] + C_R, \quad i = 1, 2, \dots, n. \quad \blacksquare$$

Assim, as funções específicas da família exponencial, para o caso em que a resposta de origem tem distribuição normal, ficam modificadas da seguinte forma:

$$b_r(\theta_i) = \frac{\{\sigma\Phi^{-1}[1 - F_{Y_i}(C_Y)] + C_R\}^2}{2}, \quad a_r(\phi) = \sigma^2 \quad \text{e} \quad c_r(r_i, \phi) = -\frac{1}{2}\left[\frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2)\right].$$



### 2.3.1 O caso particular de Suissa e Blais (1995): Modelo de regressão logística com resposta normal

A partir da motivação de estudos clínicos em medicina, Suissa e Blais (1995) consideraram  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  seguindo originalmente uma distribuição normal  $N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , e consideraram uma constante arbitrária  $C_R$  ( $C_R \in \mathbb{R}$ ). Adotaram também variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  binárias cujo valor da constante arbitrária adotado foi de  $C_Y = 0$ . Obedecendo a relação dada em (2.2.1), observamos que  $Y_i > 0$ , se  $R_i > C_R$ , e  $Y_i \leq 0$ , se  $R_i \leq C_R$ ,  $i = 1, 2, \dots, n$ .

Como as variáveis  $Y_i$  são dicotômicas verificamos que a probabilidade  $P(Y_i > 0)$  é igual à probabilidade de sucesso da variável  $Y_i$ , ou seja,  $P(Y_i = 1)$ , e a probabilidade  $P(Y_i \leq 0)$  é a probabilidade de fracasso da variável  $Y_i$ , ou seja  $P(Y_i = 0)$ . Dessa forma, temos que  $P(R_i > C_R) = P(Y_i = 1) = \pi_i$  e  $P(R_i \leq C_R) = P(Y_i = 0) = 1 - \pi_i$ ,  $i = 1, 2, \dots, n$  e segue que

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad i = 1, 2, \dots, n,$$

em que  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, 2, \dots, n$ , é a probabilidade de sucesso do modelo de regressão logística.

**Corolário 2.3.1** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição de Bernoulli com parâmetro  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, 2, \dots, n$ , então, pela proposição (1), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo logístico com resposta normal é tal que*

$$R_i \sim N\{\sigma\Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] + C_R, \sigma^2\}, \quad i = 1, 2, \dots, n.$$

**Prova.** Conforme visto anteriormente,

$$P(R_i > C_R) = \Phi\left(\frac{\mu_i - C_R}{\sigma}\right),$$

e, com relação à variável aleatória  $Y_i$  temos que

$$P(Y_i > C_Y) = P(Y_i > 0) = P(Y_i = 1) = \pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$$

Pela relação expressa em (2.2.1), segue que

$$\Phi\left(\frac{\mu_i - C_R}{\sigma}\right) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}),$$

o que implica

$$\mu_i = \sigma\Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] + C_R, \quad i = 1, 2, \dots, n.$$

■

Dessa forma temos

$$g(\pi_i) = g\left[F\left(\frac{\mu_i - C_R}{\sigma}\right)\right] = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, \quad i = 1, 2, \dots, n, \quad (2.3.2)$$

em que  $g[F(\cdot)]$  é a função de ligação composta que liga a média da variável resposta  $R_i$  ao preditor linear  $\mathbf{x}_i^T \boldsymbol{\beta}$ . Assim, as funções específicas da família exponencial ficam modificadas da seguinte forma:

$$b_r(\theta_i) = \frac{\{\sigma \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] + C_R\}^2}{2}; \quad a_r(\phi) = \sigma^2; \quad c_r(r_i, \phi) = -\frac{1}{2} \left[ \frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right].$$

Fazendo  $\gamma_i = (\mu_i - C_R)/\sigma$  e assumindo  $\sigma$  conhecido, este modelo pertence à classe dos modelos lineares generalizados com componente aleatório representado por um conjunto de variáveis aleatórias independentes com distribuição  $N(\gamma_i, 1)$ , função de ligação composta  $g[F(\cdot)]$  e o preditor linear  $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ ,  $i = 1, 2, \dots, n$ .

### 2.3.2 Modelo exponencial com resposta normal

Nessa seção construímos o modelo exponencial com resposta normal. Diversas situações práticas podem justificar essa abordagem. Para ilustrar, suponha que  $Y$  seja o tempo de recuperação cirúrgica e  $R$  é a idade do indivíduo. É possível assumir que a probabilidade do tempo de recuperação cirúrgica  $Y$  de um indivíduo ser maior do que cinco meses esteja relacionada com a probabilidade deste indivíduo ter uma idade  $R$  maior do que 60 anos. Dessa maneira a informação da variável resposta normal idade  $R$  pode ser incorporada no ajuste do tempo de recuperação  $Y$ .

Considere  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes com distribuição exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com função densidade dada por

$$\begin{aligned} f(y_i) &= \lambda_i \exp(-\lambda_i y_i) \\ &= \exp\{-\lambda_i y_i + \ln(\lambda_i)\}, \quad \lambda_i > 0, \quad y_i > 0 \text{ e } i = 1, 2, \dots, n. \end{aligned} \quad (2.3.3)$$

**Corolário 2.3.2** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição exponencial com parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com densidade dada por (2.3.3), então pela proposição (1), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo exponencial com resposta normal é tal que*

$$R_i \sim N\{\sigma \Phi^{-1}[\exp(-\lambda_i C_Y)] + C_R, \sigma^2\}, \quad i = 1, 2, \dots, n,$$

em que  $\Phi^{-1}$  é a função distribuição inversa da distribuição normal padrão avaliada no ponto  $\exp(-\lambda_i C_Y)$ .

**Prova.** A partir da relação dada em (2.2.1) temos que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \\ \text{e } \Phi\left(\frac{\mu_i - C_R}{\sigma}\right) &= \exp(-\lambda_i C_Y), \end{aligned}$$

o que implica

$$\mu_i = \sigma \Phi^{-1}[\exp(-\lambda_i C_Y)] + C_R, \quad i = 1, 2, \dots, n. \quad (2.3.4)$$

■

Logo, as funções específicas da família exponencial, para o caso em que temos um modelo exponencial com resposta normal, ficam modificadas da seguinte forma:

$$b_r(\theta_i) = \frac{\{\sigma\Phi^{-1}[\exp(-\lambda_i C_Y)] + C_R\}^2}{2}, \quad a_r(\phi) = \sigma^2 \quad \text{e} \quad c_r(r_i, \phi) = -\frac{1}{2} \left[ \frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right].$$

### 2.3.3 Modelo geométrico com resposta normal

Nessa seção construímos o modelo geométrico com resposta normal. Para ilustrar, suponha que a probabilidade de um motorista apresentar um número  $Y_i$  de sinistros de pequeno valor maior do que cinco até a ocorrência de um sinistro de alto valor, na seguradora de veículos, está relacionada com a probabilidade do indivíduo ter uma idade  $R_i$  acima de 70 anos. Considere  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes com distribuição geométrica de parâmetro  $p_i$ ,  $i = 1, 2, \dots, n$ , com a seguinte distribuição de probabilidades:

$$P(Y_i = y_i) = p_i(1 - p_i)^{y_i}, \quad i = 1, 2, \dots, n \quad \text{e} \quad y_i = 0, 1, 2, \dots \quad (2.3.5)$$

**Corolário 2.3.3** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição geométrica com parâmetro  $p_i$ ,  $i = 1, 2, \dots, n$ , então pela proposição (1), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo geométrico com resposta normal é tal que*

$$R_i \sim N \left\{ \sigma\Phi^{-1} \left[ (1 - p_i)^{C_Y+1} \right] + C_R, \sigma^2 \right\}, \quad i = 1, 2, \dots, n,$$

em que  $\Phi^{-1}$  é a inversa da função de distribuição de uma variável aleatória normal padrão.

**Prova.** Como  $Y_i \sim \text{Geométrica}(p_i)$ , com distribuição de probabilidades da forma expressa em (2.3.5) temos inicialmente que

$$\begin{aligned} P(Y_i > C_Y) &= 1 - P(Y_i \leq C_Y) \\ &= 1 - [P(Y_i = 0) + P(Y_i = 1) + \dots + P(Y_i = C_Y)] \\ &= 1 - \left[ p_i(1 - p_i)^0 + p_i(1 - p_i)^1 + p_i(1 - p_i)^2 + \dots + p_i(1 - p_i)^{C_Y} \right] \\ &= 1 - p_i \left[ \frac{(1 - p_i)^{C_Y+1} - 1}{(1 - p_i) - 1} \right], \end{aligned}$$

e, finalmente,

$$P(Y_i > C_Y) = (1 - p_i)^{C_Y+1}, \quad i = 1, 2, \dots, n.$$

Com a relação dada em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \quad \text{e} \\ \Phi \left( \frac{\mu_i - C_R}{\sigma} \right) &= (1 - p_i)^{C_Y+1}, \end{aligned}$$

o que implica

$$\mu_i = E(R_i) = \sigma\Phi^{-1} \left[ (1 - p_i)^{C_Y+1} \right] + C_R, \quad i = 1, 2, \dots, n. \quad (2.3.6)$$

■

Dessa forma, as funções específicas da família exponencial, para o caso em que temos o modelo geométrico com resposta normal, ficam modificadas da seguinte forma:

$$b_r(\theta_i) = \frac{\left\{ \sigma \Phi^{-1} \left[ (1 - p_i)^{C_Y + 1} \right] + C_R \right\}^2}{2}, \quad a_r(\phi) = \sigma^2 \quad \text{e} \quad c_r(r_i, \phi) = -\frac{1}{2} \left[ \frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right].$$

### 2.3.4 Modelo Poisson com resposta normal

Nessa seção construímos o modelo de Poisson com resposta normal. Ilustrativamente, considere agora que  $Y_i$  é o número de sinistros de pequeno valor em um determinado período e  $R_i$  é a idade do indivíduo. É razoável supor que a probabilidade de um motorista apresentar um número  $Y_i$  de sinistros de pequeno valor maior do que cinco em um determinado período está relacionada com a probabilidade do indivíduo ter uma idade  $R_i$  acima de 70 anos. Considere  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes com distribuição de Poisson com parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , isto é,

$$P(Y_i = y_i) = \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}, \quad i = 1, 2, \dots, n \quad \text{e} \quad y_i = 0, 1, 2, \dots \quad (2.3.7)$$

**Corolário 2.3.4** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_i, \sigma^2)$ ,  $i = 1, 2, \dots, n$ , e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição de Poisson com parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , então, pela proposição (1), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo de Poisson com resposta normal é tal que*

$$R_i \sim N \left\{ \sigma \Phi^{-1} \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \right] + C_R \right\}.$$

**Prova.** Pela distribuição de probabilidades dada em (2.3.7) temos que

$$P(Y_i > C_Y) = 1 - P(Y_i \leq C_Y) = 1 - \sum_{y_i=0}^{C_Y} P(Y_i = y_i) = 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!},$$

e seguem as relações

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \quad \text{e} \\ \Phi \left( \frac{\mu_i - C_R}{\sigma} \right) &= 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}, \end{aligned}$$

o que implica

$$\mu_i = E(R_i) = \sigma \Phi^{-1} \left\{ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \right\} + C_R, \quad i = 1, 2, \dots, n. \quad (2.3.8)$$

■

Assim, as funções específicas da família exponencial, para o caso em que temos o modelo de Poisson com resposta normal, ficam modificadas da seguinte forma:

$$b_r(\theta_i) = \frac{1}{2} \left\{ \sigma \Phi^{-1} \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \right] + C_R \right\}^2, \quad a_r(\phi) = \sigma^2 \quad \text{e} \quad c_r(r_i, \phi) = -\frac{1}{2} \left[ \frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right].$$

### 2.3.5 Modelo log-normal com resposta normal

Nessa seção construímos o modelo log-normal com resposta normal. Consideremos, inicialmente, o seguinte resultado. Se  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes positivas contínuas com distribuição log-normal com parâmetros  $\mu_{Y_i}$  e  $\sigma_Y^2$ ,  $i = 1, 2, \dots, n$ , então  $\log(Y_1), \log(Y_2), \dots, \log(Y_n)$  são  $n$  variáveis aleatórias independentes com distribuição normal com parâmetros  $\mu_{Y_i}$ ,  $i = 1, 2, \dots, n$ , e variância  $\sigma_Y^2$ . Portanto, os métodos de estimação para o modelo log-normal são similares aos do modelo normal, bastando substituir  $Y_i$  por  $\log(Y_i)$  e a constante  $C_Y$  por  $\log(C_Y)$ .

**Corolário 2.3.5** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_{R_i}, \sigma_R^2)$ ,  $i = 1, 2, \dots, n$ , e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição  $LN(\mu_{Y_i}, \sigma_Y^2)$ ,  $i = 1, 2, \dots, n$ , então pela proposição (1), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo log-normal com resposta normal é tal que*

$$R_i \sim N \left\{ \sigma_R \Phi_R^{-1} \left[ \Phi_Y \left( \frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y} \right) \right] + C_R, \sigma_R^2 \right\}, \quad i = 1, 2, \dots, n,$$

em que  $\Phi_R^{-1}[\cdot]$  é a inversa da função distribuição da variável aleatória normal padrão referente a variável  $R_i$  e  $\Phi_Y$  é a função distribuição da variável aleatória normal padrão referente à variável  $Y_i$ .

**Prova.** Pela relação dada em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P[\log(Y_i) > \log(C_Y)] \quad \text{e} \\ \Phi_R \left( \frac{\mu_{R_i} - C_R}{\sigma_R} \right) &= \Phi_Y \left( \frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y} \right), \end{aligned}$$

o que implica

$$\mu_{R_i} = \sigma_R \Phi_R^{-1} \left[ \Phi_Y \left( \frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y} \right) \right] + C_R, \quad i = 1, 2, \dots, n. \quad (2.3.9)$$

■

As funções específicas da família exponencial, para o caso em que temos o modelo log-normal com resposta normal, ficam modificadas da seguinte forma

$$\begin{aligned} b_r(\theta_i) &= \frac{1}{2} \left\{ \sigma_R \Phi_R^{-1} \left[ \Phi_Y \left( \frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y} \right) \right] + C_R \right\}^2; \\ a_r(\phi) &= \sigma^2; \\ c_r(r_i, \phi) &= -\frac{1}{2} \left[ \frac{r_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right]. \end{aligned}$$

### 2.3.6 Função de verossimilhança

Considere  $r_1, r_2, \dots, r_n$  realizações das variáveis aleatórias  $R_1, R_2, \dots, R_n$  e  $y_1, y_2, \dots, y_n$  realizações das variáveis  $Y_1, Y_2, \dots, Y_n$ . A função de verossimilhança para os modelos desenvolvidos com resposta normal, segundo a proposição (1), em que  $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$ , é da

seguinte forma:

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{r}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \quad (2.3.10)$$

$$\times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [r_i - \sigma \Phi^{-1}(1 - F_{Y_i}(C_Y)) - C_R]^2 \right\},$$

que, para facilitar a obtenção das funções derivadas de primeira e segunda ordem, necessárias para a obtenção dos escores, (2.3.10) pode ser reescrita como

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{r}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \quad (2.3.11)$$

$$\times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (r_i - C_R)^2 + \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) - \frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\}.$$

Cada um dos possíveis modelos de regressão assume uma expressão para o termo  $\psi_i$  em (2.3.11), conforme a Tabela 2.3.6.

Tabela 2.3.6. Expressões do termo  $\psi_i$  para os modelos de regressão.

Logístico	$\psi_i = \Phi^{-1}(\pi_i) = \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ , $i = 1, 2, \dots, n$ .
Exponencial	$\psi_i = \Phi^{-1}[\exp(-\lambda_i C_Y)]$ , $i = 1, 2, \dots, n$ , em que $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Geométrico	$\psi_i = \Phi^{-1}[(1 - p_i)^{C_Y + 1}]$ , $i = 1, 2, \dots, n$ , em que $p_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Poisson	$\psi_i = \Phi^{-1} \left\{ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \right\}$ , $i = 1, 2, \dots, n$ , em que $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Log-normal	$\psi_i = \Phi_R^{-1} \left[ \Phi_Y \left( \frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y} \right) \right]$ , $i = 1, 2, \dots, n$ , em que $\mu_{Y_i} = \mathbf{x}_i^T \boldsymbol{\beta}$ .

## 2.4 Modelos da família exponencial com resposta exponencial

Nessa seção apresentamos os modelos da família exponencial com resposta exponencial. Sejam  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  tal que  $R_i \sim \text{Exponencial}(\lambda_i)$ ,  $i = 1, 2, \dots, n$ , com função densidade dada por

$$f(r_i) = \lambda_i \exp(-\lambda_i r_i) \quad (2.4.1)$$

$$= \exp\{-\lambda_i r_i + \ln(\lambda_i)\}, \lambda_i > 0 \text{ e } i = 1, 2, \dots, n,$$

que, fazendo  $\theta_i = -\lambda_i$ , fornece as seguintes funções específicas da família exponencial:

$$b_r(\theta_i) = \ln(-\theta_i); \quad a_r(\phi_i) = 1; \quad c_r(r_i, \phi) = 0.$$

**Proposição 2** Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  tal que  $R_i \sim \text{Exponencial}(\lambda_i)$ ,  $i = 1, 2, \dots, n$ , com densidade dada por (2.4.1), e considere outro conjunto de  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  com distribuições de probabilidade ou funções densidade de probabilidade pertencentes à família exponencial. Incorporando a informação da variável resposta exponencial, segundo a relação dada em (2.2.1), no ajuste de  $Y_i$  temos que a distribuição de probabilidades de  $R_i$  é tal que

$$R_i \sim \text{Exponencial} \left\{ -\frac{\ln[1 - F_{Y_i}(C_Y)]}{C_R} \right\}, \quad i = 1, 2, \dots, n, \quad (2.4.2)$$

em que  $F_{Y_i}(C_Y)$  é a função distribuição de  $Y_i$  acumulada no ponto  $C_Y$ .

Dessa maneira, segue que a esperança e a variância são tais que

$$E(R_i) = -\frac{C_R}{\ln[1 - F_{Y_i}(C_Y)]} \quad (2.4.3)$$

$$\text{e } Var(R_i) = \left\{ -\frac{C_R}{\ln[1 - F_{Y_i}(C_Y)]} \right\}^2, \quad i = 1, 2, \dots, n. \quad (2.4.4)$$

**Prova.** Dada a relação (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \\ \text{e } \exp(-\lambda_i C_R) &= 1 - F_{Y_i}(C_Y), \end{aligned}$$

e, portanto, segue que o parâmetro  $\lambda_i$  no contexto da resposta exponencial, é tal que

$$\lambda_i = -\frac{\ln[1 - F_{Y_i}(C_Y)]}{C_R}. \quad (2.4.5)$$

Pela densidade expressa em (2.4.1) temos que  $E(R_i) = 1/\lambda_i$  e  $Var(R_i) = 1/\lambda_i^2$ . Considerando (2.4.5) segue imediatamente os resultados em (2.4.3) e (2.4.4). ■

As funções específicas modificadas da família exponencial dos modelos com resposta exponencial são

$$b_r(\theta_i) = \ln \left\{ -\frac{\ln[1 - F_{Y_i}(C_Y)]}{C_R} \right\}, \quad a_r(\phi_i) = 1 \quad \text{e} \quad c_r(r_i, \phi) = 0.$$

### 2.4.1 Modelo logístico com resposta exponencial

Nessa seção construímos o modelo logístico com resposta normal. Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  seguindo originalmente uma distribuição normal com distribuição exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com densidade expressa em (2.4.1) e considere  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  binárias cujo valor da constante arbitrária é de  $C_Y = 0$ . Obedecendo a relação dada em (2.2.1), observamos que  $Y_i > 0$ , se  $R_i > C_R$ , e  $Y_i \leq 0$ , se  $R_i \leq C_R$ ,  $i = 1, 2, \dots, n$ . Como as variáveis  $Y_i$  são dicotômicas verificamos que a probabilidade  $P(Y_i > 0) = P(Y_i = 1)$ , ou seja, é a probabilidade de sucesso da variável  $Y_i$ , e a probabilidade  $P(Y_i \leq 0) = P(Y_i = 0)$ , que é a probabilidade de fracasso da variável  $Y_i$ . Dessa forma temos que  $P(Y_i = 1) = P(R_i > C_R) = \pi_i$  e  $P(Y_i = 0) = P(R_i \leq C_R) = 1 - \pi_i$ ,  $i = 1, 2, \dots, n$  e segue que

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad i = 1, 2, \dots, n,$$

em que  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, 2, \dots, n$ .

**Corolário 2.4.1** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com densidade expressa em (2.4.1) e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição de Bernoulli com parâmetro  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, 2, \dots, n$ , então, pela proposição (2), a distribuição de*

probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo logístico com resposta exponencial é tal que

$$R_i \sim \text{Exponencial} \left\{ -\frac{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{C_R} \right\}, \quad i = 1, 2, \dots, n.$$

Logo, segue que a esperança e a variância são tais que

$$E(R_i) = -\frac{C_R}{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]} \quad \text{e} \quad (2.4.6)$$

$$\text{Var}(R_i) = \left\{ -\frac{C_R}{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]} \right\}^2, \quad i = 1, 2, \dots, n. \quad (2.4.7)$$

**Prova.** Pela relação vista anteriormente em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > 0) \quad \text{e} \\ \exp(-\lambda_i C_R) &= P(Y_i = 1) = \pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), \end{aligned}$$

o que implica

$$\lambda_i = -\frac{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{C_R}, \quad i = 1, 2, \dots, n,$$

e seguem imediatamente os resultados em (2.4.6) e (2.4.7). ■

As funções específicas modificadas da família exponencial, para o caso em que temos o modelo de regressão logística com resposta exponencial, são

$$b_r(\theta_i) = \ln \left\{ -\frac{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{C_R} \right\}, \quad a_r(\phi_i) = 1 \quad \text{e} \quad c_r(r_i, \phi) = 0.$$

## 2.4.2 Modelo normal com resposta exponencial

Nessa seção construímos o modelo normal com resposta exponencial. Diversas situações práticas podem justificar essa abordagem. Para ilustrar, suponha que  $Y$  seja a pressão arterial sistólica do indivíduo e que  $R$  seja o tempo diário que o indivíduo permanece em uma única posição no trabalho. É perfeitamente plausível que a probabilidade do indivíduo apresentar uma pressão arterial sistólica maior do que  $120 \text{ mmHg}$  esteja relacionada com a probabilidade do indivíduo permanecer mais do que cinco horas em uma única posição no trabalho.

Considere  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes com distribuição Normal de parâmetros  $\mu_i, i = 1, 2, \dots, n$ , e  $\sigma^2$ , com função densidade dada por

$$\begin{aligned} f(y_i | \mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \mu_i)^2 \right\} \\ &= \exp \left\{ \frac{1}{\sigma^2} \left( y_i \mu_i - \frac{\mu_i^2}{2} \right) - \frac{1}{2} \left[ \frac{y_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right] \right\}, \quad i = 1, 2, \dots, n, \end{aligned} \quad (2.4.8)$$

fornecendo as seguintes funções específicas, com  $\theta_i = \mu_i$  e  $\phi = \sigma$ , da forma

$$b_y(\theta_i) = \theta_i^2/2, \quad a_y(\phi) = \sigma^2 \quad \text{e} \quad c_y(y_i, \phi) = -\frac{1}{2} \left[ \frac{y_i^2}{\sigma^2} + \ln(2\pi\sigma^2) \right].$$



**Corolário 2.4.2** Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com densidade expressa em (2.4.1) e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_i, \sigma^2)$ , com densidade dada por (2.4.8), então pela proposição (2), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo normal com resposta exponencial é tal que

$$R_i \sim \text{Exponencial} \left\{ -\frac{1}{C_R} \ln \left[ \Phi_Y \left( \frac{\mu_i - C_Y}{\sigma} \right) \right] \right\}, \quad i = 1, 2, \dots, n.$$

A esperança e a variância são tais que

$$E(R_i) = -\frac{C_R}{\ln \left[ \Phi_Y \left( \frac{\mu_i - C_Y}{\sigma} \right) \right]} \quad \text{e} \quad (2.4.9)$$

$$\text{Var}(R_i) = \left\{ -\frac{C_R}{\ln \left[ \Phi_Y \left( \frac{\mu_i - C_Y}{\sigma} \right) \right]} \right\}^2 \cdot i = 1, 2, \dots, n. \quad (2.4.10)$$

**Prova.** Pela relação proposta em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \quad \text{e} \\ \exp(-\lambda_i C_R) &= \Phi_Y \left( \frac{\mu_i - C_Y}{\sigma} \right), \end{aligned}$$

o que implica

$$\lambda_i = -\frac{1}{C_R} \ln \left[ \Phi_Y \left( \frac{\mu_i - C_Y}{\sigma} \right) \right], \quad i = 1, 2, \dots, n,$$

e segue imediatamente os resultados em (2.4.9) e (2.4.10). ■

As funções específicas modificadas da família exponencial do modelo Normal com resposta exponencial são

$$b_r(\theta_i) = \ln \left\{ -\frac{\ln \left[ \Phi_Y \left( \frac{\mu_i - C_Y}{\sigma} \right) \right]}{C_R} \right\}, \quad a_r(\phi_i) = 1 \quad \text{e} \quad c_r(r_i, \phi) = 0.$$

### 2.4.3 Modelo geométrico com resposta exponencial

Nessa seção construímos o modelo geométrico com resposta exponencial. Considere  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes com distribuição geométrica de parâmetro  $p_i$ ,  $i = 1, 2, \dots, n$ , com a seguinte distribuição de probabilidades

$$P(Y_i = y_i) = p_i (1 - p_i)^{y_i}, \quad i = 1, 2, \dots, n \quad \text{e} \quad y_i = 0, 1, 2, \dots \quad (2.4.11)$$

**Corolário 2.4.3** Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com função densidade na forma (2.4.1) e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição geométrica com parâmetro  $p_i$ ,  $i = 1, 2, \dots, n$ , na forma (2.4.11), então pela proposição (2), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo geométrico com resposta exponencial é tal que

$$R_i \sim \text{Exponencial} \left\{ -\frac{1}{C_R} \ln \left[ (1 - p_i)^{C_Y + 1} \right] \right\}, \quad i = 1, 2, \dots, n.$$

A esperança e a variância são tais que

$$E(R_i) = -\frac{C_R}{\ln[(1-p_i)^{C_Y+1}]} \quad \text{e} \quad (2.4.12)$$

$$\text{Var}(R_i) = \left\{ -\frac{C_R}{\ln[(1-p_i)^{C_Y+1}]} \right\}^2, \quad i = 1, 2, \dots, n. \quad (2.4.13)$$

**Prova.** Como  $Y_i \sim \text{Geométrica}(p_i)$ , com distribuição de probabilidades da forma expressa em (2.4.11) temos primeiramente temos que

$$\begin{aligned} P(Y_i > C_Y) &= 1 - P(Y_i \leq C_Y) \\ &= 1 - [P(Y_i = 0) + P(Y_i = 1) + \dots + P(Y_i = C_Y)] \\ &= 1 - [p_i(1-p_i)^0 + p_i(1-p_i)^1 + p_i(1-p_i)^2 + \dots + p_i(1-p_i)^{C_Y}] \\ &= 1 - p_i \left[ \frac{(1-p_i)^{C_Y+1} - 1}{(1-p_i) - 1} \right], \end{aligned}$$

e finalmente,

$$P(Y_i > C_Y) = (1-p_i)^{C_Y+1}, \quad i = 1, 2, \dots, n.$$

Pela relação proposta em (2.2.1) temos que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \quad \text{e} \\ \exp(-\lambda_i C_R) &= (1-p_i)^{C_Y+1}, \end{aligned}$$

o que implica

$$\lambda_i = -\frac{1}{C_R} \ln[(1-p_i)^{C_Y+1}], \quad i = 1, 2, \dots, n,$$

e seguem imediatamente os resultados em (2.4.12) e (2.4.13). ■

As funções específicas modificadas da família exponencial do modelo geométrico com resposta exponencial são

$$b_r(\theta_i) = \ln \left\{ -\frac{\ln[(1-p_i)^{C_Y+1}]}{C_R} \right\}, \quad a_r(\phi_i) = 1 \quad \text{e} \quad c_r(r_i, \phi) = 0.$$

#### 2.4.4 Modelo de Poisson com resposta exponencial

Nessa seção construímos o modelo de Poisson com resposta exponencial. Considere  $Y_1, Y_2, \dots, Y_n$  variáveis aleatórias independentes com distribuição de Poisson com parâmetro  $\alpha_i$ ,  $i = 1, 2, \dots, n$ , isto é,

$$P(Y_i = y_i) = \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!}, \quad i = 1, 2, \dots, n \quad \text{e} \quad y_i = 0, 1, 2, \dots \quad (2.4.14)$$

**Corolário 2.4.4** Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com função densidade na forma (2.4.1) e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição de Poisson com parâmetro  $\alpha_i$ ,  $i = 1, 2, \dots, n$ , na forma (2.4.14), então pela proposição (2), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo de Poisson com resposta exponencial é tal que

$$R_i \sim \text{Exponencial} \left\{ -\frac{1}{C_R} \ln \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!} \right] \right\}, \quad i = 1, 2, \dots, n.$$

A esperança e a variância são tais que

$$E(R_i) = \frac{C_R}{\ln \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!} \right]} \quad e \quad (2.4.15)$$

$$Var(R_i) = \left\{ \frac{C_R}{\ln \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!} \right]} \right\}^2, \quad i = 1, 2, \dots, n. \quad (2.4.16)$$

**Prova.** Pela distribuição de probabilidades dada em (2.4.14) temos que

$$P(Y_i > C_Y) = 1 - P(Y_i \leq C_Y) = 1 - \sum_{y_i=0}^{C_Y} P(Y_i = y_i) = 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!},$$

e segue a relação

$$P(R_i > C_R) = P(Y_i > C_Y)$$

e

$$\exp(-\lambda_i C_R) = 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!},$$

o que implica

$$\lambda_i = -\frac{1}{C_R} \ln \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!} \right], \quad i = 1, 2, \dots, n. \quad (2.4.17)$$

o que resulta imediatamente as expressões (2.4.15) e (2.4.16). ■

Dessa forma, as funções específicas modificadas da família exponencial do modelo de Poisson com resposta exponencial são

$$b_r(\theta_i) = \ln \left\{ -\frac{1}{C_R} \ln \left[ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i) \alpha_i^{y_i}}{y_i!} \right] \right\}, \quad a_r(\phi_i) = 1 \quad e \quad c_r(r_i, \phi) = 0.$$

### 2.4.5 Modelo log-normal com resposta exponencial

Nessa seção construímos o modelo log-normal com resposta exponencial. Para ilustrar, suponha que  $Y$  seja o tempo de recuperação de indivíduos acidentados, e que  $R$  seja o tempo diário de exercícios fisioterápicos que o indivíduo realiza. É admissível que a probabilidade de um indivíduo demorar mais de 6 meses para se recuperar esteja relacionada com a probabilidade do indivíduo realizar mais de 30 minutos de exercícios fisioterápicos diários.

Consideremos, primeiramente, o seguinte resultado. Se  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes positivas contínuas com distribuição log-normal com parâmetros  $\mu_{Y_i}$  e  $\sigma_Y^2$ ,  $i = 1, 2, \dots, n$ , então  $\log(Y_1), \log(Y_2), \dots, \log(Y_n)$  são  $n$  variáveis aleatórias independentes com distribuição normal com parâmetros  $\mu_{Y_i}$ ,  $i = 1, 2, \dots, n$ , e variância  $\sigma_Y^2$ .

**Corolário 2.4.5** *Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original exponencial de parâmetro  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , com densidade expressa em (2.4.1) e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição original log-normal com parâmetros  $\mu_{Y_i}$  e  $\sigma_Y^2$ ,  $i = 1, 2, \dots, n$ , então pela proposição (2), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo log-normal com resposta exponencial é tal que*

$$R_i \sim \text{Exponencial} \left\{ -\frac{1}{C_R} \ln \left[ \Phi_Y \left( \frac{\mu_i - \log(C_Y)}{\sigma} \right) \right] \right\}, \quad i = 1, 2, \dots, n.$$

A esperança e a variância são tais que

$$E(R_i) = \frac{C_R}{\ln \left[ \Phi_Y \left( \frac{\mu_i - \log(C_Y)}{\sigma} \right) \right]} e \quad (2.4.18)$$

$$\text{Var}(R_i) = \left\{ \frac{C_R}{\ln \left[ \Phi_Y \left( \frac{\mu_i - \log(C_Y)}{\sigma} \right) \right]} \right\}^2, \quad i = 1, 2, \dots, n. \quad (2.4.19)$$

**Prova.** Pela relação proposta em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \\ \exp(-\lambda_i C_R) &= \Phi_Y \left( \frac{\mu_i - \log(C_Y)}{\sigma} \right), \end{aligned}$$

o que implica

$$\lambda_i = -\frac{1}{C_R} \ln \left[ \Phi_Y \left( \frac{\mu_i - \log(C_Y)}{\sigma} \right) \right], \quad i = 1, 2, \dots, n,$$

e segue imediatamente os resultados em (2.4.18) e (2.4.19). ■

Dessa maneira, as funções específicas modificadas da família exponencial do modelo log-normal com resposta exponencial são

$$b_r(\theta_i) = \ln \left\{ -\frac{1}{C_R} \ln \left[ \Phi_Y \left( \frac{\mu_i - \log(C_Y)}{\sigma} \right) \right] \right\}, \quad a_r(\phi_i) = 1 \quad \text{e} \quad c_r(r_i, \phi) = 0.$$

## 2.4.6 Função de verossimilhança para os modelos propostos com resposta exponencial

Considere  $r_1, r_2, \dots, r_n$  realizações das variáveis aleatórias  $R_1, R_2, \dots, R_n$  e  $y_1, y_2, \dots, y_n$  realizações das variáveis  $Y_1, Y_2, \dots, Y_n$ . A função de verossimilhança para os modelos desenvolvidos com resposta normal, segundo a proposição (2), em que  $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$ , é da seguinte forma

$$L(\boldsymbol{\beta} \mid \mathbf{r}) = \prod_{i=1}^n \left\{ -\frac{\ln[\psi_i] \times [\psi_i]^{r_i/C_R}}{C_R} \right\}. \quad (2.4.20)$$

Assim como nos modelos com resposta normal, cada um dos possíveis modelos de regressão com resposta exponencial assume uma expressão para o termo  $\psi_i$  em (2.4.20), conforme a Tabela 2.4.6.

Tabela 2.4.6. Expressões do termo  $\psi_i$  para os modelos de regressão.

Logístico	$\psi_i = \pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ , $i = 1, 2, \dots, n$ .
Normal	$\psi_i = F_Y\left(\frac{\mu_i - C_Y}{\sigma}\right)$ , $i = 1, 2, \dots, n$ , em que $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ .
Geométrico	$\psi_i = (1 - p_i)^{C_Y + 1}$ , $i = 1, 2, \dots, n$ , em que $p_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Poisson	$\psi_i = 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}$ , $i = 1, 2, \dots, n$ , em que $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Log-normal	$\psi_i = \Phi_R^{-1}\left[\Phi_Y\left(\frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y}\right)\right]$ , $i = 1, 2, \dots, n$ , em que $\mu_{Y_i} = \mathbf{x}_i^T \boldsymbol{\beta}$ .

## 2.5 O modelo de regressão logística com distribuições diversas para a variável resposta

Nessa seção apresentamos o modelo de regressão logística com resposta de origem usando duas distribuições discretas pertencentes à família exponencial, geométrica e de Poisson, no contexto dos modelos lineares generalizados e, também, o modelo normal considerando uma estrutura heteroscedástica multiplicativa.

### 2.5.1 Modelo de regressão logística com resposta geométrica

Consideremos  $n$  variáveis aleatórias  $R_1, R_2, \dots, R_n$  independentes seguindo uma distribuição geométrica com parâmetro  $p_i$ , isto é,  $R_i \sim \text{Geométrica}(p_i)$ ,  $i = 1, 2, \dots, n$ , tal que sua distribuição de probabilidades seja

$$\begin{aligned} P(R_i = r_i) &= p_i (1 - p_i)^{r_i} \\ &= \exp\{\ln(p_i) + r_i \ln(1 - p_i)\}, \quad i = 1, 2, \dots, n \text{ e } r_i = 0, 1, 2, \dots, \end{aligned} \quad (2.5.1)$$

sendo  $r_i = 0, 1, 2, \dots$  o número de fracassos até a ocorrência do primeiro sucesso em que

$$\begin{aligned} E(R_i) &= (1 - p_i) / p_i \\ \text{Var}(R_i) &= (1 - p_i) / p_i^2. \end{aligned}$$

Fazendo  $\theta_i = \ln(1 - p_i)$  e  $\phi_i = 1$ , temos que a expressão (2.5.1) fornece as seguintes funções específicas na forma da família exponencial

$$b_r(\theta_i) = -\ln[1 - \exp(\theta_i)]; \quad a_r(\phi_i) = 1; \quad c_r(r_i, \phi) = 0.$$

**Proposição 3** Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  tal que  $R_i \sim \text{Geométrica}(p_i)$ ,  $i = 1, 2, \dots, n$ , com distribuição de probabilidades na forma (2.5.1), e considere outro conjunto de  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  com distribuição de probabilidade de Bernoulli, isto é,  $Y_i \sim \text{Bernoulli}(\pi_i)$ ,  $i = 1, 2, \dots, n$ . Segundo a relação dada em (2.2.1), incorporando a informação da variável resposta geométrica no ajuste de  $Y_i$  temos que a distribuição de probabilidades de  $R_i$  é tal que

$$R_i \sim \text{Geométrica} \left\{ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{C_R+1} \right\}$$

**Prova.** Temos primeiramente que, com relação as variáveis aleatórias binárias  $Y_i$ ,

$$P(Y_i > C_Y) = P(Y_i > 0) = P(Y_i = 1) = \pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}),$$

e com relação as variáveis aleatórias  $R_i$  que

$$\begin{aligned} P(R_i > C_R) &= 1 - P(R_i \leq C_R) \\ &= 1 - [P(R_i = 0) + P(R_i = 1) + \dots + P(R_i = C_R)] \\ &= 1 - \left[ p_i(1-p_i)^0 + p_i(1-p_i)^1 + p_i(1-p_i)^2 + \dots + p_i(1-p_i)^{C_R} \right] \\ &= 1 - p_i \left[ \frac{(1-p_i)^{C_R+1} - 1}{(1-p_i) - 1} \right], \end{aligned}$$

e assim

$$P(R_i > C_R) = (1-p_i)^{C_R+1}, \quad i = 1, 2, \dots, n.$$

Dessa maneira, a partir da relação dada em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > 0) \\ (1-p_i)^{C_R+1} &= g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), \end{aligned}$$

e finalmente temos que

$$p_i = 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{C_R+1}, \quad i = 1, 2, \dots, n. \quad (2.5.2)$$

■

Dessa maneira as funções específicas da família exponencial, em que temos o modelo logístico com resposta geométrica, ficam modificadas da seguinte forma

$$b_r(\theta_i) = \ln \left[ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{C_R+1} \right], \quad a_r(\phi_i) = 1 \quad \text{e} \quad c_r(r_i, \phi) = 0,$$

e a densidade de  $R_i$  fica assim definida por meio das funções específicas modificadas com a incorporação da variável resposta da seguinte forma

$$f(r_i | \boldsymbol{\beta}) = \exp \left\{ r_i \frac{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{C_R + 1} + \ln \left[ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{C_R+1} \right] \right\}.$$

A esperança é dada por

$$E(R_i) = \mu_i = b'(\theta_i) = \frac{\exp(\theta_i)}{1 - \exp(\theta_i)}, \quad (2.5.3)$$

que, recordando que o parâmetro canônico é expresso por  $\theta_i = \ln(1 - p_i)$  e, a partir de (2.5.2) temos

$$E(R_i) = \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}, \quad i = 1, 2, \dots, n.$$

A variância, por sua vez, é dada por

$$Var(R_i) = V(\mu_i) a_i(\phi) = \frac{d\mu_i}{d\theta_i} a_i(\phi) = b''(\theta_i) a_i(\phi), \quad (2.5.4)$$

e como  $a_i(\phi) = 1$  segue que

$$Var(R_i) = V(\mu_i) = \frac{d\mu_i}{d\theta_i} = \frac{\exp(\theta_i)}{[1 - \exp(\theta_i)]^2},$$

e finalmente

$$Var(R_i) = \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{\left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\}^2}, \quad i = 1, 2, \dots, n.$$

Dessa forma, a partir de (2.5.2), a função de verossimilhança de  $R_1, R_2, \dots, R_n$  é expressa por

$$L(\boldsymbol{\beta} | \mathbf{r}) = \prod_{i=1}^n \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\} \left\{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{r_i}{C_R+1}}\right\}, \quad r_i = 0, 1, 2, \dots$$

e o logaritmo da função de verossimilhança é tal que

$$l(\boldsymbol{\beta} | \mathbf{r}) = \sum_{i=1}^n \ln \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\} + \sum_{i=1}^n \left(\frac{r_i}{C_R+1}\right) \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})], \quad r_i = 0, 1, 2, \dots \quad (2.5.5)$$

## 2.5.2 Escores para o modelo logístico com resposta geométrica

Nessa Seção desenvolvemos os escores para o modelo logístico com resposta geométrica. Considerando a função de log-verossimilhança dada em (2.5.5) e usando os resultados de funções derivadas expostas no Apêndice C, as equações escore são tais que

$$\frac{\partial l(\boldsymbol{\beta} | \mathbf{r})}{\partial \beta_j} = 0, \quad j = 0, 1, 2, \dots, p.$$

Na forma da família exponencial temos  $a_i(\phi) = 1$  e os termos

$$V(\mu_i) = \frac{d\mu_i}{d\theta_i} = Var(R_i) = \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{\left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\}^2}, \quad (2.5.6)$$

e

$$\begin{aligned} \frac{d\mu_i}{d\eta_i} &= \frac{d}{d(\mathbf{x}_i^T \boldsymbol{\beta})} \left\{ \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}} \right\} \\ &= \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{(C_R+1) [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\}^2}, \quad i = 1, 2, \dots, n. \end{aligned} \quad (2.5.7)$$

Portanto, o escore, a partir de (2.5.6) e (2.5.7) é dada por

$$U_j = \sum_{i=1}^n \left\{ r_i - \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}} \right\} \frac{x_{ij}}{(C_R + 1) [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} = 0, \quad j = 0, 1, 2, \dots, p.$$

Para o modelo logístico com resposta geométrica a matriz observada de informação de Fisher  $\mathbf{J}$  com dimensão  $(p + 1) \times (p + 1)$ , é dada por

$$\mathbf{J}_{\beta\beta} = \begin{pmatrix} J_{\beta_0\beta_0} & J_{\beta_0\beta_1} & \cdots & J_{\beta_0\beta_p} \\ J_{\beta_1\beta_0} & J_{\beta_1\beta_1} & \cdots & J_{\beta_1\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ J_{\beta_p\beta_0} & J_{\beta_p\beta_1} & \cdots & J_{\beta_p\beta_p} \end{pmatrix},$$

cujos elementos são dados por

$$\begin{aligned} J_{\beta_j\beta_j} &= \frac{\partial^2 l(\boldsymbol{\beta} | \mathbf{r})}{\partial \beta_j^2} \\ &= \sum_{i=1}^n \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \left\{ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \right\} (C_R + 1) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{(C_R + 1)^2 [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2 \left\{ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \right\}^2} \\ &\quad \times x_{ij}^2 \\ &\quad - \sum_{i=1}^n \left( \frac{r_i}{C_R + 1} \right) \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij}^2, \\ J_{\beta_j\beta_k} &= \frac{\partial^2 l(\boldsymbol{\beta} | \mathbf{r})}{\partial \beta_j \partial \beta_k} \\ &= \sum_{i=1}^n \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \left\{ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \right\} (C_R + 1) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{(C_R + 1)^2 [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2 \left\{ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \right\}^2} \\ &\quad \times x_{ij} x_{ik} \\ &\quad - \sum_{i=1}^n \left( \frac{r_i}{C_R + 1} \right) \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij} x_{ik}, \end{aligned}$$

para  $j = 0, 1, 2, \dots, p$  e  $k \neq j$ .

### 2.5.3 Modelo de regressão logística com resposta Poisson

Consideremos  $n$  variáveis aleatórias  $R_1, R_2, \dots, R_n$  independentes seguindo uma distribuição de Poisson com parâmetro  $\lambda_i$ , isto é,  $R_i \sim \text{Poisson}(\lambda_i)$ ,  $i = 1, 2, \dots, n$ , tal que sua distribuição de probabilidades seja

$$\begin{aligned} P(R_i = r_i) &= \frac{\exp(-\lambda_i) \lambda_i^{r_i}}{r_i!} \\ &= \exp\{r_i \ln(\lambda_i) - \lambda_i - \ln(r_i!)\}, \quad i = 1, 2, \dots, n \text{ e } r_i = 0, 1, 2, \dots, \end{aligned} \quad (2.5.8)$$

e

$$E(R_i) = \text{VAR}(R_i) = \lambda_i, \quad i = 1, 2, \dots, n.$$



Fazendo  $\theta_i = \ln(\lambda_i)$  e  $\phi = 1$  temos as seguintes funções específicas na forma da família exponencial

$$b_r(\theta_i) = \exp(\theta_i), \quad a_r(\phi) = 1 \quad \text{e} \quad c_r(r_i, \phi) = -\ln(r_i!).$$

**Proposição 4** *Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  tal que  $R_i \sim \text{Poisson}(\lambda_i)$ ,  $i = 1, 2, \dots, n$ , com distribuição de probabilidades na forma (2.5.8), e considere outro conjunto de  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  com distribuição de probabilidade de Bernoulli, isto é,  $Y_i \sim \text{Bernoulli}(\pi_i)$ ,  $i = 1, 2, \dots, n$ . Segundo a relação dada em (2.2.1), incorporando a informação da variável resposta Poisson no ajuste de  $Y_i$  temos  $R_i$  tem distribuição de Poisson cujo parâmetro  $\lambda_i$  é a solução da equação*

$$\sum_{r_i=0}^{C_R} \frac{\exp(-\lambda_i)\lambda_i^{r_i}}{r_i!} = 1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$$

Em particular, se  $C_R = 0$ , temos

$$R_i \sim \text{Poisson} \{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \}.$$

**Prova.** Sabemos que, com relação às variáveis binárias  $Y_i$  temos

$$P(Y_i > C_Y) = P(Y_i > 0) = P(Y_i = 1) = \pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}),$$

e com relação às variáveis aleatórias  $R_i$  que

$$P(R_i > C_R) = 1 - P(R_i \leq C_R) = 1 - \sum_{r_i=0}^{C_R} \frac{\exp(-\lambda_i)\lambda_i^{r_i}}{r_i!}.$$

A partir da relação (2.2.1) temos que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y) \\ 1 - \sum_{r_i=0}^{C_R} \frac{\exp(-\lambda_i)\lambda_i^{r_i}}{r_i!} &= g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), \end{aligned}$$

e assim temos a seguinte relação para o modelo de regressão logística com resposta Poisson

$$\sum_{r_i=0}^{C_R} \frac{\exp(-\lambda_i)\lambda_i^{r_i}}{r_i!} = 1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$$

Dessa maneira não há uma forma explícita para o parâmetro  $\lambda_i$ , a menos que  $C_R = 0$ , e segue que trata-se da seguinte solução

$$\exp(-\lambda_i) = 1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{1}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]}, \quad i = 1, 2, \dots, n.$$

Neste caso temos

$$\lambda_i = -\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] = \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})], \quad i = 1, 2, \dots, n, \quad (2.5.9)$$

e

$$R_i \sim \text{Poisson} \{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \}.$$

■

A as funções específicas da família exponencial, para o caso em que temos o modelo de regressão logística com resposta Poisson, ficam modificadas da seguinte forma

$$b_r(\theta_i) = -\ln \{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}; \quad a_r(\phi_i) = 1; \quad c_r(r_i, \phi) = -\ln(r_i!),$$

e a densidade de  $R_i$  fica assim definida por meio das funções específicas modificadas com a incorporação da variável resposta da seguinte forma

$$P(R_i = r_i) = \exp \{r_i \ln \{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \} - \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] - \ln(r_i!)\},$$

com

$$E(R_i) = Var(R_i) = \lambda_i = \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})], \quad i = 1, 2, \dots, n.$$

Dessa forma a função de verossimilhança de  $R_1, R_2, \dots, R_n$ , a partir da proposição (4) é

$$L(\boldsymbol{\beta} | \mathbf{r}) = \prod_{i=1}^n \frac{1}{r_i!} \{-\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}^{r_i} [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})], \quad r_i = 0, 1, 2, \dots$$

e o logaritmo da função de verossimilhança é tal que

$$l(\boldsymbol{\beta} | \mathbf{r}) = \sum_{i=1}^n \ln \left( \frac{1}{r_i!} \right) + \sum_{i=1}^n r_i \ln \{-\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\} + \sum_{i=1}^n \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]. \quad (2.5.10)$$

#### 2.5.4 Escores para o modelo de regressão logística com resposta Poisson

Nessa Seção desenvolvemos a inferência sobre o modelo logístico com resposta Poisson. Considerando a função de log-verossimilhança dada em (2.5.10) e usando os resultados de funções derivadas expostas no Apêndice C, os escores são tais que

$$\frac{\partial l(\boldsymbol{\beta} | \mathbf{r})}{\partial \beta_j} = 0, \quad j = 0, 1, 2, \dots, p. \quad (2.5.11)$$

Considerando as funções específicas da família exponencial e como  $a_r(\phi) = 1$  segue que

$$V(\mu_i) = \frac{d\mu_i}{d\theta_i} = Var(R_i) = \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})],$$

e

$$\frac{d\mu_i}{d\eta_i} = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \quad i = 1, 2, \dots, n,$$

temos que (2.5.11) é dado por

$$U_j = \sum_{i=1}^n \{r_i - \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]\} \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{\ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} x_{ij}, \quad j = 0, 1, 2, \dots, p.$$

Para o modelo logístico com resposta Poisson, considerando a constante  $C = 0$ , a matriz observada de informação de Fisher  $J$  com dimensão  $(p + 1) \times (p + 1)$ , é dada por

$$\mathbf{J}_{\beta\beta} = \begin{pmatrix} J_{\beta_0\beta_0} & J_{\beta_0\beta_1} & \cdots & J_{\beta_0\beta_p} \\ J_{\beta_1\beta_0} & J_{\beta_1\beta_1} & \cdots & J_{\beta_1\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ J_{\beta_p\beta_0} & J_{\beta_p\beta_1} & \cdots & J_{\beta_p\beta_p} \end{pmatrix},$$

cujos elementos são dados por

$$\begin{aligned} J_{\beta_j\beta_j} &= \frac{\partial^2 l(\boldsymbol{\beta} | \mathbf{r})}{\partial \beta_j^2} = \sum_{i=1}^n r_i \frac{\{1 + \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}}{\{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]\}^2} x_{ij}^2 - \sum_{i=1}^n \frac{r_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij}^2 \\ J_{\beta_j\beta_k} &= \frac{\partial^2 l(\boldsymbol{\beta} | \mathbf{r})}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n r_i \frac{\{1 + \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}}{\{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]\}^2} x_{ij} x_{ik} - \sum_{i=1}^n \frac{r_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} \\ &\quad \times x_{ij} x_{ik}, \end{aligned}$$

para  $j = 0, 1, 2, \dots, p$  e  $k \neq j$ .

### 2.5.5 Modelo de regressão logística com resposta normal em uma estrutura heteroscedástica

Para garantir a eficiência e a qualidade dos estimadores, no contexto da teoria clássica dos modelos de regressão, uma das suposições iniciais é que os erros seguem uma distribuição normal com variância constante. Quando isso não ocorre, os erros são heteroscedásticos. Em algumas situações é possível alcançar a hipótese de homoscedasticidade através de transformações na variável resposta (Box e Cox, 1964). Porém, como isso nem sempre é possível ou viável, torna-se preferível considerar uma análise com modelagem explícita da variância, que pode ser desenvolvida incluindo possíveis explicações da heterogeneidade da variância por meio de covariáveis.

Nessa seção supomos que a variável resposta segue originalmente uma distribuição normal, isto é,  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição  $N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$ , com função densidade dada por

$$\begin{aligned} f(r_i | \mu_i, \sigma_i^2) &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2} (r_i - \mu_i)^2\right\} \\ &= \exp\left\{\frac{1}{\sigma_i^2} \left(r_i \mu_i - \frac{\mu_i^2}{2}\right) - \frac{1}{2} \left[\frac{r_i^2}{\sigma_i^2} + \ln(2\pi\sigma_i^2)\right]\right\}, \quad i = 1, 2, \dots, n, \end{aligned} \quad (2.5.12)$$

fornecendo as seguintes funções específicas, com  $\theta_i = \mu_i$  e  $\phi_i = \phi/\omega_i = \sigma_i^2$ , da forma

$$b_r(\theta_i) = \mu_i^2/2; \quad a_r(\phi_i) = \phi/\omega_i = \sigma_i^2; \quad c_r(r_i, \phi) = -\frac{1}{2} \left[\frac{r_i^2}{\sigma_i^2} + \ln(2\pi\sigma_i^2)\right].$$

Quando temos uma situação de heterogeneidade é conveniente assumir a variância como uma função de uma ou mais variáveis regressoras, ou seja,

$$\sigma_i^2 = \phi_i = \phi/\omega_i = h(\mathbf{v}_i, \boldsymbol{\gamma}), \quad i = 1, 2, \dots, n,$$

em que  $h$  é uma função das covariáveis  $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{in})^T$  e do vetor de parâmetros de regressão  $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$  associados ao modelo, tal que  $h(\mathbf{v}_i, \boldsymbol{\gamma}) > 0$ .

**Proposição 5** Se  $R_1, R_2, \dots, R_n$  são  $n$  variáveis aleatórias independentes com distribuição original  $N(\mu_i, \sigma_i^2)$ ,  $i = 1, 2, \dots, n$ , e  $Y_1, Y_2, \dots, Y_n$  são  $n$  variáveis aleatórias independentes com distribuição de Bernoulli com parâmetro  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, 2, \dots, n$ , então, pela relação dada em (2.2.1), a distribuição de probabilidades de  $R_1, R_2, \dots, R_n$  no contexto do modelo logístico com resposta normal considerando uma estrutura heteroscedástica é tal que

$$R_i \sim N \left\{ \sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})} \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] + C_R; h(\mathbf{v}_i, \boldsymbol{\gamma}) \right\}, i = 1, 2, \dots, n.$$

**Prova.** Com relação as variáveis binárias  $Y_i$  sabemos que

$$P(Y_i > C_Y) = P(Y_i > 0) = P(Y_i = 1) = \pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}),$$

e com relação as variáveis aleatórias  $R_i$  que

$$P(R_i > C_R) = P \left[ Z_i > \frac{C_R - \mu_i}{\sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})}} \right] = P \left[ Z_i < \frac{\mu_i - C_R}{\sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})}} \right] = \Phi \left( \frac{\mu_i - C_R}{\sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})}} \right),$$

em que  $Z_i$  é uma variável aleatória com distribuição normal padrão. Usando a relação dada em (2.2.1) segue que

$$\begin{aligned} P(R_i > C_R) &= P(Y_i > C_Y), \\ \Phi \left( \frac{\mu_i - C_R}{\sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})}} \right) &= g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), \end{aligned}$$

o que implica em

$$\mu_i = E(R_i) = \sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})} \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] + C_R, i = 1, 2, \dots, n, \quad (2.5.13)$$

■

Temos, então, que  $\mu_i$  coincide com o parâmetro canônico  $\theta_i$  e dessa maneira, as funções específicas da família exponencial, para o caso em que temos o modelo de regressão logística com resposta normal considerando estrutura heteroscedástica, ficam modificadas da seguinte forma

$$\begin{aligned} b_r(\theta_i) &= \frac{\left\{ \sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})} \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] + C_R \right\}^2}{2} \\ a_r(\phi_i) &= \phi / \omega_i = h(\mathbf{v}_i, \boldsymbol{\gamma}) \\ c_r(r_i, \phi) &= -\frac{1}{2} \left[ \frac{r_i^2}{h(\mathbf{v}_i, \boldsymbol{\gamma})} + \ln(2\pi h(\mathbf{v}_i, \boldsymbol{\gamma})) \right]. \end{aligned}$$

A densidade de  $R_i$  fica, assim, definida por meio das funções específicas modificadas com a incorporação da variável resposta, na forma da família exponencial, da seguinte maneira

$$\begin{aligned} &f(r_i | \boldsymbol{\beta}, \sigma, \lambda) \\ &= \exp \left\{ \frac{1}{h(\mathbf{v}_i, \boldsymbol{\gamma})} \left[ r_i \left( \sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})} \psi_i + C_R \right) - \frac{\left( \sqrt{h(\mathbf{v}_i, \boldsymbol{\gamma})} \psi_i + C_R \right)^2}{2} \right] \right. \\ &\quad \left. - \frac{1}{2} \left[ \frac{r_i^2}{h(\mathbf{v}_i, \boldsymbol{\gamma})} + \ln(2\pi h(\mathbf{v}_i, \boldsymbol{\gamma})) \right] \right\}, \end{aligned}$$

em que  $\psi_i = \Phi^{-1} [g^{-1} (\mathbf{x}_i^T \boldsymbol{\beta})]$  para  $i = 1, 2, \dots, n$ .

Considerando  $r_1, r_2, \dots, r_n$  realizações das variáveis  $R_1, R_2, \dots, R_n$  independentes com distribuição dada pela proposição (5), a função de log-verossimilhança para  $\boldsymbol{\beta}$  e  $\sigma^2$ , com  $\mathbf{r} = (r_1, r_2, \dots, r_n)$ , pode ser escrita como

$$l(\boldsymbol{\beta}, \sigma^2 | \mathbf{r}) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n \ln[h(\mathbf{v}_i, \gamma)] \quad (2.5.14)$$

$$- \frac{1}{2} \sum_{i=1}^n \frac{\left[ r_i - \sqrt{h(\mathbf{v}_i, \gamma)} F^{-1} [g^{-1} (\mathbf{x}_i^T \boldsymbol{\beta})] - C \right]^2}{h(\mathbf{v}_i, \gamma)}.$$

Um caso particular, adotado nesse trabalho, para a estrutura heteroscedástica é

$$h(\mathbf{v}_i, \gamma) = \exp(\mathbf{v}_i^T \boldsymbol{\gamma}) = \sigma_i^2, \quad i = 1, 2, \dots, n, \quad (2.5.15)$$

denominado de estrutura heteroscedástica multiplicativa, em que a variância da variável resposta normal é proporcional a uma potência desconhecida de uma das variáveis explicativas. Se reescrevermos  $\boldsymbol{\gamma} = (\ln(\sigma^2), \lambda)^T$  e  $\mathbf{v}_i^T = (1, \ln(x_i))^T$ , obtemos

$$Var(R_i) = \sigma_i^2 = \sigma^2 x_i^\lambda, \quad i = 1, 2, \dots, n, \quad (2.5.16)$$

em que o parâmetro  $\lambda$  representa o grau ou o nível de heteroscedasticidade,  $x_i$  é uma das variáveis explicativas ou regressoras do modelo e  $\sigma^2$  é um parâmetro desconhecido e comum a todas as variâncias, também interpretada como a variância de  $R$  quando  $\lambda = 0$ . Neste caso a distribuição de  $R_i$  é dada por

$$R_i \sim N \left\{ \sigma x_i^{\lambda/2} \Phi^{-1} [g^{-1} (\mathbf{x}_i^T \boldsymbol{\beta})] + C_R; \sigma^2 x_i^\lambda \right\}, \quad i = 1, 2, \dots, n, \quad (2.5.17)$$

e (2.5.14) fica definida como

$$l(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\lambda}{2} \sum_{i=1}^n \ln(x_i) \quad (2.5.18)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{\left[ r_i - \sigma x_i^{\lambda/2} \Phi^{-1} [g^{-1} (\mathbf{x}_i^T \boldsymbol{\beta})] - C_R \right]^2}{x_i^\lambda}.$$

Para facilitar a obtenção das equações escore para  $\sigma$ ,  $\lambda$  e  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$  a expressão (2.5.18) pode ser reescrita como

$$l(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\lambda}{2} \sum_{i=1}^n \ln(x_i) \quad (2.5.19)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} - \frac{1}{2} \sum_{i=1}^n \psi_i^2,$$

em que

$$\psi_i = \Phi^{-1} [g^{-1} (\mathbf{x}_i^T \boldsymbol{\beta})] = \Phi^{-1} \left[ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right], \quad i = 1, 2, \dots, n. \quad (2.5.20)$$

### 2.5.6 Escores do modelo logístico com resposta normal considerando uma estrutura heteroscedástica

Nessa seção desenvolvemos a inferência sobre o modelo logístico com resposta normal em uma estrutura heteroscedástica. Apresentamos as equações escore para  $\sigma$ ,  $\lambda$  e  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$  obtidas de resultados de funções derivadas apresentadas no Apêndice C. Considere  $l(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r})$  a função de log-verossimilhança, dada em (2.5.19), os escores são tais que

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r})}{\partial \sigma} = 0; \quad \frac{\partial l(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r})}{\partial \lambda} = 0; \quad \frac{\partial l(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r})}{\partial \beta_j} = 0, \quad j = 0, 1, \dots, p.$$

Os estimadores de máxima verossimilhança podem ser obtidos pela maximização direta de (2.5.19) por meio do algoritmo BFGS. Verificadas certas condições de regularidade, a distribuição assintótica de  $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$  é uma distribuição normal multivariada  $N_{p+3}(\boldsymbol{\theta}, I(\boldsymbol{\theta})^{-1})$ , em que  $I(\boldsymbol{\theta})^{-1}$  é a matriz inversa de informação, que pode ser aproximada pela matriz de informação observada de Fisher,  $J(\boldsymbol{\theta})$ . Os elementos do vetor escore são tais que

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r})}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} - \frac{1}{\sigma^2} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}}, \quad (2.5.21)$$

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r})}{\partial \lambda} = -\frac{1}{2} \sum_{i=1}^n \ln(x_i) + \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2 \ln(x_i)}{x_i^\lambda} - \frac{1}{2\sigma} \sum_{i=1}^n \frac{(r_i - C_R) \psi_i \ln(x_i)}{x_i^{\lambda/2}}$$

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{r})}{\partial \beta_j} = \sum_{i=1}^n \left\{ \frac{1}{x_i^{\lambda/2}} \left( \frac{r_i - C_R}{\sigma} - \psi_i \right) \right\} \frac{1}{f\{\psi_i\} [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij},$$

em que  $f$  é a função densidade de probabilidade de uma variável aleatória normal padrão no ponto  $\psi_i = \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ ,  $i = 1, 2, \dots, n$  e  $j = 0, 1, \dots, p$ .

Com relação a equação escore dada em (2.5.21), o estimador de máxima verossimilhança  $\hat{\sigma}$ , para  $\sigma$ , dados  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$  e  $\lambda$ , tem forma analítica e é tal que

$$\hat{\sigma} = \frac{1}{2n} \left\{ -\sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} + \left[ \left\{ \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\}^2 + 4n \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \right]^{1/2} \right\} \quad (2.5.22)$$

em que  $\psi_i = \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ ,  $i = 1, 2, \dots, n$ .

Para o modelo logístico com resposta normal em uma estrutura heteroscedástica multiplicativa, a matriz de informação observada de Fisher,  $\mathbf{J}$ , com dimensão  $(p+3) \times (p+3)$ , dada por

$$\mathbf{J} = \begin{pmatrix} J_{\beta_0\beta_0} & J_{\beta_0\beta_1} & \dots & J_{\beta_0\beta_p} & J_{\beta_0\sigma} & J_{\beta_0\lambda} \\ J_{\beta_1\beta_0} & J_{\beta_1\beta_1} & \dots & J_{\beta_1\beta_p} & J_{\beta_1\sigma} & J_{\beta_1\lambda} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ J_{\beta_p\beta_0} & J_{\beta_p\beta_1} & \dots & J_{\beta_p\beta_p} & J_{\beta_p\sigma} & J_{\beta_p\lambda} \\ J_{\sigma\beta_0} & J_{\sigma\beta_1} & \dots & J_{\sigma\beta_p} & J_{\sigma\sigma} & J_{\sigma\lambda} \\ J_{\lambda\beta_0} & J_{\lambda\beta_1} & \dots & J_{\lambda\beta_p} & J_{\lambda\sigma} & J_{\lambda\lambda} \end{pmatrix},$$

cujos elementos são dados por

$$\begin{aligned}
J_{\beta_j \beta_j} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2 | \mathbf{r})}{\partial \beta_j^2} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2 f\{\psi_i\}} x_{ij}^2 \\
&\times \left\{ \frac{1}{\sigma x_i^{\lambda/2}} - \frac{(r_i - C_R) \frac{df\{\psi_i\}}{d\beta_j}}{\sigma x_i^{\lambda/2} f\{\psi_i\}} + \frac{(r_i - C_R)}{\sigma x_i^{\lambda/2}} \left[ \frac{1 - \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right] + \frac{2\psi_i}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} \right\}, \\
J_{\beta_j \beta_k} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2 | \mathbf{r})}{\partial \beta_j \partial \beta_k} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2 f\{\psi_i\}} x_{ij} x_{ik} \\
&\times \left\{ \frac{1}{\sigma x_i^{\lambda/2}} - \frac{(r_i - C_R) \frac{df\{\psi_i\}}{d\beta_j}}{\sigma x_i^{\lambda/2} f\{\psi_i\}} + \frac{(r_i - C_R)}{\sigma x_i^{\lambda/2}} \left[ \frac{1 - \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right] + \frac{2\psi_i}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} \right\}, \\
J_{\sigma\sigma} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r})}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + \frac{2}{\sigma^3} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}}, \\
J_{\lambda\lambda} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r})}{\partial \lambda^2} = -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2 [\ln(x_i)]^2}{x_i^\lambda} + \frac{1}{2\sigma} \sum_{i=1}^n \frac{(r_i - C_R) \psi_i [\ln(x_i)]^2}{2x_i^{\lambda/2}}, \\
J_{\sigma\lambda} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r})}{\partial \sigma \partial \lambda} = -\frac{1}{\sigma^3} \sum_{i=1}^n \frac{(r_i - C_R)^2 \ln(x_i)}{x_i^\lambda} + \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R) \psi_i \ln(x_i)}{x_i^{\lambda/2}}, \\
J_{\lambda\beta_j} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2 | \mathbf{r})}{\partial \lambda \partial \beta_j} = -\frac{1}{\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)}{f\{\psi_i\}} x_{ij} \quad \text{e} \\
J_{\sigma\beta_j} &= \frac{\partial^2 l(\boldsymbol{\beta}, \sigma^2 | \mathbf{r})}{\partial \lambda \partial \beta_j} = -\frac{1}{2\sigma} \sum_{i=1}^n \frac{(r_i - C_R) \ln(x_i)}{x_i^{\lambda/2} f\{\psi_i\}} x_{ij},
\end{aligned}$$

para  $j = 0, 1, \dots, p$  e  $j \neq k$ .

## 2.6 Diagnósticos para o modelo com resposta de origem

O conhecimento a respeito da relação entre as diversas variáveis envolvidas em um problema real, em muitas vezes, é limitado. Uma vez que os valores observados de tais variáveis podem e devem ser encarados como resultados de um experimento aleatório, então assumimos um modelo matemático por meio do qual as variáveis estejam relacionadas, como um processo que gerou os dados. Entretanto, uma fase necessária da análise de regressão é a validação do modelo, já que tal é uma aproximação da realidade. Dessa forma, a análise de diagnóstico avalia a qualidade dessa aproximação.

Nesse sentido, há o interesse em verificar e avaliar possíveis afastamentos das suposições assumidas para o modelo, como por exemplo, a distribuição de probabilidades dos dados observados. Sob outro aspecto também há o interesse em verificar a robustez do modelo sob determinadas perturbações nas formulações iniciais, com o objetivo de avaliar a estabilidade dos resultados inferenciais obtidos. Caso pequenas perturbações na construção original do modelo produzam resultados significativamente diferentes, então o modelo é considerado não robusto.

Neste contexto destacamos a distância de Cook (1977), as matrizes de alavancagem e as medidas de influência local. A expressão pontos de alavanca se da ao fato de tais pontos

exercerem uma influencia desproporcional ao próprio valor ajustado. Para o caso dos modelos lineares a medida de alavancagem está associada a matriz de projeção da solução de mínimos quadrados da regressão linear de  $Y$  contra  $X$ , dada por  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$  (Hoaglin e Welsh, 1978). De forma geral, como destacaram alguns autores, tais como Yoshizoe (1991), St. Laurent e Cook (1992), entre outros, uma medida de alavancagem deve refletir mais diretamente a influencia de  $Y_t$  ao próprio valor ajustado.

A distância de Cook (1977) tem como objetivo medir o impacto de uma observação particular nas estimativas dos coeficientes de regressão a partir da exclusão do conjunto de dados, verificando assim a interferência em resultados inferenciais.

### 2.6.1 Resíduos padronizados para a variável de interesse

Nessa Seção apresentamos os resíduos para resposta de interesse  $Y_i$ . Uma etapa inicial da análise de diagnóstico é a análise de resíduos a fim de detectar pontos mal ajustados ou aberrantes (*outliers*) e verificar indícios de afastamentos das suposições sobre o modelo em questão, como também verificar a adequação da distribuição de probabilidades proposta para a variável resposta. Tal etapa pode se basear nos resíduos ordinários e suas possíveis padronizações como também nos resíduos construídos a partir dos componentes da função desvio (McCullagh e Nelder, 1989) comumente utilizados em modelos lineares generalizados. Neste trabalho os preditos  $\hat{Y}_i$  são as esperanças estimadas  $\hat{E}(Y_i)$ , segundo os coeficientes de regressão estimados pelo modelo com resposta de origem. Dessa forma os resíduos padronizados são expressos por

$$rp_i = \frac{y_i - \hat{E}(Y_i)}{\sqrt{\widehat{Var}(Y_i)}}, i = 1, 2, \dots, n.$$

A tabela 2.6.1 apresenta um resumo dos dos modelos de regressão e seus preditos  $\hat{Y}_i = \hat{E}(Y_i)$ .



Tabela 2.6.1. Quadro resumo dos modelos de regressão e seus preditos.

Resposta de interesse $Y_i$	Resposta de origem $R_i$	Parâmetro em $R_i$	Parâmetro em $Y_i$	Predito $\hat{Y}_i = \hat{E}(Y_i)$
Logístico	Normal	$\mu_i = \sigma\Phi^{-1}[\pi_i] + C_R$	$\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Exponencial	Normal	$\mu_i = \sigma\Phi^{-1}[\exp(-\lambda_i C_Y)] + C_R$	$\lambda_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = [\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})]^{-1}$
Lognormal	Normal	$\mu_i = \sigma_R\Phi_R^{-1}\left[\Phi_Y\left(\frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y}\right)\right] + C_R$	$\mu_{Y_i} = \mathbf{x}_i^T\boldsymbol{\beta}$	$\hat{Y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$
Geométrico	Normal	$\mu_i = \sigma\Phi^{-1}\left[(1-p_i)^{C_Y+1}\right] + C_R$	$p_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = [\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})]^{-1}$
Poisson	Normal	$\mu_i = \sigma\Phi^{-1}\left[1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right] + C_R$	$\lambda_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Logístico	Exponencial	$\lambda_i = -\frac{\ln[\pi_i]}{C_R}$	$\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Normal	Exponencial	$\lambda_i = -\frac{1}{C_R} \ln\left[\Phi_Y\left(\frac{\mu_i - C_Y}{\sigma}\right)\right]$	$\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$	$\hat{Y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$
Lognormal	Exponencial	$\lambda_i = -\frac{1}{C_R} \ln\left[\Phi_Y\left(\frac{\mu_i - \log(C_Y)}{\sigma}\right)\right]$	$\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$	$\hat{Y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$
Geométrico	Exponencial	$\lambda_i = -\frac{1}{C_R} \ln\left[(1-p_i)^{C_Y+1}\right]$	$p_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = [\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})]^{-1}$
Poisson	Exponencial	$\lambda_i = -\frac{1}{C_R} \ln\left[1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\alpha_i)\alpha_i^{y_i}}{y_i!}\right]$	$\alpha_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Logístico	Geométrica	$p_i = 1 - [\pi_i]^{\frac{1}{C_R+1}}$	$\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Logístico	Poisson	$\lambda_i = -\ln[1 - \pi_i]$	$\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Logístico	Normal Hetero	$\mu_i = \sqrt{h(\mathbf{v}_i, \gamma)}\Phi^{-1}[\pi_i] + C_R$	$\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$
Logístico	GZI	$p_i = 1 - \left[\frac{\pi_i}{1-\omega}\right]^{\frac{1}{C_R+1}}$	$\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$	$\hat{Y}_i = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})$

## 2.6.2 Pontos influentes

Nesta seção abordamos a performance do ajuste dos modelos por meio da detecção e deleção de observações influentes. Considerando a matriz de projeção, ou também chamada de matriz “chapéu” dada por

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T,$$

temos que o  $i$ -ésimo elemento da diagonal de  $\mathbf{H}$  é dado por

$$0 \leq h_{ii} = \mathbf{X}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i^T \leq 1, \quad i = 1, 2, \dots, n,$$

e é usado para determinar se o  $i$ -ésimo ponto amostral é influente, com relação a  $\mathbf{X}$ , ou seja, é uma medida de distância entre o  $\mathbf{X}_i$  e o valor central de todos os outros casos. Valores de  $h_{ii}$  próximos de zero indicam que o ponto amostral não é outlier ou influente. Valores de  $h_{ii}$  próximos de um indicam que o ponto é influente.

Consideramos como pontos influentes todos os pontos tais que

$$h_{ii} > 2\bar{h} = \frac{2\sum h_{ii}}{n} = \frac{2p}{n},$$

em que  $p$  é o número de coeficientes de regressão no modelo considerado.

### 2.6.3 Teste de hipóteses para os coeficientes de regressão dos modelos com resposta de origem normal

Frequentemente há o interesse em testar se um modelo mais simples, isto é, com menos parâmetros, é adequado. Em outras palavras, há o interesse em verificar se algumas das covariáveis  $X_1, X_2, \dots, X_p$  podem ser retiradas do modelo. A formulação da hipótese pode ser expressa por

$$\begin{aligned} H_0 & : \beta_0 = \beta_1 = \dots = \beta_p = 0, \text{ contra} \\ H_1 & : \text{ pelo menos um dos parâmetros } \beta_0, \beta_1, \dots, \beta_p \text{ não é nulo.} \end{aligned}$$

A estatística da razão de verossimilhanças para testar  $H_0$  é

$$\xi_{RV} = 2 \left\{ \log L \left( \hat{\boldsymbol{\theta}}; \mathbf{r} \right) - \log L \left( \hat{\boldsymbol{\theta}}^0; \mathbf{r} \right) \right\},$$

em que  $\hat{\boldsymbol{\theta}}^0 = \left( \hat{\boldsymbol{\beta}}^0, \hat{\sigma}^0 \right)$  são, respectivamente, as estimativas de máxima verossimilhança de  $\boldsymbol{\beta}$  e  $\sigma$  sob a hipótese  $H_0$ , ou seja, as estimativas obtidas no ajuste do modelo reduzido, no qual não estão incluídas determinada(s) covariável(eis). Substituindo  $\hat{\boldsymbol{\beta}}^0$  e  $\hat{\sigma}^0$  em (2.3.11) obtemos  $\log L \left( \hat{\boldsymbol{\theta}}^0; \mathbf{r} \right)$ . O termo  $\log L \left( \hat{\boldsymbol{\theta}}; \mathbf{r} \right)$  é obtido substituindo em (2.3.11) as estimativas de máxima verossimilhança de  $\boldsymbol{\beta}$  e  $\sigma$ ,  $\hat{\boldsymbol{\beta}}$  e  $\hat{\sigma}$ , resultantes do ajuste do modelo completo, ou seja, com todas as covariáveis. Assintoticamente, sob a hipótese  $H_0$ , a estatística  $\xi_{RV}$  segue distribuição qui-quadrado com  $p-q$  graus de liberdade, em que  $p-q$  é o número de parâmetros considerados no modelo com resposta normal.

**Observação:** Para os modelos com resposta exponencial, o procedimento é análogo fazendo  $\hat{\boldsymbol{\theta}}^0 = \hat{\boldsymbol{\beta}}^0$ .

## 3 Ilustração com dados artificiais

### 3.1 Introdução

Para ilustrar a metodologia apresentamos neste capítulo um estudo de simulação para alguns modelos propostos. Comparamos as estimativas obtidas dos parâmetros de regressão com as estimativas do modelo de regressão usual considerado. Tal estudo de simulação envolve o modelo de regressão logística com resposta de origem seguindo distribuição normal, log-normal, exponencial, normal em um estrutura heteroscedástica multiplicativa, geométrica e de Poisson. Também apresentamos um estudo com o modelo de regressão geométrica com resposta normal e exponencial. Diversas métricas foram comparadas tais como o vício, erro padrão e erro quadrático médio simulados e a medida de eficiência adotada por Suissa e Blais (1995), que se trata do quociente entre o erro quadrático médio das estimativas dos parâmetros do modelo com resposta de origem e o erro quadrático médio das estimativas dos parâmetros do modelo logístico usual. O desempenho dos estimadores do modelo logístico com a informação de origem é invariavelmente melhor que o modelo usual no que se refere a tais métricas. Além disso, o estudo de simulação segundo os modelos propostos apresentam probabilidades de cobertura dos intervalos de confiança dos parâmetros mais próximas do real considerando os casos nominais 90%, 95% e 99%. Aspectos sobre diagnósticos de ajuste e análise de resíduos também são apresentados.

### 3.2 Uma revisão sobre o estudo de simulação de Suissa e Blais (1995)

Suissa e Blais (1995) realizaram um estudo de simulação, reproduzido posteriormente por Araújo (2002), para estudar a eficiência dos estimadores dos coeficientes de regressão  $\beta_0$  e  $\beta_1$  segundo o método proposto, isto é, incorporando a informação da distribuição original da variável resposta, considerando os modelos normal e log-normal. Tal estudo considerou apenas uma variável explicativa  $X$  que assumiu os valores 0, 1, 2, 3 e 4. Foram considerados tamanhos amostrais  $n = 20, 30, 50$  e  $100$  e geradas 100 amostras de tamanho  $n$ . A variável resposta  $R$  foi gerada da distribuição normal com média  $\mu = \sigma\Phi^{-1}[g^{-1}(\beta_0 + \beta_1 X)] + C_R$  e variância conhecida  $\sigma^2 = 1$ , em que diferentes valores para  $C_R$  foram fixados, de tal forma que, no ponto  $\bar{X} = 2$  ocorresse  $C_R = E(R)$ ,  $C_R = E(R) + 0,67$ ,  $C_R = E(R) + 1$  ou  $C_R = E(R) + 1,25$ . Para o caso de dados simulados as estimativas de máxima verossimilhança dos parâmetros do modelo foram de 25% a 85% mais eficientes que os estimadores de máxima verossimilhança dos parâmetros do modelo logístico usual.

### 3.3 Estudo de simulação para o modelo de regressão logística segundo a família de regressão proposta

Nessa Seção analisamos o desempenho das estimativas do modelo logístico com resposta de origem por meio de um estudo com dados artificiais e comparamos com o desempenho do modelo de regressão logística usual. Como fase preliminar realizamos este estudo para os modelos normal e log-normal apresentados por Suissa e Blais (1995) considerando

três variáveis explicativas. Consideramos os tamanhos de amostra  $n = 50$ ,  $n = 100$ ,  $n = 200$  e  $n = 500$  e geramos 10.000 amostras de tamanho  $n$  com o objetivo de monitorar a acurácia e a probabilidade de cobertura dos intervalos de confiança de 90%, 95% e 99% de cada um dos coeficientes de regressão. Tal procedimento foi realizado considerando as variáveis respostas normal, log-normal, exponencial, normal considerando uma estrutura de heteroscedasticidade multiplicativa, geométrico e Poisson. Particularmente para o modelo logístico com resposta discreta geométrica adotamos a constante  $C_R = 0$ . Em diversas situações práticas há o interesse na constante  $C_R = 0$  como, por exemplo:

- Número de defeitos que um determinado item apresenta numa linha de produção. Basta que o item apresente um único defeito e o mesmo é considerado impróprio para o consumo. Então  $Y_i = 1$  se  $R_i = 1, 2, 3, \dots$  e  $Y_i = 0$  se  $R_i = 0$ .
- Número de cáries dentárias que o indivíduo apresenta. Basta que tenha uma única cárie para que seja necessário o seu tratamento, isto é, para que se configure em “sucesso”,  $Y_i = 1$ .
- Número de dias entre o pagamento da fatura do cartão de crédito e seu vencimento. Em geral, se o cliente realizar o pagamento mínimo da fatura do cartão em no máximo 60 dias ele é considerado adimplente e, num período maior do que 60 dias ele é considerado inadimplente. Desta forma  $Y_i = 1$  se  $R_i > 60$  dias e  $Y_i = 0$  se  $R_i \leq 60$  dias.

Para cada um dos casos estudados a seguir, as Tabelas de resultados apresentam as seguintes métricas de avaliação:

**Estimativa:** denota a média obtida das estimativas dos coeficientes de regressão, segundo ambos os modelos, considerando todas as 10.000 amostras geradas.

**EP:** denota a média obtida do erro-padrão, segundo ambos os modelos, considerando todas as 10.000 amostras geradas.

**90%, 95% e 99%** denotam respectivamente a probabilidade de cobertura nominal para cada parâmetro de ambos os modelos.

**Eficiência:** é o quociente entre o erro-quadrático-médio das estimativas dos parâmetros do modelo com resposta de origem e o erro-quadrático-médio das estimativas dos parâmetros do modelo logístico usual. Quando a eficiência assume um valor menor que 1 então o primeiro é menor que o segundo.

O algoritmo usado foi o BFGS, cujos valores iniciais para os coeficientes de regressão são as estimativas obtidas segundo o modelo logístico usual.

### 3.4 Modelo de regressão logística com resposta normal

Nesta seção vamos analisar a performance do modelo logístico com resposta normal e, assim, estudar a eficiência das estimativas dos coeficientes de regressão  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo com dados artificiais. Para a geração dos dados supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(20; 16); R_i \sim N(\mu_i, \sigma^2),$$

com  $\sigma = 1.000$  e

$$\mu_i = \sigma \Phi^{-1} [g^{-1} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + C_R, i = 1, 2, \dots, n.$$

Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\mu_i, i = 1, 2, \dots, n$ , foram  $\beta_0 = -4,5, \beta_1 = 0,5, \beta_2 = 0,35$  e  $\beta_3 = 0,02$ .

Consideramos o ponto de corte  $C_R = 11.000$ , isto é, todos os valores de  $R_i$  acima de 11.000 foram considerados sucesso, ou seja, tiveram a resposta igual a 1 e todos os valores de  $R_i$  com valores abaixo de 11.000 foram considerados fracasso, ou seja, tiveram a resposta igual a 0.

A Tabela 3.4 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta normal, para cada tamanho amostral, bem como as métricas de avaliação.

Tabela 3.4. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros	Modelo com resposta de origem (n=50)					Modelo logístico usual (n=50)						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-4,5	-4,2688	2,4529	92,1	96,3	99,4	-5,5258	4,4295	87,8	93,8	98,4	0,2936
$\beta_1$	0,50	0,4765	0,3046	92,5	96,8	99,5	0,6296	0,5130	86,7	92,6	98,0	0,3332
$\beta_2$	0,35	0,3299	0,1446	92,2	96,5	99,3	0,4141	0,2555	87,6	93,2	98,3	0,3069
$\beta_3$	0,02	0,0195	0,0728	92,0	96,8	99,6	0,0244	0,1271	87,6	93,0	98,1	0,3272
$\sigma$	1000	1055,68	127,35	98,9	99,4	99,8	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem (n=100)					Modelo logístico usual (n=100)						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-4,5	-4,4676	1,4889	90,8	95,4	99,1	-4,7781	2,2681	89,0	94,3	98,8	0,4248
$\beta_1$	0,50	0,4939	0,1964	90,5	95,3	99,1	0,5356	0,3004	88,5	94,2	98,5	0,4219
$\beta_2$	0,35	0,3483	0,0970	90,8	95,7	99,4	0,3712	0,1502	88,9	94,1	98,7	0,4087
$\beta_3$	0,02	0,0203	0,0431	90,1	95,3	99,2	0,0205	0,0639	88,7	93,8	98,5	0,4634
$\sigma$	1000	1011,21	75,55	95,3	97,4	99,2	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem (n=200)					Modelo logístico usual (n=200)						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-4,5	-4,5561	1,1931	90,1	94,8	99,0	-4,6343	1,7827	89,9	94,9	98,8	0,4464
$\beta_1$	0,50	0,5071	0,1415	90,2	95,2	98,9	0,5183	0,2136	89,1	94,5	98,7	0,4370
$\beta_2$	0,35	0,3541	0,0680	89,5	94,8	99,0	0,3602	0,1044	89,1	94,4	98,9	0,4182
$\beta_3$	0,02	0,0200	0,0328	89,5	94,9	98,9	0,0200	0,0488	89,3	94,4	98,8	0,4583
$\sigma$	1000	995,60	50,25	88,0	95,1	98,6	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem (n=500)					Modelo logístico usual (n=500)						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-4,5	-4,5322	0,7477	90,2	94,9	99,0	-4,5502	1,1081	89,9	94,9	99,1	0,4552
$\beta_1$	0,50	0,5037	0,0867	89,7	95,1	98,9	0,5059	0,1331	90,3	94,8	98,9	0,4213
$\beta_2$	0,35	0,3523	0,0435	90,2	95,4	99,0	0,3544	0,0658	90,4	95,2	99,0	0,4419
$\beta_3$	0,02	0,0201	0,0202	89,7	94,7	98,9	0,0199	0,0309	90,1	95,0	99,1	0,4000
$\sigma$	1000	996,44	31,54	89,6	94,8	98,8	—	—	—	—	—	—

Observamos por meio da Tabela 3.4 que o modelo de regressão logística com a informação da resposta normal produziu estimativas pontuais dos coeficientes mais próximas

dos valores reais para todos os tamanhos amostrais. O erro padrão obtido foi menor para todos os parâmetros produzindo estimativas intervalares com menores amplitudes, ou seja, mais precisas. Verificamos também que, a medida que o tamanho da amostra aumenta as estimativas pontuais de ambos os modelos se aproximam e, no entanto, o erro padrão continua sendo menor para todos os tamanhos de amostra considerados.

Considerando as probabilidades de coberturas obtidas dos intervalos de confiança, verificamos que para os tamanhos de amostra  $n = 50$  e  $n = 100$  o modelo logístico usual subestima os três casos considerados de probabilidade de cobertura (90%, 95% e 99%). Por outro lado, o modelo logístico com resposta de origem apresentou uma cobertura dos intervalos mais próximos para todos os casos considerados. A medida que o tamanho da amostra aumenta ( $n = 200$  e  $n = 500$ ) a probabilidade de cobertura dos intervalos de confiança de ambos os modelos se aproximam cada vez mais.

Com relação à eficiência, o erro quadrático médio das estimativas segundo o modelo logístico com resposta de origem é menor que o erro quadrático médio das estimativas do modelo logístico usual. O erro quadrático médio das estimativas do modelo com resposta de origem é de 2,2 a 3,4 vezes menor que o erro quadrático médio das estimativas do modelo logístico usual.

### 3.5 Modelo de regressão logística com resposta log-normal

Nesta seção vamos analisar a performance do modelo logístico com resposta normal e, assim, estudar a eficiência das estimativas dos coeficientes de regressão  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo com dados artificiais. Para a geração dos dados supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(25; 25); R_i \sim LN(\mu_i, \sigma^2),$$

em que  $LN$  denota a distribuição log-normal, e

$$\mu_i = \sigma F^{-1} [\Phi^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + \log(C_R), \quad i = 1, 2, \dots, n,$$

com  $\sigma^2 = 0,50 \Rightarrow \sigma = 0,71$ . Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\mu_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -12,0$ ,  $\beta_1 = 1,0$ ,  $\beta_2 = 0,44$  e  $\beta_3 = 0,05$ . Consideramos o ponto de corte  $C_R = 10$ , isto é, todos os valores de  $R_i$  acima de 10 foram considerados sucessos, ou seja, tiveram a resposta igual a 1 e todos os valores de  $R_i$  com valores abaixo de 10 foram considerados fracassos, ou seja, tiveram a resposta igual a 0. A Tabela 3.5 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta log-normal, para cada tamanho amostral, bem como as métricas de avaliação.

Tabela 3.5. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros	Modelo com resposta de origem ( $n=50$ )					Modelo logístico usual ( $n=50$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-12,0	-13,0908	3,0009	89,0	94,3	99,0	-14,5979	7,0029	86,0	92,3	97,9	0,1827
$\beta_1$	1,0	1,0915	0,3547	89,0	94,4	99,1	1,1986	0,9171	87,1	92,8	98,1	0,1524
$\beta_2$	0,44	0,4785	0,1493	88,9	94,5	99,0	0,5414	0,2470	86,1	92,3	97,8	0,3338
$\beta_3$	0,05	0,0550	0,0581	88,4	94,5	98,9	0,0626	0,0915	87,6	93,3	98,0	0,4000
$\sigma$	0,71	0,6743	0,0673	82,2	87,8	94,5	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem ( $n=100$ )					Modelo logístico usual ( $n=100$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-12,0	-12,5194	1,9883	89,8	94,6	99,0	-13,1894	3,3153	88,3	93,9	98,5	0,3404
$\beta_1$	1,0	1,0399	0,2252	89,7	94,8	98,9	1,0922	0,3637	88,4	93,9	98,6	0,3714
$\beta_2$	0,44	0,4593	0,1036	89,6	94,9	99,2	0,4858	0,1656	88,6	94,2	98,7	0,3763
$\beta_3$	0,05	0,0528	0,0379	89,3	94,6	98,9	0,0562	0,0613	88,8	94,3	99,0	0,3684
$\sigma$	0,71	0,6911	0,0488	85,8	91,6	97,1	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem ( $n=200$ )					Modelo logístico usual ( $n=200$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-12,0	-12,2426	1,4962	89,5	94,8	99,0	-12,4977	2,3360	89,7	94,5	98,9	0,4027
$\beta_1$	1,0	1,0213	0,1573	90,1	94,8	99,0	1,0417	0,2393	89,7	94,8	98,9	0,4271
$\beta_2$	0,44	0,4489	0,0704	89,7	94,7	98,8	0,4586	0,1065	89,6	94,5	98,8	0,4274
$\beta_3$	0,05	0,0509	0,0276	89,2	94,4	98,8	0,0523	0,0424	89,8	94,7	98,9	0,4444
$\sigma$	0,71	0,6995	0,0349	88,2	93,4	98,2	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem ( $n=500$ )					Modelo logístico usual ( $n=500$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\beta_0$	-12,0	-12,0955	0,9249	89,9	94,9	99,0	-12,1930	1,4037	89,4	94,6	99,0	0,4307
$\beta_1$	1,0	1,0084	0,0938	90,2	95,1	99,0	1,0178	0,1390	89,6	94,8	99,1	0,4541
$\beta_2$	0,44	0,4433	0,0451	90,5	95,2	99,0	0,4459	0,0672	90,3	95,2	99,0	0,4348
$\beta_3$	0,05	0,0504	0,0159	89,8	94,7	98,9	0,0511	0,0231	89,6	94,6	98,9	0,6000
$\sigma$	0,71	0,7040	0,0222	90,2	95,0	98,8	—	—	—	—	—	—

Assim como no caso do modelo com resposta normal, observamos por meio da Tabela 3.5 que o modelo de regressão logística com a informação da resposta log-normal produziu estimativas pontuais dos coeficientes mais próximas dos valores reais para todos os tamanhos amostrais. O erro padrão obtido foi menor para todos os parâmetros produzindo estimativas intervalares com menores amplitudes, ou seja, mais precisas.

Considerando as probabilidades de coberturas obtidas dos intervalos de confiança, o modelo logístico com resposta de origem apresentou uma cobertura dos intervalos mais próximos para todos os casos considerados.

Quanto à eficiência, o erro quadrático médio das estimativas segundo o modelo logístico com resposta de origem é menor que o erro quadrático médio das estimativas do modelo logístico usual. O erro quadrático médio das estimativas do modelo com resposta de origem é de 1,7 a 6,6 vezes menor que o erro quadrático médio das estimativas do modelo logístico usual.

### 3.6 Modelo de regressão logística com resposta normal em uma estrutura de heteroscedasticidade

Nesta seção analisamos a performance do modelo de regressão logística com resposta normal em uma estrutura de heteroscedasticidade multiplicativa. Estudamos a eficiência das estimativas dos coeficientes de regressão  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo com dados artificiais. Para a geração dos dados supomos três variáveis explicativas da seguinte forma:

$$\begin{aligned} X_{i1} &\sim N(10; 4); X_{i2} \sim N(5; 1); X_{i3} \sim N(15; 4); R_i \sim N(\mu_i, \sigma_i^2), \\ \mu_i &= \sigma x_{1i}^{\lambda/2} \Phi^{-1} [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + C_R, i = 1, 2, \dots, n, \\ \sigma_i^2 &= x_{1i}^\lambda \sigma^2, i = 1, 2, \dots, n. \end{aligned}$$

Os valores atribuídos para o vetor  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\mu_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -32, 0$ ,  $\beta_1 = 0, 77$ ,  $\beta_2 = 2, 2$  e  $\beta_3 = 1, 0$ . Para os parâmetros perturbadores atribuímos  $\sigma^2 = 4000$  e  $\lambda = 3$ . Consideramos o ponto de corte  $C_R = 10.000$ , isto é, todos os valores de  $R_i$  acima de 10.000 foram considerados sucessos, ou seja, tiveram a resposta igual a 1 e todos os valores de  $R_i$  com valores abaixo de 10.000 foram considerados fracassos, ou seja, tiveram a resposta igual a 0. A Tabela 3.6 mostra as estimativas obtidas dos coeficientes de regressão considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta normal heteroscedástica, para cada tamanho amostral.



Tabela 3.6. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros		Modelo com resposta de origem ( $n=50$ )					Modelo logístico usual ( $n=50$ )					Eficiência
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	
$\beta_0$	-32,0	-35,4379	6,2886	87,6	94,0	99,0	-42,0181	14,5304	82,9	89,9	97,2	0,1649
$\beta_1$	0,77	0,8533	0,2167	87,7	93,8	98,8	0,9908	0,4253	84,1	91,0	97,8	0,2348
$\beta_2$	2,20	2,4315	0,4924	88,2	94,1	99,0	2,9052	1,1064	84,2	90,6	97,3	0,1720
$\beta_3$	1,00	1,1075	0,2314	88,1	94,0	98,9	1,3197	0,5118	84,2	90,5	97,5	0,1788
$\sigma$	63,25	62,27	63,92	87,1	88,6	90,9	—	—	—	—	—	—
$\lambda$	3	3,10	0,9325	93,5	96,1	98,6	—	—	—	—	—	—

Parâmetros		Modelo com resposta de origem ( $n=100$ )					Modelo logístico usual ( $n=100$ )					Eficiência
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	
$\beta_0$	-32,0	-33,6722	4,1026	89,1	94,4	99,1	-36,1181	7,7844	86,5	92,5	98,1	0,2531
$\beta_1$	0,77	0,8075	0,1390	88,9	94,6	98,9	0,8608	0,2449	88,2	93,7	98,4	0,3035
$\beta_2$	2,20	2,3126	0,3232	89,6	94,8	98,9	2,4883	0,6007	87,4	93,0	98,0	0,2637
$\beta_3$	1,00	1,0539	0,1503	88,9	94,4	99,1	1,1316	0,2789	86,9	92,9	98,2	0,2681
$\sigma$	63,25	63,49	30,31	86,3	89,3	93,4	—	—	—	—	—	—
$\lambda$	3	3,05	0,4265	91,6	95,6	98,8	—	—	—	—	—	—

Parâmetros		Modelo com resposta de origem ( $n=200$ )					Modelo logístico usual ( $n=200$ )					Eficiência
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	
$\beta_0$	-32,0	-32,7754	2,9175	89,7	95,0	99,0	-33,7131	5,0296	88,9	94,2	98,7	0,3228
$\beta_1$	0,77	0,7886	0,0960	89,9	95,0	99,2	0,8118	0,1568	89,0	94,3	98,7	0,3650
$\beta_2$	2,20	2,2526	0,2191	89,6	94,7	99,1	2,3145	0,3738	89,2	94,3	98,8	0,3325
$\beta_3$	1,00	1,0243	0,1066	89,8	94,8	99,0	1,0537	0,1812	88,9	94,5	98,8	0,3361
$\sigma$	63,25	63,78	23,33	87,9	91,2	95,2	—	—	—	—	—	—
$\lambda$	3	3,03	0,3209	90,8	95,3	99,0	—	—	—	—	—	—

Parâmetros		Modelo com resposta de origem ( $n=500$ )					Modelo logístico usual ( $n=500$ )					Eficiência
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	
$\beta_0$	-32,0	-32,3190	1,8372	90,1	94,9	99,2	-32,6908	3,2049	89,6	94,8	99,2	0,3235
$\beta_1$	0,77	0,7777	0,0602	89,7	94,8	99,1	0,7874	0,0995	89,7	94,8	99,1	0,3627
$\beta_2$	2,20	2,2215	0,1410	90,6	95,5	99,1	2,2464	0,2443	90,0	95,0	99,0	0,3285
$\beta_3$	1,00	1,0099	0,0665	89,6	94,8	99,1	1,0209	0,1134	90,0	95,0	99,0	0,3383
$\sigma$	63,25	63,88	14,26	90,4	94,1	97,6	—	—	—	—	—	—
$\lambda$	3	3,01	0,1939	90,6	95,4	99,0	—	—	—	—	—	—

Observamos por meio da Tabela 3.6 que o modelo de regressão logística com a informação da resposta normal heteroscedástica produziu estimativas pontuais dos coeficientes mais próximas dos valores reais para todos os tamanhos amostrais. Assim como nos casos anteriores, o erro padrão obtido foi menor para todos os parâmetros produzindo estimativas intervalares com menores amplitudes, ou seja, mais precisas. O modelo logístico com resposta de origem apresenta uma cobertura dos intervalos bem mais próximos para todos os casos considerados em que  $n = 200$  e  $n = 500$ .

Com relação à eficiência, percebemos que o erro quadrático médio das estimativas segundo o modelo logístico com resposta de origem é sempre menor que o erro quadrático médio das estimativas do modelo logístico usual. Pelos valores de eficiência obtidos notamos que o erro quadrático médio das estimativas do modelo com resposta de origem é de 2,7 a 6,1 vezes menor que o erro quadrático médio das estimativas do modelo logístico usual.

### 3.7 Modelo de regressão logística com resposta exponencial

Nesta seção vamos analisar a performance do modelo de regressão logística com resposta exponencial e, assim, estudar a eficiência das estimativas dos coeficientes de regressão  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo com dados artificiais. Para a geração dos dados supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(5; 1); \quad X_{i2} \sim N(15; 4); \quad X_{i3} \sim N(45; 25); \quad R_i \sim \text{Exponencial}(\lambda_i),$$

e

$$\lambda_i = -\frac{\ln [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})]}{C_R}, \quad i = 1, 2, \dots, n.$$

Os valores atribuídos para o vetor  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -24, 0$ ,  $\beta_1 = 1, 2$ ,  $\beta_2 = 0, 61$  e  $\beta_3 = 0, 19$ .

Consideramos o ponto de corte  $C_R = 40.000$ , isto é, todos os valores de  $R_i$  acima de 40.000 foram considerados sucessos, ou seja, tiveram a resposta igual a 1 e todos os valores de  $R_i$  com valores abaixo de 40.000 foram considerados fracassos, ou seja, tiveram a resposta igual a 0. A Tabela 3.7 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta exponencial, para cada tamanho amostral, bem como as métricas de avaliação.

Tabela 3.7. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros	Modelo com resposta de origem (n=50)					Modelo logístico usual (n=50)					Eficiência	
	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%		
$\beta_0$	-24,0	-24,1787	3,2770	90,3	94,9	98,9	-28,8988	8,5071	85,2	91,1	97,5	0,1118
$\beta_1$	1,20	1,2082	0,3005	90,3	95,0	99,1	1,3954	0,6266	87,2	93,2	98,4	0,2098
$\beta_2$	0,61	0,6064	0,1150	90,7	95,2	99,2	0,7624	0,2877	85,9	91,9	97,9	0,1245
$\beta_3$	0,19	0,1925	0,0491	90,5	95,2	99,1	0,2254	0,1029	86,4	92,3	97,9	0,2034
Parâmetros	Modelo com resposta de origem (n=100)					Modelo logístico usual (n=100)					Eficiência	
	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%		
$\beta_0$	-24,0	-24,0982	2,4471	90,5	95,5	99,1	-26,1547	5,4356	88,2	93,9	98,5	0,1754
$\beta_1$	1,20	1,2017	0,1883	89,9	94,8	99,0	1,3052	0,3669	88,7	94,1	98,6	0,2437
$\beta_2$	0,61	0,6084	0,0841	90,4	95,3	99,0	0,6660	0,1718	88,3	94,2	98,6	0,2171
$\beta_3$	0,19	0,1916	0,0355	90,2	95,4	99,1	0,2069	0,0654	88,9	93,9	98,8	0,2826
Parâmetros	Modelo com resposta de origem (n=200)					Modelo logístico usual (n=200)					Eficiência	
	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%		
$\beta_0$	-24,0	-24,0321	1,8410	90,0	95,0	99,0	-24,9480	3,7810	89,5	94,4	98,7	0,2231
$\beta_1$	1,20	1,2018	0,1275	90,0	95,2	99,2	1,2502	0,2427	88,8	94,2	98,8	0,2655
$\beta_2$	0,61	0,6102	0,0569	89,9	94,9	99,0	0,6351	0,1102	88,9	94,4	98,8	0,2500
$\beta_3$	0,19	0,1900	0,0248	90,1	94,9	98,9	0,1969	0,0451	90,0	94,8	99,0	0,2857
Parâmetros	Modelo com resposta de origem (n=500)					Modelo logístico usual (n=500)					Eficiência	
	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%		
$\beta_0$	-24,0	-24,0087	1,0861	90,3	94,9	99,0	-24,3364	2,2935	89,0	94,7	98,9	0,2195
$\beta_1$	1,20	1,1996	0,0721	89,6	94,9	99,0	1,2160	0,1433	89,3	94,9	98,9	0,2500
$\beta_2$	0,61	0,6094	0,0360	90,0	94,9	98,9	0,6189	0,0712	89,7	94,9	98,8	0,2549
$\beta_3$	0,19	0,1903	0,0134	89,9	95,1	99,0	0,1927	0,0264	89,0	94,2	98,8	0,2857

Assim como nos casos anteriores, observamos por meio da Tabela 3.7 que o modelo de regressão logística com a informação da resposta exponencial produziu estimativas pontuais dos coeficientes de regressão mais próximas dos valores reais para todos os tamanhos amostrais e o erro padrão obtido foi menor para todos os parâmetros, produzindo estimativas intervalares com amplitudes menores, isto é, mais precisas.

Com relação a probabilidade de cobertura dos intervalos de confiança obtida, verificamos que para os tamanhos de amostra  $n = 50$  e  $n = 100$  o modelo logístico usual subestima os três casos considerados de probabilidade de cobertura (90%, 95% e 99%) e, por sua vez, o modelo logístico com resposta de origem apresentou uma cobertura dos intervalos bem mais próximos para todos os casos considerados.

Com relação à eficiência, percebemos que o erro quadrático médio das estimativas segundo o modelo logístico com resposta de origem é sempre menor que o erro quadrático médio das estimativas do modelo logístico usual. Pelos valores de eficiência obtidos percebemos que o erro quadrático médio das estimativas do modelo com resposta de origem é de 3,5 a 8,9 vezes menor que o erro quadrático médio das estimativas do modelo logístico usual.

### 3.8 Modelo de regressão logística com resposta geométrica

Nesta seção atribuímos a resposta discreta de origem geométrica para o modelo de regressão logística. Analisamos a performance do modelo logístico com resposta geométrica e estudamos a eficiência das estimativas dos coeficientes de regressão  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo com dados artificiais. Supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(35; 25); \quad X_{i2} \sim N(5; 1); \quad X_{i3} \sim N(10; 4); \quad R_i \sim Geometrica(p_i),$$

e

$$p_i = \left[ 1 - \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})} \right]^{C_R + 1}, \quad i = 1, 2, \dots, n.$$

Os valores atribuídos para o vetor  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $p_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -20, 0$ ;  $\beta_1 = 0, 2$ ;  $\beta_2 = 1, 0$ ;  $\beta_3 = 0, 7$ ; e uma constante  $C_R = 0$ . A Tabela 3.8 mostra as estimativas obtidas para os modelos de regressão logística usual e regressão logística com a informação da variável resposta geométrica, para cada tamanho amostral, bem como as métricas de avaliação.

Tabela 3.8. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros		Modelo com resposta de origem ( $n=50$ )					Modelo logístico usual ( $n=50$ )					
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência
$\beta_0$	-20,0	-20,8384	4,1270	91,0	95,8	99,2	-24,6214	7,6179	85,5	91,8	97,7	0,2234
$\beta_1$	0,20	0,2095	0,0808	90,4	95,6	99,2	0,2447	0,1334	87,9	93,4	98,5	0,3333
$\beta_2$	1,00	1,0119	0,2744	91,3	95,9	99,2	1,2583	0,5595	86,9	92,2	98,1	0,1985
$\beta_3$	0,70	0,7264	0,1682	91,0	96,0	99,2	0,8524	0,3039	86,3	92,1	97,9	0,2509
Parâmetros		Modelo com resposta de origem ( $n=100$ )					Modelo logístico usual ( $n=100$ )					
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência
$\beta_0$	-20,0	-20,3412	3,0487	90,1	95,0	98,9	-21,9305	4,9193	87,6	93,6	98,5	0,3370
$\beta_1$	0,20	0,2031	0,0515	89,6	95,0	99,1	0,2195	0,0759	87,8	93,7	98,4	0,4426
$\beta_2$	1,00	1,0080	0,2017	90,5	95,6	99,2	1,1011	0,3371	88,2	94,0	98,7	0,3285
$\beta_3$	0,70	0,7106	0,1262	90,4	95,0	99,1	0,7649	0,2006	87,7	93,4	98,7	0,3596
Parâmetros		Modelo com resposta de origem ( $n=200$ )					Modelo logístico usual ( $n=200$ )					
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência
$\beta_0$	-20,0	-20,1139	2,0515	90,2	95,0	99,0	-20,8096	3,2918	88,8	93,9	98,8	0,3674
$\beta_1$	0,20	0,2009	0,0331	90,5	95,2	99,0	0,2085	0,0493	88,9	94,5	98,8	0,4400
$\beta_2$	1,00	1,0057	0,1310	90,2	95,2	99,0	1,0430	0,2118	89,3	94,3	98,8	0,3683
$\beta_3$	0,70	0,7018	0,0825	90,1	95,1	99,0	0,7260	0,1320	89,1	94,4	98,9	0,3757
Parâmetros		Modelo com resposta de origem ( $n=500$ )					Modelo logístico usual ( $n=500$ )					
Valores reais		Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência
$\beta_0$	-20,0	-20,0279	1,1189	89,8	94,9	99,0	-20,3040	1,9783	89,6	94,7	98,8	0,3127
$\beta_1$	0,20	0,2003	0,0174	90,1	95,1	99,0	0,2030	0,0287	89,6	94,8	99,1	0,3750
$\beta_2$	1,00	1,0012	0,0832	90,4	95,1	99,1	1,0160	0,1368	90,4	95,1	98,8	0,3632
$\beta_3$	0,70	0,7000	0,0427	89,8	94,8	99,0	0,7101	0,0786	89,8	94,8	99,0	0,2857

Novamente observamos por meio da Tabela 3.8 que o modelo de regressão logística com resposta geométrica produz estimativas pontuais dos coeficientes de regressão mais próximas dos valores reais para todos os tamanhos amostrais. O erro padrão obtido foi menor para todos os parâmetros. Verificamos também que, à medida que o tamanho da amostra aumenta as estimativas pontuais de ambos os modelos se aproximam e, no entanto, o erro padrão continua sendo menor para todos os tamanhos de amostra considerados. Nesse sentido, o modelo logístico com resposta de origem produz estimativas mais precisas que o modelo logístico usual.

Considerando a probabilidade de cobertura dos intervalos de confiança obtidas, verificamos que para os tamanhos de amostra  $n = 50$  e  $n = 100$  o modelo logístico usual subestima os três casos considerados de probabilidade de cobertura (90%, 95% e 99%). Por outro lado, o modelo logístico com resposta de origem apresenta uma cobertura dos intervalos mais próximos dos casos considerados.

Com relação à eficiência, percebemos que o erro quadrático médio das estimativas segundo o modelo logístico com resposta de origem é sempre menor que o erro quadrático médio das estimativas do modelo logístico usual. Pelos valores de eficiência obtidos observamos que o erro quadrático médio das estimativas do modelo com resposta de origem é de 2, 3 a 5 vezes menor que o erro quadrático médio das estimativas do modelo logístico usual.

### 3.9 Modelo de regressão logística com resposta Poisson

Nesta seção vamos analisar a performance do modelo de regressão logística com a informação da variável resposta Poisson e, assim, estudar a eficiência das estimativas dos coeficientes de regressão  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo com dados artificiais. Supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(35; 25); X_{i2} \sim N(5; 1); X_{i3} \sim N(10; 4); R_i \sim Poisson(\lambda_i);$$

Considerando  $C_R = 0$  temos que

$$\lambda_i = -\ln \left[ 1 - \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})} \right], i = 1, 2, \dots, n.$$

Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -10, 0$ ,  $\beta_1 = 0, 16$ ,  $\beta_2 = 1, 22$  e  $\beta_3 = 0, 48$ .

A Tabela 3.9 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com a informação da variável resposta Poisson, para cada tamanho amostral bem como as métricas de avaliação.

Tabela 3.9. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros	Modelo com resposta de origem ( $n=50$ )					Modelo logístico usual ( $n=50$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-15,0	-15,3912	2,9938	90,3	95,3	99,2	-18,0694	6,3559	86,6	92,5	97,9	0,1830
$\hat{\beta}_1$	0,16	0,1620	0,0643	90,6	95,4	99,2	0,1894	0,1057	86,9	92,6	98,0	0,3417
$\hat{\beta}_2$	1,22	1,2525	0,2839	90,2	95,3	99,3	1,4759	0,5759	85,9	92,2	98,0	0,2057
$\hat{\beta}_3$	0,48	0,4911	0,1363	90,1	95,6	99,1	0,5854	0,2685	87,0	92,9	98,3	0,2248
Parâmetros	Modelo com resposta de origem ( $n=100$ )					Modelo logístico usual ( $n=100$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-15,0	-15,3141	2,1320	90,6	95,4	99,1	-16,2968	3,5696	88,4	93,8	98,6	0,3220
$\hat{\beta}_1$	0,16	0,1632	0,0413	89,8	95,0	99,0	0,1728	0,0627	89,0	94,0	98,8	0,4146
$\hat{\beta}_2$	1,22	1,2404	0,1973	90,7	95,3	99,1	1,3265	0,3530	88,0	93,4	98,6	0,2890
$\hat{\beta}_3$	0,48	0,4879	0,0954	90,0	95,1	99,1	0,5229	0,1525	88,6	94,0	98,6	0,3665
Parâmetros	Modelo com resposta de origem ( $n=200$ )					Modelo logístico usual ( $n=200$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-15,0	-15,1431	1,5782	90,0	95,0	98,9	-15,6214	2,6195	89,4	94,6	98,8	0,3465
$\hat{\beta}_1$	0,16	0,1612	0,0281	90,0	95,0	99,0	0,1663	0,0440	89,2	94,4	98,9	0,4000
$\hat{\beta}_2$	1,22	1,2300	0,1297	90,2	95,3	99,0	1,2711	0,2297	89,1	94,6	98,9	0,3051
$\hat{\beta}_3$	0,48	0,4839	0,0691	89,9	94,9	99,0	0,4994	0,1111	90,1	95,1	98,9	0,3780
Parâmetros	Modelo com resposta de origem ( $n=500$ )					Modelo logístico usual ( $n=500$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-15,0	-15,0475	0,9760	90,3	95,5	99,0	-15,2667	1,6464	90,0	95,0	98,8	0,3432
$\hat{\beta}_1$	0,16	0,1605	0,0169	89,8	94,8	99,0	0,1629	0,0277	89,4	94,6	98,8	0,3750
$\hat{\beta}_2$	1,22	1,2230	0,0849	90,2	95,0	99,0	1,2412	0,1465	90,3	95,2	99,0	0,3288
$\hat{\beta}_3$	0,48	0,4811	0,0413	90,4	95,2	99,1	0,4878	0,0691	89,9	95,1	99,1	0,3542

Assim como no caso geométrico, observamos novamente por meio da Tabela 3.9 que o modelo de regressão logística com a informação da variável resposta Poisson produziu estimativas pontuais dos coeficientes de regressão mais próximas dos valores reais para todos os tamanhos amostrais e o erro padrão obtido foi menor para todos os parâmetros produzindo estimativas intervalares mais precisas. A medida que o tamanho da amostra aumenta as estimativas pontuais de ambos os modelos se aproximam e o erro padrão continua sendo menor para todos os tamanhos de amostra considerados. Novamente o modelo logístico com resposta de origem produz assim estimativas mais precisas que o modelo logístico usual.

Considerando a probabilidade de cobertura dos intervalos de confiança obtidas, verificamos que para os tamanhos de amostra  $n = 50$  e  $n = 100$  o modelo logístico usual subestima os três casos considerados de probabilidade de cobertura (90%, 95% e 99%) e o modelo logístico com resposta Poisson apresentou uma cobertura dos intervalos mais próximos dos casos considerados.

Pelos valores de eficiência obtidos percebemos que o erro quadrático médio das estimativas do modelo com resposta de origem é de 2,4 a 5,5 vezes menor que o erro quadrático médio das estimativas do modelo logístico usual.

### 3.10 Modelo geométrico com resposta Normal

A exemplo do modelo de regressão logística com resposta geométrica, nesta seção analisamos a performance do modelo de regressão geométrica com resposta normal e estudamos a eficiência das estimativas dos coeficientes de regressão  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  por meio de um estudo de simulação. Para a geração dos dados supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(20; 16); R_i \sim N(\mu_i, \sigma^2),$$

com  $\sigma = 1.000$  e

$$\mu_i = \sigma \Phi^{-1} [1 - g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})]^{C_Y + 1} + C_R, \quad i = 1, 2, \dots, n.$$

Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\mu_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -4,5$ ;  $\beta_1 = 0,5$ ;  $\beta_2 = 0,35$ ;  $\beta_3 = 0,02$ , e o ponto de corte  $C_R = 11.000$ . Para a geração do vetor  $Y_1, Y_2, \dots, Y_n$  consideramos

$$Y_i \sim \text{Geométrica} [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})], \quad i = 1, 2, \dots, n,$$

e o ponto de corte  $C_Y = 2$ . Dessa forma temos a seguinte relação

$$P(R_i > 11.000) = P(Y_i > 2), \quad i = 1, 2, \dots, n.$$

A Tabela 3.10 mostra as estimativas obtidas considerando o modelo de regressão geométrico usual e o modelo de regressão geométrico com resposta normal, para cada tamanho amostral bem como as métricas de avaliação.

Tabela 3.10. Estimativas dos parâmetros para cada modelo.

Parâmetros	Modelo com resposta de origem ( $n=50$ )					Modelo geométrico usual ( $n=50$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,6706	1,6728	91,0	96,0	99,4	-4,7399	3,7776	86,7	91,5	94,2	0,1973
$\hat{\beta}_1$	0,50	0,5218	0,2103	91,4	96,0	99,5	0,5821	0,4392	87,2	91,9	94,4	0,2239
$\hat{\beta}_2$	0,35	0,3677	0,1049	93,0	97,0	99,5	0,3586	0,2102	86,6	90,8	94,0	0,2551
$\hat{\beta}_3$	0,02	0,0203	0,0483	89,4	94,9	99,2	0,0233	0,1094	87,0	91,6	94,4	0,1917
$\hat{\sigma}$	1000	975	102	96,6	98,2	99,4	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem ( $n=100$ )					Modelo geométrico usual ( $n=100$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,6537	1,0149	89,6	94,6	99,1	-4,4646	1,8566	90,7	95,8	99,3	0,3056
$\hat{\beta}_1$	0,50	0,5223	0,1351	89,7	94,6	98,9	0,5163	0,2580	91,1	95,8	99,5	0,2799
$\hat{\beta}_2$	0,35	0,3640	0,0694	90,2	95,2	99,2	0,3517	0,1261	90,3	95,5	99,2	0,3145
$\hat{\beta}_3$	0,02	0,0204	0,0285	89,7	94,9	98,9	0,0193	0,0526	90,9	95,6	99,4	0,2857
$\hat{\sigma}$	1000	978	69	87,6	92,7	97,4	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem ( $n=200$ )					Modelo geométrico usual ( $n=200$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,5726	0,8024	90,0	95,4	99,1	-4,4353	1,4745	89,8	95,2	99,2	0,2980
$\hat{\beta}_1$	0,50	0,5087	0,0962	90,0	95,1	99,1	0,4998	0,1841	90,3	95,7	99,2	0,2743
$\hat{\beta}_2$	0,35	0,3568	0,0482	90,1	94,9	99,0	0,3521	0,0898	90,0	95,3	99,1	0,2963
$\hat{\beta}_3$	0,02	0,0205	0,0215	89,4	94,9	98,8	0,0191	0,0410	90,4	95,5	99,3	0,2941
$\hat{\sigma}$	1000	989	49	89,1	94,3	98,3	—	—	—	—	—	—
Parâmetros	Modelo com resposta de origem ( $n=500$ )					Modelo geométrico usual ( $n=500$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,5300	0,5025	90,4	95,2	99,1	-4,4536	0,9067	90,0	94,9	98,9	0,3074
$\hat{\beta}_1$	0,50	0,5036	0,0588	90,3	94,9	98,9	0,4979	0,1154	90,2	95,3	99,2	0,2632
$\hat{\beta}_2$	0,35	0,3528	0,0307	90,6	95,3	99,1	0,3501	0,0555	90,4	95,1	99,0	0,3226
$\hat{\beta}_3$	0,02	0,0202	0,0131	89,9	94,7	98,8	0,0194	0,0267	90,2	95,3	99,1	0,2857
$\hat{\sigma}$	1000	996	31	91,1	95,6	98,9	—	—	—	—	—	—

A partir da Tabela 3.10 notamos que o modelo de regressão geométrico com a informação da variável resposta normal fornece estimativas pontuais dos coeficientes de regressão mais próximas dos verdadeiros valores para todos os tamanhos amostrais. A medida que o tamanho  $n$  da amostra aumenta, as estimativas pontuais segundo o modelo com resposta de origem tendem a se aproximar das estimativas pontuais segundo modelo de regressão usual. O erro padrão obtido foi menor para todos os parâmetros do modelo produzindo estimativas intervalares mais precisas, para todos os tamanhos de amostra.

Com relação as probabilidades de coberturas obtidas dos intervalos de confiança, o modelo de regressão geométrico com resposta normal apresenta uma cobertura dos intervalos mais próxima do real para todos os casos considerados.

Quanto a eficiência, os valores obtidos sugerem que o erro quadrático médio das estimativas segundo o modelo geométrico com resposta normal é de 3 a 5 vezes menor que o erro quadrático médio segundo o modelo geométrico usual.

### 3.11 Modelo geométrico com resposta exponencial

Por meio de dados simulados, analisamos nesta seção a performance do modelo geométrico com resposta exponencial e estudamos a eficiência das estimativas dos coeficientes de regressão  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ . Para a geração dos dados supomos três variáveis explicativas da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(20; 16); \lambda_i \sim \text{Exponencial}(\lambda_i),$$

com

$$\lambda_i = -\frac{1}{C_R} \ln \left[ (1 - p_i)^{C_Y + 1} \right], \quad i = 1, 2, \dots, n,$$

e  $p_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$ .

Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\lambda_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -4,5$ ;  $\beta_1 = 0,5$ ;  $\beta_2 = 0,35$ ;  $\beta_3 = 0,02$  e os pontos de corte  $C_R = 11.000$  e  $C_Y = 1$ . Para a geração do vetor  $Y_1, Y_2, \dots, Y_n$  consideramos

$$Y_i \sim \text{Geométrica} \left[ g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) \right], \quad i = 1, 2, \dots, n,$$

A Tabela 3.11 mostra as estimativas obtidas considerando o modelo de regressão geométrico usual e o modelo de regressão geométrico com resposta normal, para cada tamanho amostral, bem como as métricas de avaliação.

Tabela 3.11. Estimativas dos parâmetros para cada modelo e métricas de avaliação.

Parâmetros	Modelo com resposta de origem ( $n=50$ )					Modelo geométrico usual ( $n=50$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,2789	3,0978	91,2	95,8	99,4	-5,1554	4,6514	88,7	91,1	96,2	0,4371
$\hat{\beta}_1$	0,50	0,5130	0,3701	91,6	96,0	99,3	0,6071	0,5431	88,2	91,7	95,4	0,4475
$\hat{\beta}_2$	0,35	0,3466	0,1669	91,0	95,9	99,4	0,4228	0,2650	86,6	92,4	96,0	0,3695
$\hat{\beta}_3$	0,02	0,0150	0,0719	91,4	96,0	99,3	0,0186	0,1036	87,2	91,3	96,1	0,4860
Parâmetros	Modelo com resposta de origem ( $n=100$ )					Modelo geométrico usual ( $n=100$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,3463	1,8559	90,3	95,5	99,3	-4,4189	2,2417	90,8	95,7	99,3	0,6892
$\hat{\beta}_1$	0,50	0,5049	0,2385	90,3	95,0	99,2	0,5198	0,2927	91,4	96,1	99,4	0,6609
$\hat{\beta}_2$	0,35	0,3445	0,1231	90,9	95,6	99,2	0,3588	0,1543	91,0	95,6	99,4	0,6360
$\hat{\beta}_3$	0,02	0,0176	0,0524	90,9	95,5	99,0	0,0145	0,0620	91,4	96,1	99,5	0,7179
Parâmetros	Modelo com resposta de origem ( $n=200$ )					Modelo geométrico usual ( $n=200$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,4022	1,3445	90,5	95,4	99,2	-4,4213	1,5775	90,4	95,7	99,4	0,7284
$\hat{\beta}_1$	0,50	0,4949	0,1546	90,5	95,4	99,0	0,5001	0,1833	90,1	95,1	99,1	0,7113
$\hat{\beta}_2$	0,35	0,3484	0,0871	90,0	95,2	99,1	0,3520	0,1036	90,5	95,5	99,1	0,7103
$\hat{\beta}_3$	0,02	0,0189	0,0394	90,7	95,6	99,2	0,0184	0,0456	90,7	95,4	99,3	0,7619
Parâmetros	Modelo com resposta de origem ( $n=500$ )					Modelo geométrico usual ( $n=500$ )						
Valores reais	Estimativa	EP	90%	95%	99%	Estimativa	EP	90%	95%	99%	Eficiência	
$\hat{\beta}_0$	-4,5	-4,4722	0,8542	90,1	95,0	99,0	-4,4794	1,0011	90,2	95,0	99,0	0,7285
$\hat{\beta}_1$	0,50	0,4997	0,0972	89,9	94,8	99,0	0,5015	0,1139	90,4	95,3	99,1	0,7231
$\hat{\beta}_2$	0,35	0,3490	0,0518	90,5	95,4	99,1	0,3503	0,0615	90,0	95,2	99,0	0,7105
$\hat{\beta}_3$	0,02	0,0200	0,0254	90,1	95,0	99,0	0,0198	0,0292	90,5	95,2	99,1	0,6667



Pelos resultados apresentados na Tabela 3.11, observamos que o modelo de regressão geométrica com a informação da variável resposta exponencial, a exemplo do caso anterior com resposta normal, fornece estimativas pontuais dos coeficientes de regressão mais próximas dos verdadeiros valores para todos os tamanhos amostrais. A medida que o tamanho  $n$  da amostra aumenta, as diferenças entre as estimativas e os valores reais tendem a zero. O erro padrão obtido é menor para todos os estimadores, produzindo estimativas intervalares mais precisas para todos os tamanhos de amostra.

Com relação as probabilidades de coberturas obtidas dos intervalos de confiança, o modelo de regressão geométrico com resposta exponencial apresenta cobertura intervalares mais próximas das nominais para todos os casos considerados.

Quanto a eficiência, os valores obtidos sugerem que o erro quadrático médio das estimativas, segundo o modelo geométrico com resposta exponencial, é de 1,3 a 1,6 vezes menor que o erro quadrático médio segundo o modelo geométrico usual.

### 3.12 Análise de resíduos da variável de interesse binária

Considerando as mesmas 10.000 amostras de tamanho  $n$  geradas no estudo com dados artificiais apresentados na seção anterior, ajustamos  $\hat{\pi}_i$  por meio de  $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ ,  $i = 1, 2, \dots, n$ , segundo as estimativas dos coeficientes do modelo logístico usual e dos coeficientes do modelo logístico com a informação da variável de origem. A tabela 3.12 apresenta os resultados obtidos, em que **% média:** é a porcentagem de amostras ajustadas pelo modelo com resposta de origem, dentre as 10.000 amostras (geradas e ajustadas segundo ambos os modelos, origem e usual), que apresentou um resíduo médio menor que o resíduo médio do ajuste logístico usual, e **% desvio:** é a porcentagem de amostras ajustadas pelo modelo com resposta de origem, dentre as 10.000 amostras (geradas e ajustadas segundo ambos os modelos, origem e usual), que apresentou um desvio-padrão dos resíduos menor que o desvio-padrão dos resíduos do ajuste logístico usual.

Tabela 3.12. Análise de resíduos das probabilidades de sucesso.

	Tamanho da amostra							
	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
Modelos:	% Média	% Desvio	% Média	% Desvio	% Média	% Desvio	% Média	% Desvio
Normal	57,60%	99,85%	55,20%	100%	55,57%	100%	54,63%	100%
Exponencial	73,72%	97,87%	72,23%	99,61%	49,99%	100%	48,99%	100%
Log-normal	60,81%	98,27%	57,93%	99,85%	54,92%	100%	52,46%	100%
Normal Het.	82,76%	96,08%	71,71%	99,37%	57,71%	99,99%	56,69%	100%
Geométrica	74,53%	97,36%	60,90%	98,75%	48,26%	99,91%	50,93%	100%
Poisson	51,40%	97,73%	49,61%	99,81%	50,38%	100%	49,51%	100%

Podemos observar pela Tabela 3.12 que, para pequenas amostras, a quantidade de amostras modeladas segundo o modelo de origem que apresenta um resíduo médio menor que o resíduo médio do modelo usual é maior. À medida que o tamanho da amostra aumenta, essa porcentagem tende a convergir para 50%, o que sugere que, para amostras grandes, o resíduo médio do modelo de origem é o mesmo do modelo logístico usual.

Verificamos, claramente, que, para todo tamanho de amostra, o desvio dos resíduos segundo o modelo com resposta de origem é, em mais de 96% das vezes, menor que o desvio dos resíduos segundo o modelo logístico usual. Para amostras grandes isso é verdade em 100% das vezes. Esse fato sugere que, os modelos com resposta de origem gozam de um melhor ajuste dos dados em pelo menos 96% das vezes.

Em síntese, a vantagem do ajuste dos dados, segundo o modelo com resposta de origem, é especialmente verificada nas pequenas amostras em que o resíduo médio é menor. Os resíduos do modelo de origem possuem uma variação menor que o modelo logístico para qualquer tamanho de amostra.

### 3.13 Análise de influência considerando os dados artificiais

Nesta Seção analisamos a performance do ajuste dos modelos por meio da detecção e deleção de observações influentes. Considerando a matriz de projeção  $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ , temos que o *i-ésimo* elemento da diagonal de  $\mathbf{H}$ ,  $0 \leq h_{ii} = \mathbf{X}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_i^T \leq 1$  é usado para determinar se o *i-ésimo* ponto amostral é influente, com relação a  $\mathbf{X}$ , ou seja, é uma medida de distância entre o  $\mathbf{X}_i$  e o valor central de todos os outros casos. Valores de  $h_{ii}$  próximos de zero indicam que o ponto amostral não é *outlier* ou influente. Valores de  $h_{ii}$  próximos de um indicam que o ponto é influente. Dessa forma, Consideramos como pontos influentes todos os pontos tais que

$$h_{ii} > 2\bar{h} = \frac{2 \sum h_{ii}}{n} = \frac{2p}{n},$$

em que  $p$  é o número de coeficientes de regressão no modelo considerado. Dessa maneira, obtemos quatro contextos de ajustes a saber:

- **Ajuste 1.** Modelo completo: ajuste do modelo logístico usual considerando todas as  $n$  observações para cada uma das 10.000 amostras geradas.
- **Ajuste 2.** Modelo completo: Ajuste do modelo logístico com resposta de origem considerando todas as  $n$  observações para cada uma das 10.000 amostras geradas.
- **Ajuste 3.** Modelo perturbado: Ajuste do modelo logístico usual com a deleção das observações influentes ( $h_{ii} > 2\bar{h}$ ) para cada uma das 10.000 amostras geradas.
- **Ajuste 4.** Modelo perturbado: Ajuste do modelo logístico com resposta de origem com a deleção das observações influentes ( $h_{ii} > 2\bar{h}$ ) para cada uma das 10.000 amostras geradas.

Fizemos o ajuste dos modelos com perturbação, isto é, o ajuste considerando a deleção das observações influentes para cada uma das 10.000 amostras geradas de tamanho  $n$ , a fim de monitorar a acurácia e a probabilidade de cobertura 90%, 95% e 99% dos intervalos de confiança. Verificamos que, para os modelos aproximadamente simétricos, a exclusão dos pontos influentes no ajuste dos dados segundo o modelo logístico com resposta de origem (modelo perturbado) não altera significativamente as estimativas obtidas para os coeficientes de regressão do modelo em comparação ao ajuste considerando todas as observações (modelo

completo). Comparando com as estimativas obtidas segundo o ajuste do modelo de regressão usual, o modelo de regressão com resposta de origem ainda apresenta estimativas pontuais e intervalares mais precisas. O mesmo ocorreu com o modelo geométrico com resposta de origem.

### 3.14 Análise de influência considerando uma amostra particular

Afim de analisar a sensibilidade das estimativas dos coeficientes de regressão com resposta de origem sem os pontos influentes, consideramos uma amostra particular de tamanho  $n = 50$ , dentre as 10.000 amostras geradas anteriormente, do modelo de regressão logística com resposta normal considerando uma estrutura heteroscedástica multiplicativa. Recordemos que, para o modelo logístico com resposta normal heteroscedástica temos a seguinte geração:

$$\begin{aligned} X_{i1} &\sim N(10; 4); X_{i2} \sim N(5; 1); X_{i3} \sim N(15; 4); R_i \sim N(\mu_i, \sigma_i^2), \\ \mu_i &= \sigma x_{1i}^{\lambda/2} \Phi^{-1} [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + C_R, i = 1, 2, \dots, n, \\ \sigma_i^2 &= x_{1i}^\lambda \sigma^2, i = 1, 2, \dots, n. \end{aligned}$$

Os valores atribuídos para os coeficientes de regressão, para a geração de  $\mu_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -32, 0$ ,  $\beta_1 = 0, 77$ ,  $\beta_2 = 2, 2$  e  $\beta_3 = 1, 0$ . Para os parâmetros perturbadores atribuímos  $\sigma^2 = 4000$  e  $\lambda = 3$ . Consideramos o ponto de corte  $C_R = 10.000$ , isto é, todos os valores de  $R_i$  acima de 10.000 foram considerados sucessos. A amostra particular, sorteada dentre todas as amostras, tem 27 sucessos.

A Tabela 3.14 mostra as estimativas obtidas dos coeficientes de regressão considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta normal heteroscedástica.

Tabela 3.14. Estimativas obtidas segundo o ajuste de cada modelo.

Resultados do modelo logístico usual				
Parâmetros	Estimativa	Erro-Padrão	Int.Conf(95%)	
$\beta_0 = -32, 0$	-47, 0237	14, 7001	-84, 3963	-24, 6777
$\beta_1 = 0, 77$	0, 7963	0, 3773	0, 1377	1, 6698
$\beta_2 = 2, 20$	3, 0105	0, 9383	1, 5381	5, 3724
$\beta_3 = 1, 00$	1, 7417	0, 6264	0, 7905	3, 3429
Resultados do modelo logístico com resposta normal heteroscedástica				
Parâmetros	Estimativa	Erro-Padrão	Int.Conf(95%)	
$\beta_0 = -32, 0$	-38, 7055	6, 6534	-51, 7459	-25, 6652
$\beta_1 = 0, 77$	0, 8955	0, 2109	0, 4821	1, 3089
$\beta_2 = 2, 20$	2, 3166	0, 4615	1, 4120	3, 2211
$\beta_3 = 1, 00$	1, 3336	0, 2628	0, 8185	1, 8488
$\sigma = 63, 25$	75, 56	34, 22	8, 4810	142, 63
$\lambda = 3$	2, 8747	0, 4127	2, 0659	3, 6835

Notamos, a partir da Tabela 3.14 que as estimativas pontuais dos coeficientes de regressão segundo o modelo de regressão com resposta de origem são mais próximas dos reais valores comparando com as estimativas pontuais segundo o modelo logístico usual. O

erro padrão das estimativas também são menores produzindo assim estimativas intervalares mais precisas.

A Figura 3.1 apresenta os resíduos padronizados das probabilidades de sucesso considerando o ajuste do modelo logístico usual.

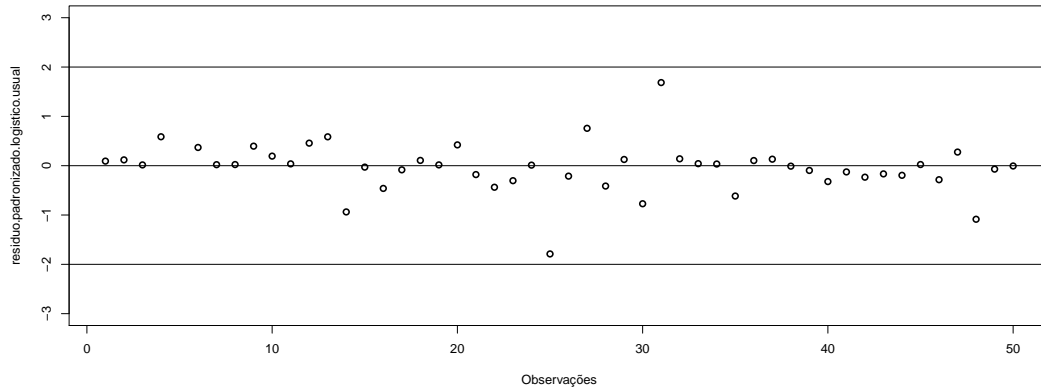


Figura 3.1: Resíduos padronizados considerando o ajuste do modelo logístico usual.

A Figura 3.2 apresenta os resíduos padronizados considerando o ajuste do modelo logístico com resposta de origem.

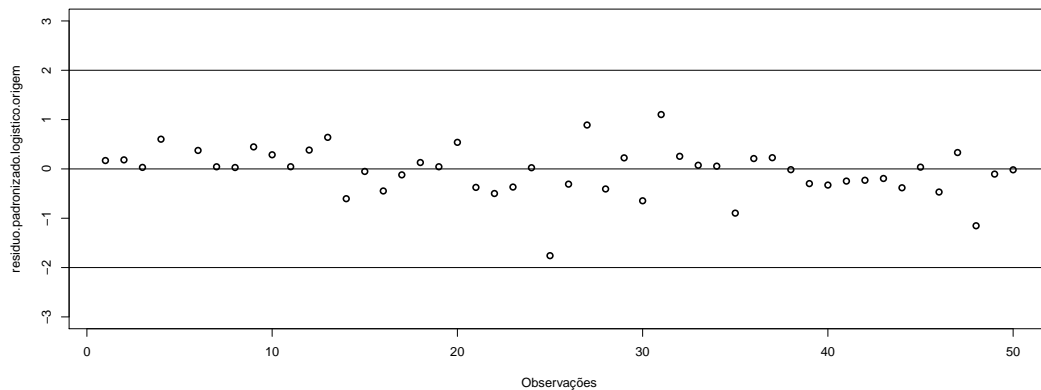


Figura 3.2: Resíduos padronizados do modelo logístico com resposta normal heteroscedástica.

Com exceção do ponto 5 em ambos os ajustes, todas as demais 49 observações forneceram um resíduo padronizado com valores entre  $-2$  e  $2$ . O modelo logístico usual forneceu o quinto resíduo padronizado igual a  $5,98$  e o modelo logístico com resposta de origem forneceu o quinto resíduo padronizado igual a  $6,68$ . Desta maneira retiramos este ponto influente do conjunto de dados e ajustamos novamente ambos os modelos. Os resultados encontram-se na Tabela 3.14.

Tabela 3.14. Estimativas segundo o ajuste de cada modelo (sem o *outlier* 5).

Resultados do modelo logístico usual				
Parâmetros	Estimativa	Erro-Padrão	Int.Conf(95%)	
$\beta_0 = -32,0$	-119,7887	62,2790	-323,7265	-47,4846
$\beta_1 = 0,77$	2,9695	1,5030	1,0124	7,4785
$\beta_2 = 2,20$	8,6396	4,6780	3,2354	23,7834
$\beta_3 = 1,00$	3,4232	1,8680	1,2608	9,6101
Resultados do modelo logístico com resposta normal heteroscedástica				
Parâmetros	Estimativa	Erro-Padrão	Int.Conf(95%)	
$\beta_0 = -32,0$	-42,7793	7,3978	-57,2788	-28,2798
$\beta_1 = 0,77$	1,0570	0,2372	0,5920	1,5219
$\beta_2 = 2,20$	2,6863	0,5151	1,6767	3,6960
$\beta_3 = 1,00$	1,3865	0,2765	0,8446	1,9285
$\sigma = 63,25$	127,54	73,94	-17,38	272,46
$\lambda = 3$	2,3665	0,5006	1,3854	3,3477

Ao retirarmos o ponto influente da amostra, verificamos pela Tabela 3.14 que as estimativas dos coeficientes de regressão segundo o modelo logístico usual são mais sensíveis. Por outro lado, o modelo logístico com resposta de origem contínua produzindo estimativas pontuais e intervalares mais precisas.

## 4 Modelo de regressão logística com resposta original pertencente à classe de distribuições série de potências inflacionadas

### 4.1 Introdução

Neste Capítulo propomos o modelo de regressão logística com resposta pertencente à classe das distribuições série de potências, que foi desenvolvida no contexto de modelagem de dados discretos de contagem tendo como casos particulares as distribuições geométrica, Poisson, binomial, binomial negativa, entre outras, bem como suas respectivas versões generalizadas (ver Murat e Szynal, 1998). Essa classe pode ser considerada para uma variedade de aplicações obtendo bons ajustes (Gupta *et al.* 1995).

### 4.2 Classe de distribuições série de potências

A modelagem de dados discretos pode ser feita por meio da classe geral das distribuições série de potências. Esta classe engloba modelos comumente utilizados na literatura, cuja forma geral da função de probabilidade é especificada como

$$P(R = r) = \frac{a(r) [g(\theta)]^r}{f(\theta)}, \quad r = 0, 1, 2, \dots$$

em que  $a(r) > 0$  e é independente de  $\theta$ ,  $\theta > 0$ ,  $f(\theta) = \sum_r a(r) [g(\theta)]^r < \infty$ ,  $g(\theta)$  é positiva, finita, diferenciável e inversível.

### 4.3 Função de verossimilhança

Consideremos  $n$  variáveis aleatórias  $R_1, R_2, \dots, R_n$  tal que  $R_i$  tem distribuição pertencente à classe das distribuições série de potências com parâmetro  $\theta_i$ ,  $i = 1, 2, \dots, n$ . Dessa forma temos a seguinte distribuição de probabilidades

$$P(R_i = r_i) = \frac{a(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)}, \quad i = 1, 2, \dots, n \text{ e } r_i = 0, 1, 2, \dots \quad (4.3.1)$$

Dessa maneira, a partir de (4.3.1), e supondo  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  um vetor de realizações das variáveis respostas  $R_1, R_2, \dots, R_n$ , a função de verossimilhança é tal que

$$L(\boldsymbol{\theta} | \mathbf{r}) = \prod_{i=1}^n \frac{a(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)},$$

e o logaritmo da função de verossimilhança é dada por

$$l(\boldsymbol{\theta} | \mathbf{r}) = \ln L(\boldsymbol{\theta} | \mathbf{r}) = \sum_{i=1}^n \ln a(r_i) + \sum_{i=1}^n r_i \ln g(\theta_i) - \sum_{i=1}^n \ln f(\theta_i).$$

As estimativas dos parâmetros do modelo são obtidas resolvendo o sistema

$$\frac{\partial l(\boldsymbol{\theta} | \mathbf{r})}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, n.$$

#### 4.4 Modelo de regressão logística com resposta original pertencente à classe de distribuições série de potências

Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  com distribuição de probabilidades dada em (4.3.1). Seja  $C_R$  uma constante arbitrária, ( $C_R \in \mathbb{R}$ ) no suporte da densidade de  $R_i$  e considere  $n$  variáveis aleatórias independentes binárias  $Y_1, Y_2, \dots, Y_n$ . Conforme já visto anteriormente, como as variáveis  $Y_i$  são dicotômicas verificamos que a probabilidade  $P(Y_i > 0)$ , é igual a probabilidade de sucesso da variável  $Y_i$ , ou seja,  $P(Y_i = 1)$ , e a probabilidade  $P(Y_i \leq 0)$  é a probabilidade de fracasso da variável  $Y_i$ , ou seja  $P(Y_i = 0)$ . Dessa forma, temos que  $P(R_i > C_R) = P(Y_i = 1) = \pi_i$  e  $P(R_i \leq C_R) = P(Y_i = 0) = 1 - \pi_i$ ,  $i = 1, 2, \dots, n$  e segue que

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad i = 1, 2, \dots, n,$$

em que  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ ,  $i = 1, 2, \dots, n$ , é a probabilidade de sucesso do modelo de regressão logística.

Para incorporarmos a informação da distribuição da variável resposta no modelo de regressão logística vamos considerar novamente a relação  $P(R_i > C_R) = P(Y_i = 1)$ . Dessa maneira, a probabilidade da variável resposta  $R_i$  ser maior que uma constante arbitrária  $C_R$ ,  $C_R \in \mathbb{R}$ , é tal que

$$\begin{aligned} P(R_i > C_R) &= 1 - P(R_i \leq C_R) \\ &= 1 - \sum_{r_i=0}^{C_R} \frac{a(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)}, \quad i = 1, 2, \dots, n, \end{aligned} \quad (4.4.1)$$

e vamos assumir que a probabilidade  $P(R_i > C_R)$  seja igual a probabilidade de sucesso  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$  da variável resposta binária  $Y_i$ , isto é

$$P(R_i > C_R) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}). \quad (4.4.2)$$

Portanto, a partir de (4.4.1) e assumindo (4.4.2), segue a igualdade

$$1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) = \sum_{r_i=0}^{C_R} \frac{a(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)},$$

o que implica

$$\mathbf{x}_i^T \boldsymbol{\beta} = g \left[ 1 - \sum_{r_i=0}^{C_R} \frac{a(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)} \right],$$

em que  $g$  é uma função monótona e diferenciável e  $g[(\cdot)]$  é a função de ligação composta que gera o preditor linear  $\mathbf{x}_i^T \boldsymbol{\beta}$ . Dessa maneira, a relação entre a média  $\mu_i$  de uma observação  $R_i$  e seu preditor linear  $\mathbf{x}_i^T \boldsymbol{\beta}$  fica assim caracterizada por meio de

$$g[(1 - F_{R_i}^{C_R})] = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, \quad i = 1, 2, \dots, n. \quad (4.4.3)$$

## 4.5 Classe de distribuições série de potências inflacionadas

### 4.5.1 Introdução

Os modelos discretos, largamente desenvolvidos na literatura, são atribuídos em diversas aplicações práticas que envolvem dados reais de contagens. Entre estes modelos podemos citar Poisson, binomial e binomial negativa. Em geral, os conjuntos de dados podem conter um número excessivo de um determinado valor, como, por exemplo, um número excessivo de zeros que não são descritos ou comportados pelo modelo proposto.

Muitas vezes, dados discretos de contagem apresentam valores inflacionados, como por exemplo, o valor zero, sendo observado com uma frequência significativamente maior que o admitido pelo modelo assumido. Nesse sentido, a classe de distribuição série de potência pode ser estendida para as distribuições inflacionadas. A distribuição de Poisson inflacionada é um caso particular desta classe de distribuições. Se o valor zero é o valor inflacionado, então trata-se do modelo Poisson inflacionado em zero (ZIP).

### 4.5.2 Excesso de valores em dados de contagens

Tais zeros excessivos podem ser classificados de quatro formas em que dois podem ser definidos como zeros verdadeiros e dois como zeros falsos (ou aleatórios). Considerando a primeira forma, os zeros verdadeiros são resultados de uma baixa frequência de ocorrência ou realmente não havia nenhum indivíduo presente no local. Na segunda forma, o indivíduo existe, pertence ao local, porém não estava presente durante o período da pesquisa ou o indivíduo pertence ao local, está presente, porém o pesquisador não o encontra por algum motivo.

Uma alternativa para explicar ou modelar os zeros verdadeiros ou falsos é utilizar uma distribuição pertencente à classe de distribuições zero inflacionadas. Essa classe de distribuições exige algum conhecimento sobre misturas de modelos, que neste contexto, considera duas distribuições: uma distribuição degenerada no ponto zero e uma distribuição que se adequaria aos dados caso não existisse o excesso de zeros. Numa situação onde os zeros inflacionados são resultados de excesso de zeros verdadeiros e falsos, não há nenhuma discussão formal na literatura de como modelar tais conjuntos de dados, pelo fato de ser difícil e, na maioria das vezes, inviável distinguir a origem desses zeros.

Numa situação onde há a incerteza sobre a sua origem nas observações, um procedimento usual, é utilizar distribuições truncadas. Em alguns conjuntos de dados o valor inflacionado não é o valor zero, porém tal classe de distribuições pode ser aplicada para solucionar esse número excessivo de valores diferentes de zero. Atualmente a classe de distribuições série de potências inflacionadas pode ser aplicada nas mais diversas áreas, como por exemplo, ecologia, atuária, controle de qualidade dentre outras.

### 4.5.3 Excesso de zero em dados de contagem

Em diversos estudos envolvendo dados discretos de contagem é comum a existência de uma grande quantidade de zero nos dados. A presença do excesso de zeros dificulta a elaboração de uma análise estatística para o problema, pois os modelos usuais não conseguem modelar tal presença excessiva de zeros. Segue abaixo alguns exemplos.



**Área financeira:** Por convenção, as instituições bancárias dão um prazo de 60 dias para o cliente pagar a fatura do cartão. Se o cliente efetuar o pagamento mínimo até o prazo de vencimento de 60 dias, ele é considerado adimplente. Se o cliente efetuar o pagamento após o prazo de vencimento de 60 dias, ele é considerado inadimplente. Considerando o número de dias entre o vencimento e o pagamento da fatura, observamos um conjunto de dados de contagem com excesso de zeros observados, pois a grande maioria dos clientes efetua o pagamento dentro do prazo de vencimento.

**Estudos em ecologia:** Os zeros podem ocorrer devido ao fato, por exemplo, da espécie ser totalmente ausente na área de amostragem ou ainda quanto a espécie pertence ao local, porém não foi observada pelo pesquisador. Nesses casos os zeros são aleatórios (falsos). Nessas aplicações em ecologia, a presença de zeros aleatórios ocorre com frequência devido aos erros humanos de medição ou vícios no método de amostragem (Martin et al. 2005). Portanto, podemos concluir que, em dados de contagem, os zeros podem ser obtidos por erro humano, por amostragem ou ser resultado de um zero verdadeiro. Todavia, na grande maioria das situações práticas é impossível separar os zeros de amostragem do zero verdadeiro.

**Controle de qualidade:** Em uma linha de produção onde há uma contagem de defeitos por item produzido, existe uma grande quantidade de itens que não apresenta nenhum defeito, devido ao fato da intensa modernização dos processos de fabricação, capacitação de funcionários e manutenção dos equipamentos. Por esse motivo, os itens fabricados com ou sem defeitos, produzidos por essa linha de produção, configuram-se então num conjunto de dados com excesso de zeros. Nessa situação os zeros são determinísticos.

#### 4.5.4 O modelo estatístico inflacionado no ponto $s$

Na análise de dados discretos inflacionados é usual considerarmos misturas de distribuições. Murat e Szynal (1998) descreveram a classe de distribuições série de potências para dados inflacionados utilizando esta metodologia. O modelo é especificado como segue

$$P[R = r \mid \Theta = (\boldsymbol{\theta}, \omega)] = \begin{cases} \omega + (1 - \omega) \frac{a(s)[g(\boldsymbol{\theta})]^s}{f(\boldsymbol{\theta})}, & \text{se } r = s \\ (1 - \omega) \frac{a(r)[g(\boldsymbol{\theta})]^r}{f(\boldsymbol{\theta})}, & \text{se } r \neq s \end{cases} \quad (4.5.1)$$

em que  $0 \leq \omega \leq 1$ ,  $r \in \mathbb{N}$  representa o conjunto dos números naturais,  $f(\boldsymbol{\theta}) = \sum_r a(r) [g(\boldsymbol{\theta})]^r < \infty$ ,  $g(\boldsymbol{\theta})$  é uma função positiva, finita, diferenciável e inversível, e  $a(r)$  é independente de  $\boldsymbol{\theta}$ .

## 4.6 Modelo de regressão logística com resposta pertencente à família de distribuições série de potências inflacionadas

Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  com distribuição de probabilidade pertencente à família de distribuições série de potências inflacionadas dada em (4.5.1). Seja  $C_R$  uma constante arbitrária,  $C \in \mathbb{R}$ , e considere  $n$  variáveis aleatórias independentes binárias  $Y_1, Y_2, \dots, Y_n$ , tal que  $Y_i = 1$  se  $R_i > C_R$ , e  $Y_i = 0$  se  $R_i \leq C_R$ ,  $i = 1, 2, \dots, n$ .

Para incorporarmos a informação da distribuição da variável resposta no modelo de regressão logística, vamos considerar a relação

$$P(R_i > C_R) = P(Y_i > 0) = P(Y_i = 1) = \pi_i, \quad i = 1, 2, \dots, n. \quad (4.6.1)$$

Dessa maneira, a probabilidade da variável resposta  $R_i$  ser maior que uma constante arbitrária  $C_R$ ,  $C_R \in \mathbb{R}$ , ou seja

$$P(R_i > C_R) = 1 - P(R_i \leq C_R) = 1 - F_{R_i}(C_R), \quad (4.6.2)$$

é igual a probabilidade  $P(R_i > C_R)$  seja igual a probabilidade de sucesso  $\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$  da variável resposta binária  $Y_i$ , isto é

$$g[(1 - F_{R_i}^{C_R})] = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, \quad i = 1, 2, \dots, n, \quad (4.6.3)$$

e portanto, a partir de (4.5.1) e assumindo (4.6.3), segue que, para o ponto inflacionado  $s \leq C_R$ ,

$$\mathbf{x}_i^T \boldsymbol{\beta} = g \left\{ 1 - \omega - (1 - \omega) \frac{a_i(s) [g(\theta_i)]^s}{f(\theta_i)} - (1 - \omega) \sum_{r_i \neq s} \frac{a_i(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)} \right\}, \quad (4.6.4)$$

e para o ponto inflacionado  $s > C_R$  temos

$$\mathbf{x}_i^T \boldsymbol{\beta} = g \left\{ 1 - (1 - \omega) \sum_{r_i \neq s} \frac{a_i(r_i) [g(\theta_i)]^{r_i}}{f(\theta_i)} \right\}. \quad (4.6.5)$$

## 4.7 Caso particular: o modelo geométrico para dados inflacionados de zeros

Consideremos  $n$  variáveis aleatórias  $R_1, R_2, \dots, R_n$  independentes seguindo uma distribuição geométrica inflacionada em zero, isto é,  $R_i \sim Geometrica(p_i, \omega)$ ,  $i = 1, 2, \dots, n$ , tal que sua distribuição de probabilidades seja

$$P(R_i = r_i) = \begin{cases} \omega + (1 - \omega) p_i, & \text{se } r_i = 0 \\ (1 - \omega) p_i (1 - p_i)^{r_i}, & \text{se } r_i \neq 0. \end{cases}$$

Seja  $A$  o conjunto dos valores de  $r_i$  iguais a zero, ou seja,  $A = \{r_i \mid r_i = 0\}$  e seja  $m$  o total de zeros, isto é,  $m = n(A)$ . Dessa forma a função de verossimilhança e log-verossimilhança são dadas por

$$\begin{aligned} L(\mathbf{p}, \omega \mid \mathbf{r}) &= \prod_{r_i \in A} [\omega + (1 - \omega) p_i] (1 - \omega)^{n-m} \prod_{r_i \notin A} [p_i (1 - p_i)^{r_i}] \\ l(\mathbf{p}, \omega \mid \mathbf{r}) &= \sum_{r_i \in A} \ln[\omega + (1 - \omega) p_i] + (n - m) \ln(1 - \omega) + \sum_{r_i \notin A} \ln(p_i) + \sum_{r_i \notin A} r_i \ln(1 - p_i), \end{aligned}$$

em que  $\mathbf{r} = (r_1, r_2, \dots, r_n)$ .

## 4.8 Modelo de regressão logística com resposta geométrica inflacionado de zeros

Nesta seção apresentamos o modelo de regressão logística com resposta geométrica zero inflacionado. Considere  $n$  variáveis aleatórias  $R_1, R_2, \dots, R_n$  independentes seguindo uma distribuição geométrica inflacionada em zero, isto é,  $R_i \sim Geometrica(p_i, \omega)$ ,  $i = 1, 2, \dots, n$ , tal que sua distribuição de probabilidades seja da seguinte forma

$$P(R_i = r_i) = \begin{cases} \omega + (1 - \omega) p_i, & \text{se } r_i = 0 \\ (1 - \omega) p_i (1 - p_i)^{r_i}, & \text{se } r_i \neq 0. \end{cases}$$

**Proposição 6** Considere  $n$  variáveis aleatórias independentes  $R_1, R_2, \dots, R_n$  com distribuição geométrica zero inflacionada, isto é,  $R_i \sim Geometrica(p_i, \omega)$ ,  $i = 1, 2, \dots, n$ , e considere outro conjunto de  $n$  variáveis aleatórias independentes  $Y_1, Y_2, \dots, Y_n$  com distribuição de probabilidade de Bernoulli, isto é,  $Y_i \sim Bernoulli(\pi_i)$ ,  $i = 1, 2, \dots, n$ . Segundo a relação dada em (4.6.1), incorporando a informação da variável resposta geométrica zero inflacionada no ajuste de  $Y_i$  temos que a distribuição de probabilidades de  $R_i$  é tal que

$$R_i \sim Geometrica \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}}, \omega \right\}, i = 1, 2, \dots, n.$$

**Prova.** Pela relação expressa em (4.6.1) temos

$$g[(1 - F_{r_i}^{C_R})] = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, i = 1, 2, \dots, n,$$

e portanto

$$\begin{aligned} \mathbf{x}_i^T \boldsymbol{\beta} &= g[1 - P(R_i \leq C_R)] \\ &= g\{1 - [P(R_i = 0) + P(R_i = 1) + \dots + P(R_i = C_R)]\} \\ &= g\left\{1 - \left[\omega + (1 - \omega) p_i + (1 - \omega) p_i (1 - p_i)^1 + \dots + (1 - \omega) p_i (1 - p_i)^{C_R}\right]\right\} \\ &= g\left\{1 - \omega - (1 - \omega) p_i \left[1 + (1 - p_i)^1 + (1 - p_i)^2 + \dots + (1 - p_i)^{C_R}\right]\right\} \\ &= g\left\{1 - \omega - (1 - \omega) p_i \left[\frac{(1 - p_i)^{C_R + 1} - 1}{(1 - p_i) - 1}\right]\right\} \\ &= g\left\{1 - \omega - (1 - \omega) \left[(1 - p_i)^{C_R + 1} - 1\right]\right\}, \end{aligned}$$

e finalmente

$$\mathbf{x}_i^T \boldsymbol{\beta} = g\left[(1 - \omega) (1 - p_i)^{C_R + 1}\right], i = 1, 2, \dots, n.$$

Dessa forma segue que

$$p_i = 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}}, i = 1, 2, \dots, n.$$

■

Como  $0 \leq p_i \leq 1$  temos a seguinte restrição

$$0 < g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) < 1 - \omega < 1.$$

O modelo de regressão logística com resposta geométrica zero inflacionado fica assim definido por

$$P(R_i = r_i) = \begin{cases} 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}}, & \text{se } r_i = 0 \\ (1 - \omega) \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}} \right\} \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{r_i}{C_R + 1}}, & \text{se } r_i \neq 0, \end{cases} \quad (4.8.1)$$

Seja  $A$  o conjunto dos valores de  $r_i$  iguais a zero, ou seja,  $A = \{r_i \mid r_i = 0\}$  e seja  $m$  o total de zeros, isto é,  $m = n(A)$ . Dessa forma a função de verossimilhança é tal que

$$L(\boldsymbol{\beta}, \omega \mid \mathbf{r}) = \prod_{r_i \in A} \left\{ 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}} \right\} \\ \times (1 - \omega)^{n-m} \prod_{r_i \notin A} \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}} \right\} \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{r_i}{C_R + 1}},$$

e, por sua vez, o logaritmo da função de verossimilhança é

$$l(\boldsymbol{\beta}, \omega \mid \mathbf{r}) = \sum_{r_i \in A} \ln \left\{ 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}} \right\} + (n - m) \ln(1 - \omega) \\ + \sum_{r_i \notin A} \ln \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_R + 1}} \right\} + \sum_{r_i \notin A} \left\{ \left( \frac{r_i}{C_R + 1} \right) \ln \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right] \right\}.$$

Os escores são tais que

$$\frac{\partial l(\boldsymbol{\beta}, \omega \mid \mathbf{r})}{\partial \omega} = 0 \quad \text{e} \quad \frac{\partial l(\boldsymbol{\beta}, \omega \mid \mathbf{r})}{\partial \beta_j} = 0, \quad j = 0, 1, \dots, p.$$

Os estimadores de máxima verossimilhança podem ser obtidos por meio do algoritmo BFGS.

## 5 Abordagem bayesiana

### 5.1 Introdução

Neste capítulo fazemos a abordagem bayesiana para os modelos de regressão segundo a metodologia proposta. A eficiência desta abordagem para a estimação de parâmetros populacionais é evidenciada, principalmente, quando o pesquisador dispõe de informações *a priori* a respeito dos parâmetros. Apresentamos as distribuições *a posteriori* bem como as distribuições condicionais para os modelos de regressão com resposta normal, com resposta exponencial e também para o modelo logístico com respostas diversas: geométrica, Poisson, normal considerando a estrutura heteroscedástica multiplicativa e geométrica inflacionada em zero. No estudo de simulação, atribuímos distribuições *a priori* vagas para os parâmetros de regressão e para os parâmetros perturbadores com o objetivo de comparar as estimativas bayesianas com as estimativas clássicas.

### 5.2 Função de verossimilhança para os modelos de regressão com resposta normal

Vimos na seção 2.3.6 que a função de verossimilhança para os modelos desenvolvidos com resposta normal, segundo a proposição (1), em que  $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$ , é da seguinte forma:

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{r}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [r_i - \sigma \Phi^{-1}(1 - F_{Y_i}(C_Y)) - C_R]^2 \right\}, \quad (5.2.1)$$

que, para facilitar a obtenção das distribuições condicionais no caso da abordagem bayesiana, (5.2.1) pode ser reescrita como

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{r}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (r_i - C_R)^2 + \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) - \frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\}. \quad (5.2.2)$$

Cada um dos possíveis modelos de regressão assume uma expressão para o termo  $\psi_i$  em 5.2.2, conforme a Tabela 5.2.

Tabela 5.2. Expressões do termo  $\psi_i$  para os modelos de regressão.

Modelo	$\psi_i$
Logístico	$\Phi^{-1}(\pi_i) = \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ , $i = 1, 2, \dots, n$ .
Exponencial	$\Phi^{-1}[\exp(-\lambda_i C_Y)]$ , $i = 1, 2, \dots, n$ , em que $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Geométrico	$\Phi^{-1}[(1 - p_i)^{C_Y + 1}]$ , $i = 1, 2, \dots, n$ , em que $p_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Poisson	$\Phi^{-1} \left\{ 1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \right\}$ , $i = 1, 2, \dots, n$ , em que $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$ .
Log-normal	$\Phi_R^{-1} \left[ \Phi_Y \left( \frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y} \right) \right]$ , $i = 1, 2, \dots, n$ , em que $\mu_{Y_i} = \mathbf{x}_i^T \boldsymbol{\beta}$ .

Consideramos o vetor de parâmetros  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a construção das distribuições *a posteriori* conforme a seguir.

### 5.3 Modelo de regressão logística bayesiano com a informação da variável resposta normal

Nesta seção estudamos o modelo de regressão logística com resposta normal sob o enfoque bayesiano. Considerando o vetor de parâmetros  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  e o parâmetro perturbador  $\sigma^2$  independentes, adotaremos as seguintes distribuições *a priori*:

$$\beta_j \sim N(a_j, b_j^2), \text{ isto é, } \pi(\beta_j) \propto \exp\left\{-\frac{1}{2b_j^2}(\beta_j - a_j)^2\right\}, \quad j = 0, 1, 2, 3. \quad (5.3.1)$$

$$\sigma^2 \sim IG(c, d), \text{ isto é, } \pi(\sigma^2) \propto (\sigma^2)^{-(c+1)} \exp\left(-\frac{d}{\sigma^2}\right), \quad (5.3.2)$$

em que *IG* denota a distribuição gama invertida e  $a_j, b_j, j = 0, 1, 2, 3, c$  e  $d$  são hiperparâmetros conhecidos. Dessa forma, como a distribuição *a posteriori* é proporcional ao produto da função de verossimilhança pela distribuição *a priori* dos parâmetros, temos que a partir de (5.2.2), (5.3.1) e (5.3.2) obtemos a seguinte distribuição *a posteriori* de  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  e  $\sigma^2$ :

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{r}) &\propto (\sigma^2)^{-\left(\frac{n}{2}+c+1\right)} & (5.3.3) \\ &\times \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (r_i - C_R)^2 + \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) - \frac{1}{2} \sum_{i=1}^n \psi_i^2\right\} \\ &\times \exp\left(-\frac{d}{\sigma^2}\right) \times \exp\left\{-\frac{1}{2} \sum_{j=0}^3 \left(\frac{\beta_j - a_j}{b_j}\right)^2\right\}. \end{aligned}$$

A partir da distribuição *a posteriori* dada em (5.3.3) temos as seguintes distribuições condicionais completas:

**i.** A distribuição condicional completa de  $\sigma^2$  dados o vetor  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  e o vetor de dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é

$$\pi(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{r}) \propto (\sigma^2)^{-\left(\frac{n}{2}+c+1\right)} \exp\left\{-\frac{1}{\sigma^2} \left[\frac{1}{2} \sum_{i=1}^n (r_i - C_R)^2 + d\right]\right\} \exp\left\{\frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R)\right\}, \quad (5.3.4)$$

em que o termo

$$(\sigma^2)^{-\left(\frac{n}{2}+c+1\right)} \exp\left\{-\frac{1}{\sigma^2} \left[\frac{1}{2} \sum_{i=1}^n (r_i - C_R)^2 + d\right]\right\}$$

é o núcleo de uma distribuição gama inversa  $IG\left\{\frac{n}{2} + c; \left[\frac{1}{2} \sum_{i=1}^n (r_i - C_R)^2 + d\right]\right\}$ .

**ii.** A distribuição condicional completa de  $\beta_j$  dados o parâmetro  $\sigma^2$ , o vetor  $\boldsymbol{\beta}_{(-j)}$  e o vetor de dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é

$$\pi(\beta_j \mid \boldsymbol{\beta}_{(-j)}, \sigma^2, \mathbf{r}) \propto \exp\left\{\frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) - \frac{1}{2} \sum_{i=1}^n \psi_i^2\right\} \exp\left\{-\frac{1}{2b_j^2}(\beta_j - a_j)^2\right\}, \quad (5.3.5)$$

para  $j = 0, 1, 2, 3$  e  $\psi_i = \Phi^{-1}\left[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})\right] = \Phi^{-1}\left[\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}\right], i = 1, 2, \dots, n.$

**Observação:** A distribuição *a posteriori* bem como as distribuições condicionais para os demais modelos de regressão com resposta normal é como em (5.3.3), (5.3.4) e (5.3.5) respectivamente, bastando substituir o termo  $\psi_i$  como apresentado na Tabela 5.2.

## 5.4 Função de verossimilhança para os modelos de regressão com resposta exponencial

Vimos na seção 2.4.6 do capítulo 2 que a função de verossimilhança para os modelos desenvolvidos com resposta exponencial, segundo a proposição (2), em que  $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$ , é da seguinte forma

$$L(\boldsymbol{\beta} | \mathbf{r}) = \prod_{i=1}^n \left\{ -\frac{\ln[\psi_i] \times [\psi_i]^{r_i/C_R}}{C_R} \right\}. \quad (5.4.1)$$

Cada um dos possíveis modelos de regressão com resposta exponencial assume uma expressão para o termo  $\psi_i$  em (5.4.1), conforme a Tabela 5.4.

Tabela 5.4. Expressões do termo  $\psi_i$  para os modelos de regressão.

Modelo	$\psi_i$
Logístico	$\pi_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), i = 1, 2, \dots, n.$
Normal	$F_Y\left(\frac{\mu_i - C_Y}{\sigma}\right), i = 1, 2, \dots, n,$ em que $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}.$
Geométrico	$(1 - p_i)^{C_Y + 1}, i = 1, 2, \dots, n,$ em que $p_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$
Poisson	$1 - \sum_{y_i=0}^{C_Y} \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}, i = 1, 2, \dots, n,$ em que $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}).$
Log-normal	$\Phi_R^{-1}\left[\Phi_Y\left(\frac{\mu_{Y_i} - \log(C_Y)}{\sigma_Y}\right)\right], i = 1, 2, \dots, n,$ em que $\mu_{Y_i} = \mathbf{x}_i^T \boldsymbol{\beta}.$

## 5.5 Modelo de regressão logística bayesiano com a informação da variável resposta exponencial

Nessa Seção vamos construir o modelo de regressão logística com resposta exponencial sob o enfoque bayesiano. Considerando os parâmetros  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  independentes, adotaremos as seguintes distribuições *a priori*

$$\beta_j \sim N(a_j, b_j^2), \text{ isto é, } \pi(\beta_j) \propto \exp\left\{-\frac{1}{2b_j^2}(\beta_j - a_j)^2\right\}, j = 0, 1, 2, 3. \quad (5.5.1)$$

Dessa forma, a partir da função de verossimilhança dada em 5.4.1 e das distribuições *a priori* dadas em 5.5.1, obtemos a seguinte distribuição *a posteriori* de  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ :

$$\pi(\boldsymbol{\beta} | \mathbf{r}) \propto \prod_{i=1}^n \left\{ -\frac{\ln[\psi_i] \times [\psi_i]^{r_i/C_R}}{C_R} \right\} \exp\left\{-\frac{1}{2} \sum_{j=0}^3 \left(\frac{\beta_j - a_j}{b_j}\right)^2\right\} \quad (5.5.2)$$

Da distribuição *a posteriori* conjunta (5.5.2) temos as seguintes distribuições condicionais completas:

**i.** A distribuição condicional de  $\beta_j$  dados o vetor de dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é dada por

$$\pi(\beta_j | \boldsymbol{\beta}_{(-j)}, \mathbf{r}) \propto \prod_{i=1}^n \left\{ -\frac{\ln[\psi_i] \times [\psi_i]^{r_i/C_R}}{C_R} \right\} \exp \left\{ -\frac{1}{2} \left( \frac{\beta_j - a_j}{b_j} \right)^2 \right\} \quad (5.5.3)$$

para  $j = 0, 1, 2, 3$ .

**Observação:** A distribuição *a posteriori* bem como as distribuições condicionais para os demais modelos de regressão com resposta exponencial é como em 5.5.2 e 5.5.3 respectivamente, bastando substituir o termo  $\psi_i$  como apresentado na Tabela 5.4.

## 5.6 Modelo logístico bayesiano com a informação da variável resposta normal heteroscedástica multiplicativa

Nesta seção estudamos o modelo logístico bayesiano com informação da variável resposta normal considerando um contexto de heteroscedasticidade multiplicativa. Conforme detalhado no Capítulo 2, seção 2.5.5, a função de verossimilhança é tal que

$$L(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \prod_{i=1}^n x_i^{-\lambda/2} \quad (5.6.1)$$

$$\times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{[r_i - \sigma x_i^{\lambda/2} \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] - C_R]^2}{x_i^\lambda} \right\}$$

Para facilitar a obtenção das distribuições condicionais, a função de verossimilhança dada em (5.6.1) pode ser reescrita como

$$L(\boldsymbol{\beta}, \sigma^2, \lambda | \mathbf{r}) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \prod_{i=1}^n x_i^{-\lambda/2} \quad (5.6.2)$$

$$\times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} - \frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\}$$

em que  $\psi_i = \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] = \Phi^{-1} \left[ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]$ , para  $i = 1, 2, \dots, n$ .

Considerando o vetor de parâmetros  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ , e os parâmetros perturbadores  $\sigma^2$  e  $\lambda$  independentes, adotaremos as seguintes distribuições *a priori*:

$$\lambda \sim \text{Gama}(c, d), \text{ isto é, } \pi(\lambda) \propto \lambda^{c-1} \exp(-d\lambda); \quad (5.6.3)$$

$$\sigma^2 \sim \text{IG}(e, f), \text{ isto é, } \pi(\sigma^2) \propto (\sigma^2)^{-(e+1)} \exp \left\{ -\frac{f}{\sigma^2} \right\}; \quad (5.6.4)$$

$$\beta_j \sim N(a_j, b_j^2), \text{ isto é, } \pi(\beta_j) \propto \exp \left\{ -\frac{1}{2b_j^2} (\beta_j - a_j)^2 \right\}, j = 0, 1, 2, 3. \quad (5.6.5)$$

Dessa maneira, a partir da função de verossimilhança dada em (5.6.2) e das distribuições *a priori* (5.6.3), (5.6.4) e (5.6.5), obtemos a seguinte distribuição *a posteriori* de



$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ ,  $\sigma^2$  e  $\lambda$ :

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2, \lambda \mid \mathbf{r}) &\propto (\sigma^2)^{-\left(\frac{n}{2}+e+1\right)} \left( \prod_{i=1}^n x_i^{-\lambda/2} \right) \\ &\times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i(r_i - C_R)}{x_i^{\lambda/2}} - \frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} \\ &\times \lambda^{c-1} \times \exp \left\{ -\frac{1}{2} \sum_{j=0}^3 \left( \frac{\beta_j - a_j}{b_j} \right)^2 - d\lambda - \frac{f}{\sigma^2} \right\} \end{aligned} \quad (5.6.6)$$

em que  $\psi_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] = \Phi^{-1} \left[ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]$ , para  $i = 1, 2, \dots, n$ .

A partir da distribuição *a posteriori* conjunta (5.6.6) temos as seguintes distribuições condicionais

i. A distribuição condicional de  $\sigma^2$  dados o parâmetro  $\lambda$ , o vetor de coeficientes  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  e o vetor de dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$ :

$$\begin{aligned} \pi(\sigma^2 \mid \boldsymbol{\beta}, \lambda, \mathbf{r}) &\propto (\sigma^2)^{-\left(\frac{n}{2}+e+1\right)} \exp \left\{ -\frac{1}{\sigma^2} \left[ \frac{1}{2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + f \right] \right\} \\ &\times \exp \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i(r_i - C_R)}{x_i^{\lambda/2}} \right\}, \end{aligned}$$

em que o termo

$$(\sigma^2)^{-\left(\frac{n}{2}+e+1\right)} \exp \left\{ -\frac{1}{\sigma^2} \left[ \frac{1}{2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + f \right] \right\}$$

denota o núcleo de uma distribuição gama inversa  $IG \left\{ \frac{n}{2} + e; \left[ \frac{1}{2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + f \right] \right\}$ .

ii. A distribuição condicional de  $\lambda$  dados o parâmetro  $\sigma^2$ , o vetor de coeficientes  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  e o vetor de dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$

$$\begin{aligned} \pi(\lambda \mid \boldsymbol{\beta}, \sigma^2, \mathbf{r}) &\propto \lambda^{c-1} \exp(-d\lambda) \times \left( \prod_{i=1}^n x_i^{-\lambda/2} \right) \\ &\times \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} + \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i(r_i - C_R)}{x_i^{\lambda/2}} \right\}. \end{aligned}$$

iii. A distribuição condicional do coeficiente de regressão  $\beta_j$ ,  $j = 0, 1, 2, 3$ , dados os parâmetros  $\lambda$ ,  $\sigma^2$ , o vetor de coeficientes  $\boldsymbol{\beta}_{(-j)}$  e o vetor de dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$

$$\pi(\beta_j \mid \boldsymbol{\beta}_{(-j)}, \sigma^2, \lambda, \mathbf{r}) \propto \exp \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i(r_i - C_R)}{x_i^{\lambda/2}} - \frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} \exp \left\{ -\frac{1}{2b_j} (\beta_j - a_j)^2 \right\},$$

para  $j = 0, 1, 2, 3$  e  $\psi_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] = \Phi^{-1} \left[ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]$ ,  $i = 1, 2, \dots, n$ .

## 5.7 Modelo de regressão logística bayesiano com a informação da variável resposta geométrica

Conforme visto anteriormente no capítulo 2, seção 2.5.1, a função de verossimilhança para o modelo logístico com informação da variável resposta geométrica é tal que

$$\begin{aligned} L(\boldsymbol{\beta} \mid \mathbf{r}) &= \prod_{i=1}^n \left\{ 1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \right\} \left\{ [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{r_i}{C_R+1}} \right\} \\ &= \prod_{i=1}^n \left\{ \frac{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{r_i}{C_R+1}}} \right\} \times \exp \left\{ \frac{1}{(C_R + 1)} \sum_{i=1}^n r_i (\mathbf{x}_i^T \boldsymbol{\beta}) \right\}. \end{aligned} \quad (5.7.1)$$

Considerando os coeficientes de regressão  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  independentes atribuímos as seguintes distribuições *a priori*

$$\beta_j \sim N(a_j, b_j^2), \text{ isto é, } \pi(\beta_j) \propto \exp \left\{ -\frac{1}{2b_j^2} (\beta_j - a_j)^2 \right\}, \quad j = 0, 1, 2, 3. \quad (5.7.2)$$

Logo, a partir da função de verossimilhança dada em (5.7.1) e das distribuições *a priori* dadas em (5.7.2) temos a seguinte distribuição *a posteriori* conjunta para os coeficientes  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ :

$$\begin{aligned} \pi(\boldsymbol{\beta} \mid \mathbf{r}) &\propto \prod_{i=1}^n \left\{ \frac{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{r_i}{C_R+1}}} \right\} \\ &\times \exp \left\{ \frac{1}{(C_R + 1)} \sum_{i=1}^n r_i (\mathbf{x}_i^T \boldsymbol{\beta}) - \frac{1}{2} \sum_{j=0}^3 \left( \frac{\beta_j - a_j}{b_j} \right)^2 \right\} \end{aligned} \quad (5.7.3)$$

Da distribuição *a posteriori* conjunta (5.7.3) temos as seguintes distribuições condicionais completas:

i. A distribuição condicional de  $\beta_0$  dados  $\boldsymbol{\beta} = \beta_1, \beta_2, \beta_3$  e os dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é dada por

$$\pi(\beta_0 \mid \boldsymbol{\beta}_{(-0)}, \mathbf{r}) \propto \prod_{i=1}^n \left\{ \frac{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{r_i}{C_R+1}}} \right\} \exp \left[ \frac{\beta_0}{C_R + 1} \sum_{i=1}^n r_i - \frac{1}{2} \left( \frac{\beta_0 - a_0}{b_0} \right)^2 \right]$$

ii. A distribuição condicional de  $\beta_j$  dados  $\boldsymbol{\beta}_{(-j)}$  e os dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é dada por

$$\pi(\beta_j \mid \beta_0, \boldsymbol{\beta}_{(-j)}, \mathbf{r}) \propto \prod_{i=1}^n \left\{ \frac{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{r_i}{C_R+1}}} \right\} \exp \left[ \frac{\beta_j}{C_R + 1} \sum_{i=1}^n x_{ij} r_i - \frac{1}{2} \left( \frac{\beta_j - a_j}{b_j} \right)^2 \right],$$

para  $j = 1, 2, 3$ .

## 5.8 Modelo de regressão logística bayesiano com informação da variável resposta Poisson

Nessa Seção vamos construir o modelo logístico bayesiano com informação da variável resposta Poisson. Conforme visto anteriormente na seção 2.5.3 do capítulo 2, a função de verossimilhança para o modelo logístico com informação da variável resposta Poisson, considerando  $C_R = 0$ , é tal que

$$L(\boldsymbol{\beta}, \mathbf{r}) = \prod_{i=1}^n \frac{1}{r_i!} \left\{ -\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}^{r_i} [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})], \quad r_i = 0, 1, 2, \dots,$$

que pode ser reescrita como

$$L(\boldsymbol{\beta}, \mathbf{r}) = \prod_{i=1}^n \frac{\left\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}^{r_i}}{r_i! [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]}, \quad r_i = 0, 1, 2, \dots \quad (5.8.1)$$

Considerando os coeficientes  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  independentes atribuímos as seguintes distribuições *a priori*

$$\beta_j \sim N(a_j, b_j^2), \quad \text{isto é, } \pi(\beta_j) \propto \exp \left\{ -\frac{1}{2b_j^2} (\beta_j - a_j)^2 \right\}, \quad j = 0, 1, 2, 3. \quad (5.8.2)$$

Logo, a partir da função de verossimilhança dada em (5.8.1) e das distribuições *a priori* dadas em (5.8.2) temos a seguinte distribuição *a posteriori* conjunta para os coeficientes  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$

$$\pi(\boldsymbol{\beta} | \mathbf{r}) \propto \prod_{i=1}^n \frac{\left\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}^{r_i}}{r_i! [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} \exp \left\{ -\frac{1}{2} \sum_{j=0}^3 \left( \frac{\beta_j - a_j}{b_j} \right)^2 \right\}, \quad (5.8.3)$$

e, portanto, a partir de (5.8.3) obtemos as seguintes distribuições condicionais

$$\pi(\beta_j | \boldsymbol{\beta}_{(-j)}, \mathbf{r}) \propto \prod_{i=1}^n \frac{\left\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \right\}^{r_i}}{r_i! [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} \exp \left\{ -\frac{1}{2} \left( \frac{\beta_j - a_j}{b_j} \right)^2 \right\},$$

para  $j = 0, 1, 2, 3$ .

## 5.9 Modelo logístico bayesiano com resposta geométrica inflacionada em zero

Nesta seção apresentamos o modelo de regressão logística bayesiano com resposta geométrica inflacionada em zero. Seja  $A$  o conjunto dos valores de  $r_i$  iguais a zero, ou seja,  $A = \{r_i | r_i = 0\}$  e seja  $m$  o total de zeros, isto é,  $m = n(A)$ . Vimos no capítulo 4, seção 4.8 que a função de verossimilhança é tal que

$$\begin{aligned} L(\boldsymbol{\beta}, \omega | \mathbf{r}) &= \prod_{r_i \in A} \left\{ 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_{R+1}}} \right\} \\ &\times (1 - \omega)^{n-m} \prod_{r_i \notin A} \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{C_{R+1}}} \right\} \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{r_i}{C_{R+1}}}, \end{aligned} \quad (5.9.1)$$

Considerando os coeficientes de regressão e o parâmetro de inflação independentes atribuímos as seguintes distribuições *a priori*

$$\beta_j \sim N(a_j, b_j^2), \text{ isto é, } \pi(\beta_j) \propto \exp\left\{-\frac{1}{2b_j^2}(\beta_j - a_j)^2\right\}, j = 0, 1, \dots, p. \quad (5.9.2)$$

$$\omega \sim \text{Beta}(c, d), \text{ isto é, } \pi(\omega) \propto \omega^{c-1}(1-\omega)^{d-1}. \quad (5.9.3)$$

Desta forma, a partir da função de verossimilhança dada em (5.9.1) e das distribuições *a priori* dadas em (5.9.2) e em (5.9.3) temos a seguinte distribuição *a posteriori* para os coeficientes de regressão  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$  e  $\omega$ :

$$\begin{aligned} \pi(\boldsymbol{\beta}, \omega \mid \mathbf{r}) &\propto \prod_{r_i \in A} \left\{ 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{c_{R+1}}} \right\} \\ &\times (1 - \omega)^{n-m} \prod_{r_i \notin A} \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{c_{R+1}}} \right\} \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{r_i}{c_{R+1}}} \\ &\times \exp\left\{-\frac{1}{2} \sum_{j=0}^p \left( \frac{\beta_j - a_j}{b_j} \right)^2\right\} \times \omega^{c-1} (1 - \omega)^{d-1}. \end{aligned} \quad (5.9.4)$$

A partir da distribuição *a posteriori* expressa em (5.9.5) temos as seguintes distribuições condicionais completas:

i. A distribuição condicional de  $\omega$  dados  $\boldsymbol{\beta}$  e os dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é dada por

$$\begin{aligned} \pi(\omega \mid \boldsymbol{\beta}, \mathbf{r}) &\propto \prod_{r_i \in A} \left\{ 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{c_{R+1}}} \right\} \times \omega^{c-1} (1 - \omega)^{d-1} \\ &\times (1 - \omega)^{n-m} \prod_{r_i \notin A} \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{c_{R+1}}} \right\} \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{r_i}{c_{R+1}}}. \end{aligned} \quad (5.9.5)$$

ii. A distribuição condicional de  $\beta_j$  dados  $\boldsymbol{\beta}_{(-j)}$ ,  $\omega$  e os dados  $\mathbf{r} = (r_1, r_2, \dots, r_n)$  é dada por

$$\begin{aligned} \pi(\beta_j \mid \boldsymbol{\beta}_{(-j)}, \omega, \mathbf{r}) &\propto \prod_{r_i \in A} \left\{ 1 - (1 - \omega) \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{c_{R+1}}} \right\} \times \exp\left\{-\frac{1}{2} \left( \frac{\beta_j - a_j}{b_j} \right)^2\right\} \\ &\times (1 - \omega)^{n-m} \prod_{r_i \notin A} \left\{ 1 - \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{1}{c_{R+1}}} \right\} \left[ \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{1 - \omega} \right]^{\frac{r_i}{c_{R+1}}} \end{aligned} \quad (5.9.6)$$

para  $j = 0, 1, \dots, p$ .

## 5.10 Simulação para os modelos propostos

Nesta seção apresentamos as estimativas bayesianas produzidas pelo modelo de regressão logística bayesiano com resposta de origem e comparamos com estimativas clássicas. A programação computacional foi desenvolvida usando o software **R**, versão 2.10.1. Usamos os mesmos tamanhos amostrais do capítulo 3 ( $n = 50, 100, 200$  e  $500$ ). Geramos 100 amostras de tamanho  $n$  e, para cada amostra geramos duas cadeias de tamanho 120000 com um descarte de 20000 e um salto de tamanho 50 fornecendo uma amostra final da distribuição *a posteriori* conjunta de tamanho 2000. Para cada uma das cadeias foram adotados valores iniciais e sementes diferentes. A convergência das cadeias geradas foi diagnosticada pelo critério de Gelman-Rubin (1992). Tal diagnóstico de convergência foi monitorada utilizando o pacote *CODA* (Best *et. al* 1995).

## 5.11 Resultados obtidos para o modelo logístico bayesiano com resposta normal

Para o caso do modelo com resposta normal, geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(20; 16); R_i \sim N(\mu_i, \sigma^2),$$

com  $\sigma = 1.000$  e  $\mu_i = \sigma \Phi^{-1} [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + C_R, i = 1, 2, \dots, n$ .

Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\mu_i, i = 1, 2, \dots, n$ , foram  $\beta_0 = -4,5, \beta_1 = 0,5, \beta_2 = 0,35$  e  $\beta_3 = 0,02$ .

Consideramos o ponto de corte  $C_R = 11.000$ , e as seguintes distribuições *a priori* vagas:

$$\sigma^2 \sim IG(2, 01; 100) \text{ e } \beta_j \sim N(0, 1000), j = 0, 1, 2, 3.$$

A Tabela 5.11 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta normal, para cada tamanho amostral.

Tabela 5.11. Estimativas bayesianas e clássicas obtidas segundo ambos os modelos.

Parâmetros	Modelo bayesiano resposta normal ( $n = 50$ )				Modelo logístico usual ( $n = 50$ )		
Valores reais	Estimativas	Int.Cred (95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-5, 5457	-10, 2996	-0, 7972	-5, 5469	-16, 9596	3, 8686
$\beta_1$	0, 50	0, 6536	0, 1073	1, 1989	0, 6531	-0, 4043	1, 9958
$\beta_2$	0, 35	0, 4568	0, 1740	0, 7390	0, 4565	-0, 0634	1, 1976
$\beta_3$	0, 02	0, 0077	-0, 1005	0, 1162	0, 0077	-0, 2504	0, 2273
$\sigma$	1000	1056	935	1229	—	—	—

Parâmetros	Modelo bayesiano resposta normal ( $n = 100$ )				Modelo logístico usual ( $n = 100$ )		
Valores reais	Estimativas	Int.Cred (95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-4, 5779	-7, 0349	-2, 1216	-4, 5812	-10, 0194	0, 4851
$\beta_1$	0, 50	0, 5453	0, 2394	0, 8514	0, 5456	-0, 0617	1, 2397
$\beta_2$	0, 35	0, 3693	0, 2083	0, 5302	0, 3692	0, 0457	0, 7329
$\beta_3$	0, 02	0, 0091	-0, 0561	0, 0743	0, 0091	-0, 1342	0, 1441
$\sigma$	1000	1214	1107	1372	—	—	—

Parâmetros	Modelo bayesiano resposta normal ( $n = 200$ )				Modelo logístico usual ( $n = 200$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-4, 3312	-6, 0732	-2, 5898	-4, 3316	-8, 1076	-0, 7201
$\beta_1$	0, 50	0, 4940	0, 3015	0, 6868	0, 4944	0, 0999	0, 9153
$\beta_2$	0, 35	0, 3506	0, 2388	0, 4622	0, 3504	0, 1221	0, 5957
$\beta_3$	0, 02	0, 0141	-0, 0338	0, 0622	0, 0142	-0, 0886	0, 1149
$\sigma$	1000	1308	1226	1433	—	—	—

Parâmetros	Modelo bayesiano resposta normal ( $n = 500$ )				Modelo logístico usual ( $n = 500$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-4, 6018	-5, 6965	-3, 5021	-4, 5991	-6, 9557	-2, 3200
$\beta_1$	0, 50	0, 4952	0, 3744	0, 6161	0, 4952	0, 2453	0, 7552
$\beta_2$	0, 35	0, 3595	0, 2928	0, 4262	0, 3596	0, 2221	0, 5033
$\beta_3$	0, 02	0, 0220	-0, 0087	0, 0527	0, 0220	-0, 0429	0, 0867
$\sigma$	1000	1390	1344	1452	—	—	—

O termo **estimativas** nas Tabelas desta seção se refere a média das 100 esperanças *a posteriori*, e o termo *Int.Cred (95%)* se refere a média dos 100 intervalos de credibilidade obtidos.

A partir dos resultados expostos na Tabela 5.11, podemos observar que as estimativas bayesianas pontuais estão próximas das estimativas clássicas. A precisão das estimativas intervalares do modelo bayesiano com resposta normal também é evidenciada por meio de intervalos de credibilidade com amplitudes menores que as dos intervalos de confiança do modelo logístico usual.

À medida que o tamanho da amostra aumenta, o modelo logístico bayesiano com resposta normal continua apresentando estimativas intervalares mais precisas.

Por outro lado, as estimativas bayesianas para o parâmetro perturbador  $\sigma$  foram sobrestimadas a medida que o tamanho da amostra aumenta.

## 5.12 Resultados obtidos para o modelo logístico bayesiano com resposta log-normal

Para o modelo bayesiano de regressão logística com resposta log-normal, geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(25; 25); R_i \sim LN(\mu_i, \sigma^2),$$

em que  $LN$  denota a distribuição log-normal, e

$$\mu_i = \sigma \Phi^{-1} [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + \log(C_R), i = 1, 2, \dots, n,$$

com  $\sigma^2 = 0,50 \Rightarrow \sigma = 0,71$ . Os valores atribuídos para os coeficientes de regressão foram  $\beta_0 = -12,0$ ,  $\beta_1 = 1,0$ ,  $\beta_2 = 0,44$  e  $\beta_3 = 0,05$ . Consideramos o ponto de corte  $C_R = 10$  e as seguintes distribuições *a priori* vaga

$$\sigma^2 \sim IG(2, 01; 100) \quad \text{e} \quad \beta_j \sim N(0, 1000), j = 0, 1, 2, 3.$$

A Tabela 5.12 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta log-normal, para cada tamanho amostral.

Tabela 5.12. Estimativas bayesianas e clássicas obtidas segundo ambos os modelos.

Parâmetros		Modelo bayesiano resposta lognormal ( $n = 50$ )			Modelo logístico usual ( $n = 50$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-12, 0	-5, 2482	-9, 7392	-0, 7571	-5, 2452	-15, 9728	3, 6902
$\beta_1$	0, 50	0, 6053	0, 0917	1, 1173	0, 6046	-0, 3998	1, 8516
$\beta_2$	0, 35	0, 4339	0, 1658	0, 7018	0, 4338	-0, 0599	1, 1327
$\beta_3$	0, 02	0, 0074	-0, 0952	0, 1101	0, 0074	-0, 2354	0, 2165
$\sigma$	0, 71	0, 8073	0, 3882	1, 2960	—	—	—

Parâmetros		Modelo bayesiano resposta lognormal ( $n = 100$ )			Modelo logístico usual ( $n = 100$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-12, 0	-4, 6233	-6, 9949	-2, 2525	-4, 6224	-9, 8725	0, 2543
$\beta_1$	0, 50	0, 5268	0, 2339	0, 8187	0, 5263	-0, 0569	1, 1883
$\beta_2$	0, 35	0, 3590	0, 2045	0, 5133	0, 3590	0, 0478	0, 7068
$\beta_3$	0, 02	0, 0136	-0, 0488	0, 0761	0, 0136	-0, 1232	0, 1438
$\sigma$	0, 71	0, 7748	0, 3826	1, 2608	—	—	—

Parâmetros		Modelo bayesiano resposta lognormal ( $n = 200$ )			Modelo logístico usual ( $n = 200$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-12, 0	-4, 3914	-6, 0723	-2, 7067	-4, 3910	-8, 0456	-0, 9076
$\beta_1$	0, 50	0, 4882	0, 3029	0, 6730	0, 4879	0, 1084	0, 8922
$\beta_2$	0, 35	0, 3426	0, 2351	0, 4504	0, 3426	0, 1227	0, 5786
$\beta_3$	0, 02	0, 0157	-0, 0304	0, 0619	0, 0157	-0, 0830	0, 1127
$\sigma$	0, 71	0, 7395	0, 3789	1, 2372	—	—	—

Parâmetros		Modelo bayesiano resposta lognormal ( $n = 500$ )			Modelo logístico usual ( $n = 500$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-12, 0	-4, 6486	-5, 7101	-3, 5840	-4, 6485	-6, 9325	-2, 4420
$\beta_1$	0, 50	0, 4928	0, 3761	0, 6092	0, 4927	0, 2513	0, 7437
$\beta_2$	0, 35	0, 3549	0, 2906	0, 4193	0, 3550	0, 2221	0, 4939
$\beta_3$	0, 02	0, 0209	-0, 0087	0, 0504	0, 0209	-0, 0416	0, 0833
$\sigma$	0, 71	0, 7026	0, 3755	1, 2450	—	—	—

A partir dos resultados obtidos na Tabela 5.12, podemos observar que as estimativas bayesianas intervalares são mais precisas em todos os casos considerados.

### 5.13 Resultados obtidos para o modelo logístico bayesiano com resposta exponencial

Para o modelo bayesiano de regressão logística com resposta exponencial, geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(15; 4); X_{i3} \sim N(45; 25); R_i \sim Exp(\lambda_i),$$

em que  $Exp$  denota a distribuição exponencial, e

$$\lambda_i = -\frac{\ln[g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})]}{C_R}, i = 1, 2, \dots, n.$$



Os valores atribuídos para os coeficientes de regressão foram  $\beta_0 = -24, 0$ ,  $\beta_1 = 1, 2$ ,  $\beta_2 = 0, 61$  e  $\beta_3 = 0, 19$ . Consideramos o ponto de corte  $C_R = 40.000$ , isto é, todos os valores de  $R_i$  acima de 40.000 foram considerados sucessos, ou seja, tiveram a resposta igual a 1. As seguintes distribuições *a priori* vagas foram consideradas

$$\beta_j \sim N(0, 1000), j = 0, 1, 2, 3.$$

A Tabela 5.13 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo bayesiano de regressão logística com resposta exponencial, para cada tamanho amostral adotado.

Tabela 5.13. Estimativas bayesianas e clássicas obtidas segundo ambos os modelos.

Parâmetros	Modelo bayesiano resposta exponencial ( $n = 50$ )				Modelo logístico usual ( $n = 50$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$ -24, 0	-29, 2335	-37, 8247	-20, 6584	-29, 2477	-51, 3905	-14, 3516	
$\beta_1$ 1, 20	1, 4716	0, 9528	1, 9907	1, 4719	0, 5152	2, 7546	
$\beta_2$ 0, 61	0, 7373	0, 4936	0, 9803	0, 7369	0, 2921	1, 3405	
$\beta_3$ 0, 29	0, 2308	0, 1398	0, 3215	0, 2307	0, 0691	0, 4623	
Parâmetros	Modelo bayesiano resposta exponencial ( $n = 100$ )				Modelo logístico usual ( $n = 100$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$ -24, 0	-25, 6346	-30, 7313	-20, 5303	-25, 6369	-37, 8271	-16, 1176	
$\beta_1$ 1, 20	1, 2921	0, 9663	1, 6183	1, 2924	0, 6635	2, 0520	
$\beta_2$ 0, 61	0, 6439	0, 4852	0, 8030	0, 6441	0, 3351	1, 0115	
$\beta_3$ 0, 29	0, 2047	0, 1451	0, 2644	0, 2048	0, 0907	0, 3448	
Parâmetros	Modelo bayesiano resposta exponencial ( $n = 200$ )				Modelo logístico usual ( $n = 200$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$ -24, 0	-24, 2320	-27, 6984	-20, 7689	-24, 2365	-32, 1467	-17, 4813	
$\beta_1$ 1, 20	1, 2292	1, 0178	1, 4404	1, 2291	0, 8108	1, 7062	
$\beta_2$ 0, 61	0, 6129	0, 4993	0, 7269	0, 6132	0, 3862	0, 8685	
$\beta_3$ 0, 29	0, 1910	0, 1499	0, 2320	0, 1910	0, 1093	0, 2832	
Parâmetros	Modelo bayesiano resposta exponencial ( $n = 500$ )				Modelo logístico usual ( $n = 500$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$ -24, 0	-24, 1465	-26, 2714	-22, 0185	-24, 1454	-28, 8484	-19, 8814	
$\beta_1$ 1, 20	1, 2140	1, 0839	1, 3438	1, 2137	0, 9504	1, 4993	
$\beta_2$ 0, 61	0, 6147	0, 5446	0, 6849	0, 6147	0, 4726	0, 7689	
$\beta_3$ 0, 29	0, 1900	0, 1648	0, 2152	0, 1900	0, 1386	0, 2450	

A partir das estimativas bayesianas obtidas na Tabela 5.13 podemos observar que as estimativas bayesianas intervalares são mais precisas, isto é, apresentaram uma menor amplitude nos intervalos de credibilidade para todos os casos considerados.

## 5.14 Resultados obtidos com o modelo logístico bayesiano com resposta geométrica

Geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte

forma:

$$X_{i1} \sim N(35; 25); X_{i2} \sim N(5; 1); X_{i3} \sim N(10; 4); R_i \sim Geom(p_i),$$

em que *Geom* denota a distribuição geométrica. Adotamos a constante  $C_R = 0$  e, portanto, temos que

$$p_i = 1 - \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}, i = 1, 2, \dots, n.$$

Os valores atribuídos para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\pi_i$ ,  $i = 1, 2, \dots, n$ , foram  $\beta_0 = -20, 0$ ,  $\beta_1 = 0, 2$ ,  $\beta_2 = 1, 0$  e  $\beta_3 = 0, 7$ .

Para cada um dos parâmetros do modelo, utilizamos uma distribuição *a priori* normal vaga tal que

$$\beta_j \sim N(0, 1000), j = 0, 1, 2, 3.$$

A Tabela 5.14 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta geométrica, para cada tamanho amostral considerado.

Tabela 5.14. Estimativas clássicas e bayesianas obtidas segundo ambos os modelos.

Parâmetros		Modelo bayesiano resposta geométrica ( $n = 50$ )			Modelo logístico usual ( $n = 50$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-20, 0	-24, 7842	-32, 4855	-17, 0678	-24, 7839	-44, 9741	-11, 5852
$\beta_1$	0, 20	0, 2600	0, 1527	0, 3671	0, 2599	0, 0609	0, 5252
$\beta_2$	1, 00	1, 1987	0, 7257	1, 6730	1, 1995	0, 3280	2, 3760
$\beta_3$	0, 70	0, 8381	0, 5466	1, 1290	0, 8378	0, 3325	1, 5874
Parâmetros		Modelo bayesiano resposta geométrica ( $n = 100$ )			Modelo logístico usual ( $n = 100$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-20, 0	-22, 1085	-26, 6488	-17, 5742	-22, 1120	-33, 0966	-13, 7420
$\beta_1$	0, 20	0, 2282	0, 1622	0, 2945	0, 2283	0, 1003	0, 3822
$\beta_2$	1, 00	1, 0207	0, 7147	1, 3270	1, 0209	0, 4223	1, 7285
$\beta_3$	0, 70	0, 7886	0, 6027	0, 9744	0, 7885	0, 4391	1, 2291
Parâmetros		Modelo bayesiano resposta geométrica ( $n = 200$ )			Modelo logístico usual ( $n = 200$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-20, 0	-20, 8065	-23, 8330	-17, 7693	-20, 8015	-27, 8085	-14, 9661
$\beta_1$	0, 20	0, 2024	0, 1587	0, 2462	0, 2024	0, 1153	0, 3006
$\beta_2$	1, 00	1, 0520	0, 8298	1, 2747	1, 0521	0, 6072	1, 5497
$\beta_3$	0, 70	0, 7416	0, 6172	0, 8657	0, 7415	0, 4978	1, 0238
Parâmetros		Modelo bayesiano resposta geométrica ( $n = 500$ )			Modelo logístico usual ( $n = 500$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-20, 0	-20, 2831	-22, 0985	-18, 4776	-20, 2815	-24, 3268	-16, 6681
$\beta_1$	0, 20	0, 2020	0, 1751	0, 2289	0, 2020	0, 1475	0, 2608
$\beta_2$	1, 00	1, 0095	0, 8753	1, 1433	1, 0094	0, 7370	1, 3027
$\beta_3$	0, 70	0, 7149	0, 6405	0, 7892	0, 7149	0, 5647	0, 8787

A partir dos resultados obtidos na Tabela 5.14 podemos observar que o modelo bayesiano com resposta geométrica apresentou intervalos de credibilidade com amplitudes menores que as dos intervalos de confiança do modelo logístico usual.

## 5.15 Resultados obtidos para o modelo logístico bayesiano com resposta Poisson

Geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte forma:

$$X_{i1} \sim N(35; 25); X_{i2} \sim N(5; 1); X_{i3} \sim N(10; 4); R_i \sim \text{Poisson}(\lambda_i).$$

Adotamos a constante  $C_R = 0$  e, portanto, temos que

$$\lambda_i = -\ln \left[ 1 - \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})} \right], i = 1, 2, \dots, n.$$

Os valores atribuídos para os coeficientes de regressão foram  $\beta_0 = -15,0$ ,  $\beta_1 = 0,16$ ,  $\beta_2 = 1,22$  e  $\beta_3 = 0,48$ . Para cada um dos parâmetros do modelo, utilizamos uma distribuição *a priori* normal vaga tal que

$$\beta_j \sim N(0, 1000), j = 0, 1, 2, 3.$$

A Tabela 5.15 mostra as estimativas obtidas considerando o modelo de regressão logística usual e o modelo de regressão logística com resposta Poisson, para cada tamanho amostral considerado.

Tabela 5.15. Estimativas clássicas e bayesianas obtidas segundo ambos os modelos.

Parâmetros		Modelo bayesiano resposta Poisson ( $n = 50$ )			Modelo logístico usual ( $n = 50$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-15,0	-16,0874	-19,7397	-12,4383	-26,6592	-34,9741	-11,5675
$\beta_1$	0,16	0,1719	0,1076	0,2362	1,0116	0,0650	1,5252
$\beta_2$	1,22	1,3397	0,9594	1,7196	3,7183	0,3334	4,3760
$\beta_3$	0,48	0,4908	0,3526	0,6290	2,8221	0,3108	3,5874
Parâmetros		Modelo bayesiano resposta Poisson ( $n = 100$ )			Modelo logístico usual ( $n = 100$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-15,0	-16,0875	-19,7345	-12,4366	-16,0840	-24,7922	-9,1990
$\beta_1$	0,16	0,1719	0,1077	0,2363	0,1719	0,0451	0,3204
$\beta_2$	1,22	1,3393	0,9579	1,7188	1,3381	0,6137	2,2381
$\beta_3$	0,48	0,4907	0,3530	0,6287	0,4908	0,2254	0,8168
Parâmetros		Modelo bayesiano resposta Poisson ( $n = 200$ )			Modelo logístico usual ( $n = 200$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-15,0	-15,4191	-17,9093	-12,9127	-15,4109	-21,0764	-10,5038
$\beta_1$	0,16	0,1657	0,1242	0,2072	0,1657	0,0819	0,2577
$\beta_2$	1,22	1,2299	0,9745	1,4859	1,2307	0,7259	1,8093
$\beta_3$	0,48	0,4976	0,3970	0,5982	0,4976	0,2976	0,7244
Parâmetros		Modelo bayesiano resposta Poisson ( $n = 500$ )			Modelo logístico usual ( $n = 500$ )		
Valores reais		Estimativas	Int.Cred(95%)		Estimativas	IC (95%)	
$\beta_0$	-15,0	-14,9844	-16,4998	-13,4653	-14,9826	-18,3293	-11,9167
$\beta_1$	0,16	0,1595	0,1344	0,1848	0,1596	0,1078	0,2143
$\beta_2$	1,22	1,2290	1,0854	1,3731	1,2291	0,9378	1,5454
$\beta_3$	0,48	0,4733	0,4093	0,5376	0,4735	0,3430	0,6137

A partir dos resultados expressos pela Tabela 5.15, assim como nos modelos anteriores, podemos observar que o modelo bayesiano com resposta geométrica apresentou intervalos de credibilidade com amplitudes menores que as dos intervalos de confiança do modelo logístico usual.

## 5.16 Resultados obtidos para o modelo geométrico bayesiano com resposta normal

Para o caso do modelo geométrico com resposta normal, geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(10; 4); X_{i3} \sim N(20; 16); R_i \sim N(\mu_i, \sigma^2),$$

com  $\sigma = 1.000$  e  $\mu_i = \sigma \Phi^{-1} [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})] + C_R, i = 1, 2, \dots, n$ .

Os valores atribuídos para o vetor  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$  para a geração de  $\mu_i, i = 1, 2, \dots, n$ , foram  $\beta_0 = -4,5, \beta_1 = 0,5, \beta_2 = 0,35$  e  $\beta_3 = 0,02$ .

Consideramos os pontos de corte  $C_R = 11.000$ , e  $C_Y = 2$ , e as seguintes distribuições *a priori* vaga

$$\sigma^2 \sim IG(1, 01; 100) \quad \text{e} \quad \beta_j \sim N(0, 1000), \quad j = 0, 1, 2, 3.$$

A Tabela 5.16 mostra as estimativas obtidas considerando o modelo de regressão geométrica e o modelo de regressão geométrica bayesiano com resposta normal, para cada tamanho amostral.

Tabela 5.16. Estimativas clássicas e bayesianas obtidas segundo ambos os modelos.

Parâmetros	Modelo bayesiano resposta normal ( $n = 50$ )				Modelo logístico usual ( $n = 50$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-5, 2801	-9, 0701	-1, 4916	-5, 2854	-13, 2698	2, 6991
$\beta_1$	0, 50	0, 5644	0, 1154	1, 0128	0, 5644	-0, 3804	1, 5093
$\beta_2$	0, 35	0, 4129	0, 1984	0, 6277	0, 4130	-0, 0394	0, 8655
$\beta_3$	0, 02	0, 0280	-0, 0575	0, 1135	0, 0279	-0, 1518	0, 2076
$\sigma$	1000	1191	832	1477	—	—	—

Parâmetros	Modelo bayesiano resposta normal ( $n = 100$ )				Modelo logístico usual ( $n = 100$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-4, 5101	-6, 5785	-2, 4452	-4, 5140	-8, 8654	-0, 1626
$\beta_1$	0, 50	0, 5193	0, 2486	0, 7907	0, 5196	-0, 0514	1, 0906
$\beta_2$	0, 35	0, 3610	0, 2179	0, 5041	0, 3610	0, 0597	0, 6624
$\beta_3$	0, 02	0, 0175	-0, 0396	0, 0745	0, 0174	-0, 1028	0, 1375
$\sigma$	1000	1044	827	1381	—	—	—

Parâmetros	Modelo bayesiano resposta normal ( $n = 200$ )				Modelo logístico usual ( $n = 200$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-4, 3947	-5, 8677	-2, 9177	-4, 3937	-7, 5020	-1, 2854
$\beta_1$	0, 50	0, 5191	0, 3461	0, 6918	0, 5191	0, 1546	0, 8836
$\beta_2$	0, 35	0, 3530	0, 2564	0, 4499	0, 3532	0, 1494	0, 5569
$\beta_3$	0, 02	0, 0128	-0, 0301	0, 0555	0, 0127	-0, 0774	0, 1029
$\sigma$	1000	997	826	1210	—	—	—

Parâmetros	Modelo bayesiano resposta normal ( $n = 500$ )				Modelo logístico usual ( $n = 500$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-4, 50	-4, 5041	-5, 4362	-3, 5682	-4, 5030	-6, 4723	-2, 5337
$\beta_1$	0, 50	0, 5016	0, 3953	0, 6082	0, 5018	0, 2778	0, 7259
$\beta_2$	0, 35	0, 3508	0, 2935	0, 4083	0, 3509	0, 2300	0, 4718
$\beta_3$	0, 02	0, 0213	-0, 0061	0, 0486	0, 0212	-0, 0365	0, 0788
$\sigma$	1000	995	826	1208	—	—	—

A exemplo do que ocorreu com os casos do modelo logístico com respostas diversas, a partir dos resultados apresentados na Tabela 5.16, verificamos que a amplitude os intervalos de credibilidades são menores em relação aos intervalos de confiança, para todos os casos considerados.

## 5.17 Resultados obtidos para o modelo geométrico bayesiano com resposta exponencial

Para o modelo bayesiano de regressão logística com resposta exponencial, geramos 100 amostras de tamanho  $n$  ( $n = 50, 100, 200$  e  $500$ ), considerando três variáveis explicativas  $X_{i1}, X_{i2}, X_{i3}$  e uma variável resposta e  $R_i, i = 1, 2, \dots, n$ , da seguinte forma:

$$X_{i1} \sim N(5; 1); X_{i2} \sim N(15; 4); X_{i3} \sim N(45; 25); R_i \sim Exp(\lambda_i),$$

em que  $Exp$  denota a distribuição exponencial, e

$$\lambda_i = -\frac{\ln [g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})]}{C_R}, \quad i = 1, 2, \dots, n.$$

Os valores atribuídos para os coeficientes de regressão foram  $\beta_0 = -24, 0$ ,  $\beta_1 = 1, 2$ ,  $\beta_2 = 0, 61$  e  $\beta_3 = 0, 19$ . Consideramos os ponto de corte  $C_R = 40.000$  e  $C_Y = 5$ . As seguintes distribuições *a priori* vagas foram consideradas

$$\beta_j \sim N(0, 1000), \quad j = 0, 1, 2, 3.$$

A Tabela 5.17 mostra as estimativas obtidas considerando o modelo de regressão geométrica e o modelo de regressão geométrica com resposta exponencial, para cada tamanho amostral adotado.

Tabela 5.17. Estimativas bayesianas e clássicas obtidas segundo ambos os modelos.

Parâmetros	Modelo bayesiano resposta exponencial ( $n = 50$ )				Modelo logístico usual ( $n = 50$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-24, 0	-24, 1104	-27, 1351	-21, 0874	-24, 1072	-30, 4781	-17, 7362
$\beta_1$	1, 20	1, 2147	0, 9651	1, 4631	1, 2143	0, 6906	1, 7381
$\beta_2$	0, 61	0, 6183	0, 5099	0, 7266	0, 6183	0, 3900	0, 8466
$\beta_3$	0, 19	0, 1893	0, 1561	0, 2225	0, 1893	0, 1192	0, 2593
Parâmetros	Modelo bayesiano resposta exponencial ( $n = 100$ )				Modelo logístico usual ( $n = 100$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-24, 0	-24, 0261	-25, 8840	-22, 1695	-24, 0277	-27, 9408	-20, 1146
$\beta_1$	1, 20	1, 1899	1, 0361	1, 3439	1, 1901	0, 8660	1, 5143
$\beta_2$	0, 61	0, 6149	0, 5351	0, 6945	0, 6148	0, 4467	0, 7829
$\beta_3$	0, 19	0, 1908	0, 1664	0, 2152	0, 1908	0, 1393	0, 2422
Parâmetros	Modelo bayesiano resposta exponencial ( $n = 200$ )				Modelo logístico usual ( $n = 200$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-24, 0	-24, 0607	-25, 4190	-22, 6973	-24, 0584	-26, 9269	-21, 1899
$\beta_1$	1, 20	1, 1986	1, 1012	1, 2958	1, 1986	0, 9936	1, 4036
$\beta_2$	0, 61	0, 6094	0, 5544	0, 6644	0, 6094	0, 4937	0, 7251
$\beta_3$	0, 19	0, 1920	0, 1733	0, 2107	0, 1920	0, 1527	0, 2313
Parâmetros	Modelo bayesiano resposta exponencial ( $n = 500$ )				Modelo logístico usual ( $n = 500$ )		
Valores reais	Estimativas	Int.Cred(95%)		Estimativas	IC (95%)		
$\beta_0$	-24, 0	-24, 0161	-24, 8712	-23, 1564	-24, 0144	-25, 8211	-22, 2076
$\beta_1$	1, 20	1, 1965	1, 1361	1, 2567	1, 1965	1, 0693	1, 3236
$\beta_2$	0, 61	0, 6131	0, 5807	0, 6457	0, 6132	0, 5448	0, 6816
$\beta_3$	0, 19	0, 1898	0, 1777	0, 2019	0, 1898	0, 1644	0, 2152

Novamente, a partir dos resultados apresentados na Tabela 5.17, verificamos que a amplitude os intervalos de credibilidades são menores em relação aos intervalos de confiança, para todos os casos considerados. Neste sentido, o modelo de regressão geométrica bayesiano com resposta de origem apresenta estimativas mais precisas.

## 6 Aplicações a dados reais

### 6.1 Introdução

Nesta seção apresentamos a aplicação da metodologia proposta por meio de dois conjuntos de dados reais. O primeiro conjunto se refere ao nível de concentração dos hidrocarbonetos policíclicos aromáticos (HPAs) estudado por Paraíba *et. al* (2010), que poderiam estar presentes em um lodo de esgoto usado como fertilizante agrícola. Para este conjunto aplicamos o modelo de regressão logística com resposta exponencial. O segundo conjunto se refere a dados do SERASA sobre informações financeiras de empresas. Para este conjunto aplicamos o modelo de regressão logística com resposta geométrica inflacionada em zeros.

### 6.2 Aplicação do modelo de regressão logística com resposta exponencial

Nessa Seção vamos ilustrar a metodologia aplicando-a a dados reais usando o caso particular do modelo de regressão logística com resposta exponencial. Paraíba *et. al* (2010) estudaram o nível de concentração dos hidrocarbonetos policíclicos aromáticos (HPAs) que, eventualmente, poderiam estar presentes em um lodo de esgoto usado como fertilizante agrícola. Os seguintes HPAs foram quantificados: naftaleno, acenaftileno, acenafteno, fluoreno, antraceno, fluoranteno, pireno, benzo(a)antraceno, criseno, benzo(b)fluoranteno, benzo(k)fluoranteno, benzo(a)pireno, indeno(1,2,3-c,d)pireno, dibenzoantraceno, benzoperileno, fenantreno, cujos níveis de concentração encontrados em lodos de esgotos podem apresentar risco de contaminação de solos. Tal contaminação foi monitorada em grãos de milho cultivados em solos tratados com lodo de esgoto que receberam tais HPAs.

O limite detectável (LD) e o limite quantificável (LQ) para o fenantreno são respectivamente  $0,28 \mu g kg^{-1}$  e  $0,89 \mu g kg^{-1}$ . Para esse conjunto real de dados temos  $n = 72$  observações para a variável resposta fenantreno, dentre as quais 19 apresentaram um nível de concentração abaixo do nível quantificável  $0,89 \mu g kg^{-1}$ .

A Figura 6.1 apresenta o histograma da variável resposta fenantreno.

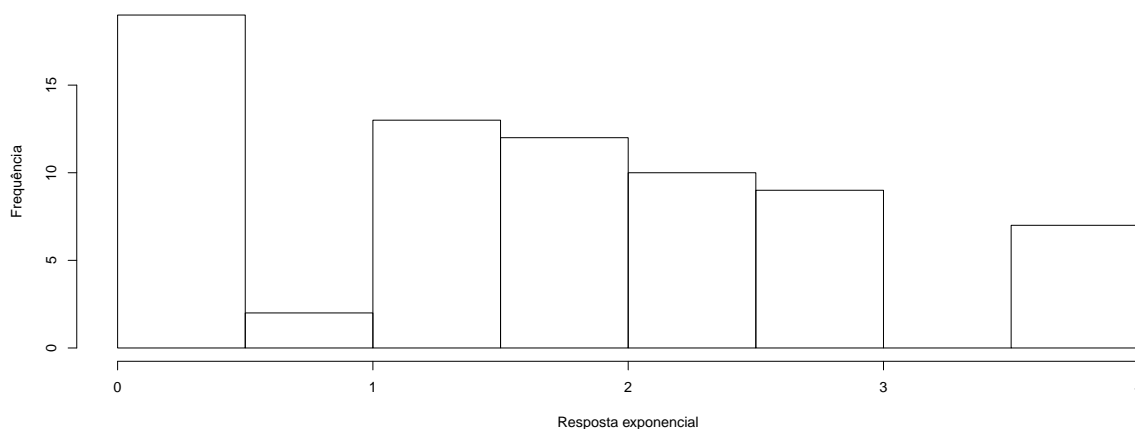


Figura 6.1: Histograma da variável resposta Fenantreno.

Segundo Paraíba *et. al* (2010), valores acima de  $2 \mu g kg^{-1}$  sugerem contaminação nos grãos de milho e, portanto, impróprio para o consumo. Desta forma consideramos o ponto de corte  $C_R = 2 \mu g kg^{-1}$ , ou seja, os valores acima de  $2 \mu g kg^{-1}$  foram considerados sucessos e abaixo foram considerados fracassos. Foram utilizados 3 tratamentos distintos: DLN1, DLN8 e FNPk, e o tratamento controle (TEST). Os tratamentos foram indicados com variáveis auxiliares conforme descrito abaixo:

Tratamentos	Auxiliares		
	$X_1$	$X_2$	$X_3$
FNPk	1	0	0
DLN1	0	1	0
DLN8	0	0	1
TEST	0	0	0

A Tabela 6.2 apresenta as estimativas obtidas para os coeficientes de regressão segundo o modelo de regressão logística usual, o modelo de regressão logística com a informação da variável resposta exponencial, bem como sua versão bayesiana. Para o modelo bayesiano foram usados como chutes iniciais as estimativas dos coeficientes estimados segundo o modelo logístico usual, e com *prioris* vagas para o vetor  $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$  da seguinte forma:

$$\beta \sim N_4(\mathbf{0}, 1000\mathbf{I}).$$

Foram geradas duas cadeias de 120.000 iterações com um salto de tamanho 20 com um descarte das 20.000 primeiras iterações totalizando uma amostra final da *posteriori* conjunta de tamanho 5.000. A convergência foi monitorada por meio do critério de Gelman e Rubin (1992).

Tabela 6.2. Estimativas obtidas segundo o ajuste de cada modelo.

Resultados do modelo logístico usual				
Coeficientes	Estimativa	Erro-Padrão	Int.Conf(95%)	
$\beta_0$	-1,2528	0,5669	-2,5135	-0,2276
$\beta_1$	-0,3567	0,8494	-2,1232	1,3152
$\beta_2$	0,2973	0,7735	-1,2257	1,8733
$\beta_3$	2,5055	0,8018	1,0278	4,2156
Resultados do modelo logístico com resposta Exponencial				
Coeficientes	Estimativa	Erro-Padrão	Int.Conf(95%)	
$\beta_0$	-1,9177	0,5555	-3,0065	-0,8289
$\beta_1$	0,1185	0,7721	-1,3948	1,6318
$\beta_2$	0,8910	0,7004	-0,4818	2,2638
$\beta_3$	1,8067	0,6487	0,5353	3,0781
Resultados do modelo logístico bayesiano com resposta Exponencial				
Coeficientes	Estimativa	Erro-Padrão	Int.Cred(95%)	
$\beta_0$	-2,3251	0,0385	-2,3630	-2,2219
$\beta_1$	-1,7795	0,2244	-2,0146	-1,1798
$\beta_2$	-1,0172	0,1914	-1,2135	-0,5212
$\beta_3$	1,0461	0,1094	0,9368	1,3379



Podemos observar a partir dos resultados obtidos pela Tabela 6.2 que as variáveis  $X_1$  e  $X_2$  não foram significativas para os modelos logístico e logístico com resposta exponencial, ou seja, o tratamento FNPk e DLN1 não interfere no nível de concentração de fenantreno no lodo de esgoto. O erro-padrão das estimativas obtidas para os coeficientes segundo o modelo de regressão logística com resposta exponencial foi menor para todos os coeficientes de regressão. Desta forma o modelo logístico com resposta exponencial forneceu estimativas intervalares mais precisas.

Com relação as estimativas bayesianas, a amplitude dos intervalos de credibilidade de todos os coeficientes de regressão foi menor que a amplitude dos intervalos de confiança do modelo logístico e do modelo logístico com resposta exponencial, mesmo utilizando distribuições *a priori* vagas. Além disso, todas as variáveis foram significativas para o modelo, isto é, os quatro tratamentos interfere no nível de concentração de fenantreno no lodo de esgoto. A Tabela 6.2 apresenta o diagnóstico de convergência segundo o critério de Gelman e Rubin (1992).

Tabela 6.2. Convergência de Gelman e Rubin (1992).

Parâmetro	Est. Pontual	Q. 97,5%
$\beta_0$	1,00	1,00
$\beta_1$	1,00	1,00
$\beta_2$	1,00	1,00
$\beta_3$	1,00	1,00

A figura 6.2 apresenta a trajetória das cadeias para cada um dos coeficientes de regressão segundo o modelo com enfoque bayesiano.

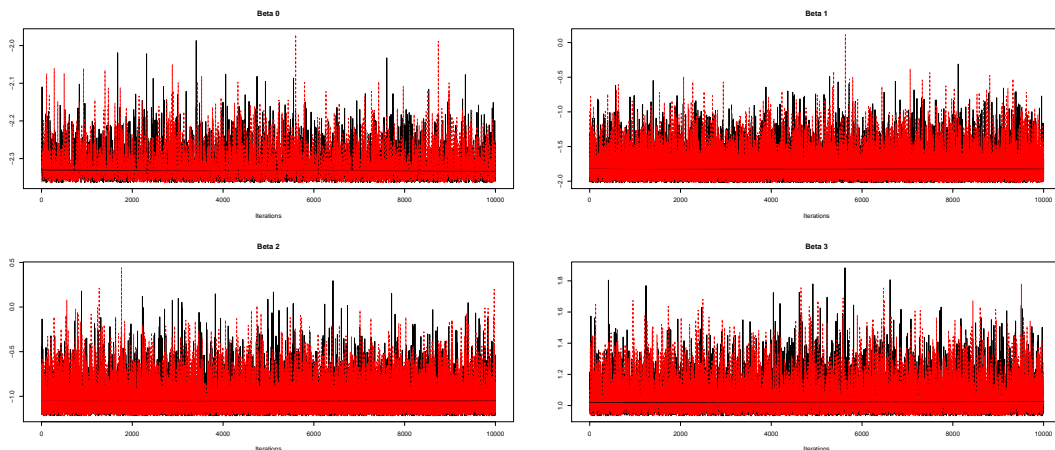


Figura 6.2: Trajetória das cadeias para cada parâmetro

A figura 6.3 apresenta o diagnóstico gráfico de convergência segundo o critério de Gelman e Rubin (1992) para cada um dos coeficientes de regressão segundo o modelo com enfoque bayesiano.

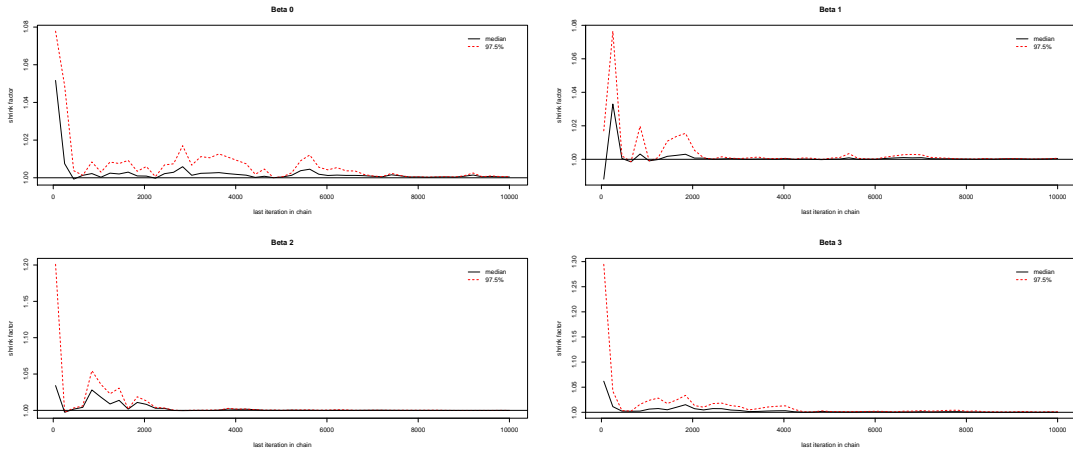


Figura 6.3: Convergência gráfica de Gelman e Rubin (1992)

A figura 6.4 apresenta a autocorrelação das cadeias para cada um dos coeficientes de regressão segundo o modelo com enfoque bayesiano.

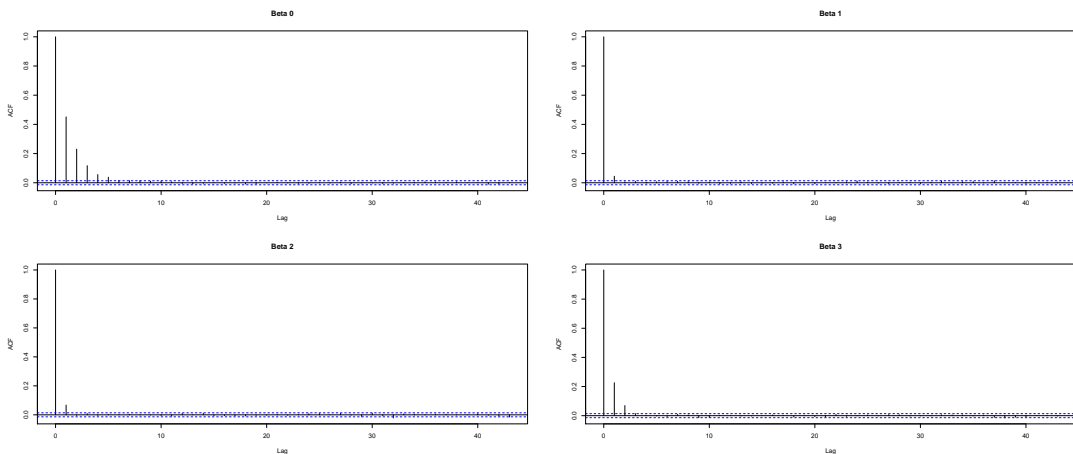


Figura 6.4: Autocorrelação das cadeias

Para os três modelos considerados usamos os seguintes resíduos padronizados para a resposta de interesse  $Y_i$ :

$$rp_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}},$$

em que  $\hat{\pi}_i = g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$ , e  $\hat{\boldsymbol{\beta}}$  são as estimativas segundo os três modelos. A Figura 6.5 apresenta os resíduos padronizados referentes a variável de interesse  $Y_i$  sobre as probabilidades de sucesso estimadas.

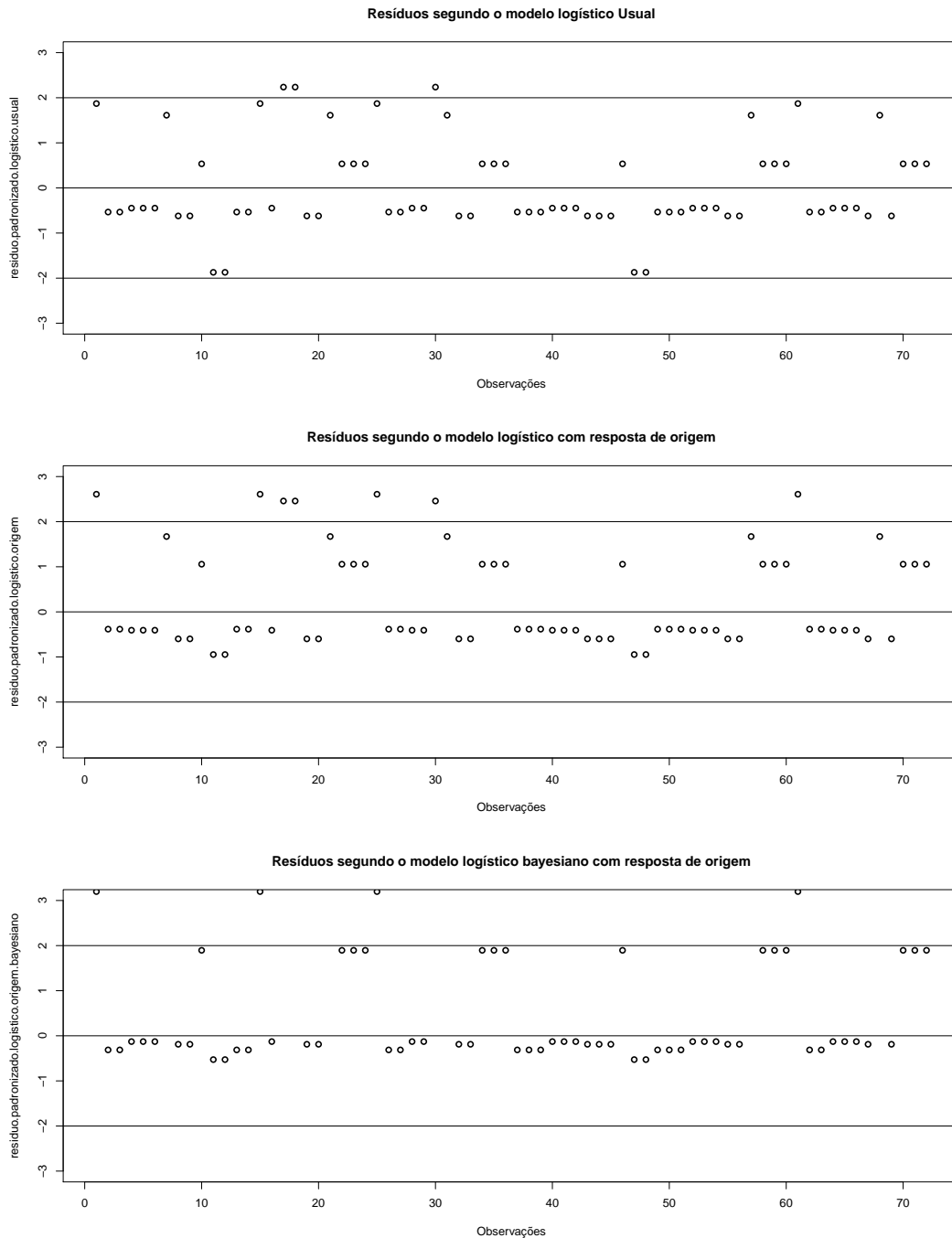


Figura 6.5: Resíduos padronizados referentes a variável de interesse  $Y_i$

### 6.3 Aplicação do modelo logístico com resposta geométrica inflacionada de zeros

Nessa Seção ilustramos a metodologia aplicando-a a dados reais usando o caso particular do modelo de regressão logística com resposta geométrica zero-inflacionada. Utilizamos um conjunto de dados do SERASA sobre informações financeiras de uma amostra de 646 empresas. Consideramos 5 variáveis preditoras  $X_{i1}, X_{i2}, \dots, X_{i5}$ , para  $i = 1, 2, \dots, 646$ , e a variável resposta  $R_i$ , conforme descrição abaixo:

Variável preditora  $X_1$  (Sigla VSOC CH30 EMP): Passagens do recheque entre 0 e 30 dias, referentes aos sócios participantes das empresas.

Variável preditora  $X_2$  (Sigla VSOC CONC60 EMP): Consultas ao CONCENTRE + CREDIT RELATO RATING + CREDIT BUREAU entre 31 e 60 dias, referentes aos sócios participantes das empresas.

Variável preditora  $X_3$  (Sigla VSOC CONC90 EMP): Consultas ao CONCENTRE + CREDIT RELATO RATING + CREDIT BUREAU entre 61 e 90 dias, referentes aos sócios participantes das empresas.

Variável preditora  $X_4$  (Sigla VSOC DIF EM30 EMP): Número de empresas diferentes que consultaram o mesmo CNPJ entre 0 e 30 dias.

Variável preditora  $X_5$  (Sigla VSOC DIF EM90 EMP): Número de empresas diferentes que consultaram o mesmo CNPJ entre 0 e 90 dias.

Variável resposta  $R$  (Sigla VSOC CH90 EMP): Passagens do recheque entre 61 e 90 dias.

O tamanho do conjunto de dados é de  $n = 646$  observações para cada variável. Com relação a variável resposta  $R$ , 579 observações assumem o valor zero, isto é, aproximadamente 90% do total, e as 67 observações restantes variaram entre 1 e 108 passagens do recheque. A Figura 6.6 apresenta o histograma da variável resposta **Passagens do recheque entre 61 e 90 dias** (Sigla VSOC CH90 EMP).

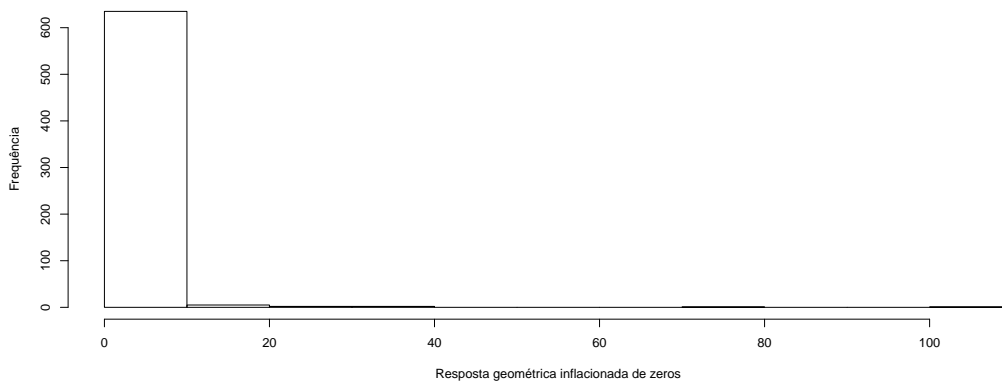


Figura 6.6: Histograma da variável resposta Passagens do recheque entre 61 e 90 dias (VSOC CH90 EMP).

Para o ajuste do modelo logístico usual consideramos uma constante  $C_R = 0$  passagens de recheque, ou seja, o valor da variável resposta binária será fracasso  $Y_i = 0$  se  $R_i = 0$  passagens do recheque, e será sucesso  $Y_i = 1$  se  $R_i > 0$  passagens do recheque, para  $i = 1, 2, \dots, 646$ . Dessa maneira, com  $C_R = 0$  temos 67 sucessos e 579 fracassos.

A Tabela 6.3 apresenta as estimativas obtidas para os coeficientes de regressão segundo o ajuste do modelo logístico usual e o ajuste do modelo com resposta geométrica inflacionada de zeros.

Tabela 6.3. Estimativas obtidas segundo o ajuste de cada modelo.

Resultados do modelo logístico usual				
Coefficientes	Estimativa	Erro padrão	Int.Conf(95%)	
$\beta_0$	-3,7431	0,2644	-4,3018	-3,2600
$\beta_1$	0,9751	0,2588	0,4686	1,5044
$\beta_2$	-0,2179	0,0557	-0,3315	-0,1130
$\beta_3$	-0,0930	0,0537	-0,1966	0,0159
$\beta_4$	-0,3400	0,0857	-0,5130	-0,1770
$\beta_5$	0,3517	0,0688	0,2220	0,4913
Resultados do modelo logístico com resposta geométrica inflacionada de zeros				
Coefficientes	Estimativa	Erro padrão	Int.Conf(95%)	
$\beta_0$	-3,5639	0,1209	-3,5639	-3,0899
$\beta_1$	0,1147	0,0196	0,0763	0,1531
$\beta_2$	0,0072	0,0082	-0,0089	0,0233
$\beta_3$	-0,0089	0,0080	-0,0246	0,0068
$\beta_4$	-0,0275	0,0107	-0,0485	-0,0065
$\beta_5$	0,0125	0,0082	-0,0036	0,0286
$\omega$	0,9812	0,0019	0,9775	0,9849

De acordo com os resultados obtidos na Tabela 6.3, o modelo de regressão logística com a informação da variável resposta geométrica inflacionada de zeros apresenta o erro padrão das estimativas dos coeficientes invariavelmente menor que o erro-padrão das estimativas do modelo logístico usual, produzindo assim estimativas intervalares mais precisas.

A Figura 6.7 apresenta os resíduos padronizados das probabilidades de sucesso considerando o ajuste do modelo logístico usual.

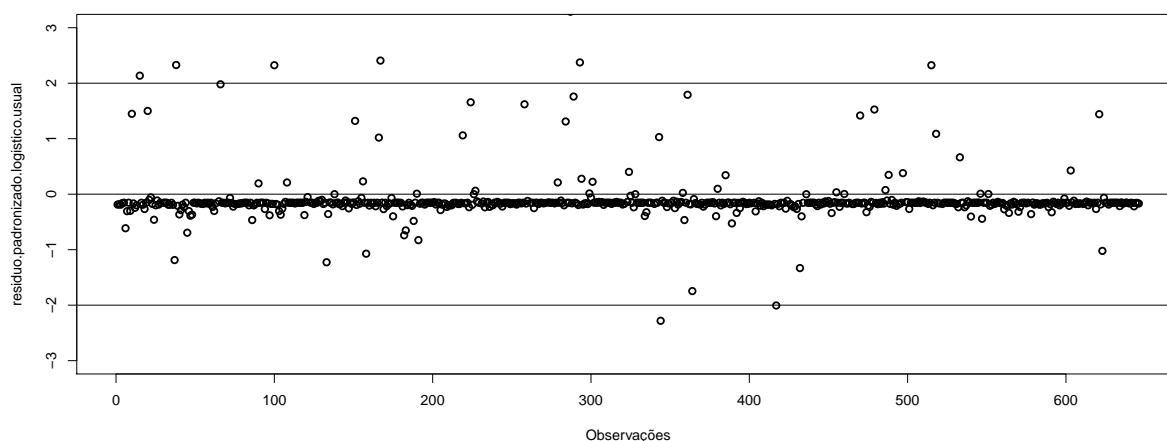


Figura 6.7: Resíduos padronizados considerando o ajuste do modelo logístico usual.

A Figura 6.8 apresenta os resíduos padronizados considerando o ajuste do modelo logístico usual.

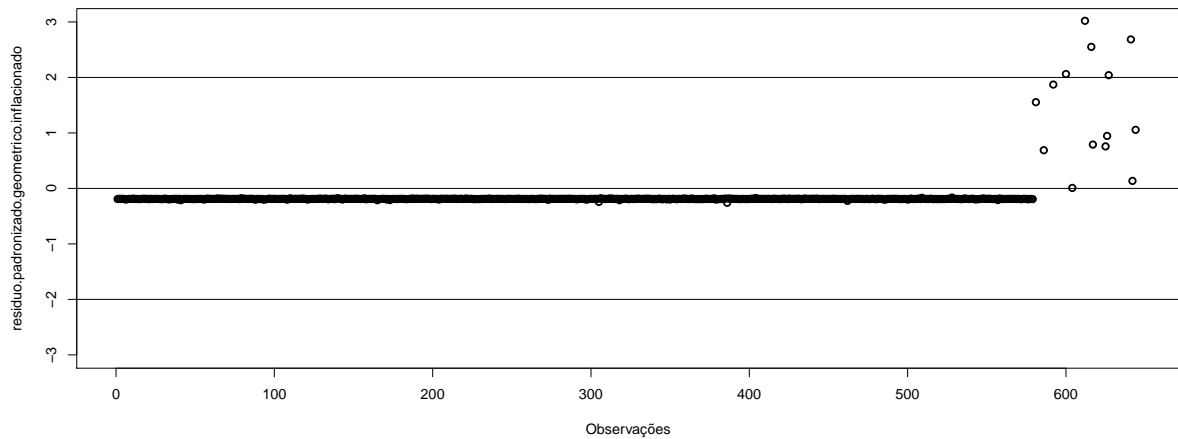


Figura 6.8: Resíduos padronizados do modelo logístico com resposta geométrica inflacionada de zeros.

Para efeito de comparação de desempenho dos modelos, a Tabela 6.3 mostra as estimativas obtidas para os coeficientes de regressão considerando diferentes valores de  $C_R$ .

Tabela 6.3. Estimativas para os coeficientes considerando diferentes valores de  $C_R$ .

	$C_R=2$ (617 Zeros e 29 Uns)				$C_R=3$ (625 Zeros e 21 Sucessos)			
	Resposta Logística		Resposta Geométrica		Resposta Logística		Resposta Geométrica	
	Estimativa	Erro-padrão	Estimativa	Erro-padrão	Estimativa	Erro-padrão	Estimativa	Erro-padrão
$\hat{\beta}_0$	-6,1378	0,7588	-5,8886	0,2708	-6,4131	0,8969	-7,5865	0,3431
$\hat{\beta}_1$	0,8510	0,3006	0,0401	0,0194	0,7882	0,2756	0,0348*	0,0211
$\hat{\beta}_2$	-0,0403*	0,0556	-0,0355*	0,0207	0,0799*	0,0731	-0,0326*	0,0220
$\hat{\beta}_3$	0,0615*	0,0697	0,0364*	0,0255	0,0671*	0,0891	0,0141*	0,0303
$\hat{\beta}_4$	-0,2389	0,0958	-0,0978	0,0349	-0,2206	0,1096	-0,1385*	0,0413
$\hat{\beta}_5$	0,1355	0,0632	0,0667	0,0226	0,0422*	0,0834	0,0953	0,0266
	$C_R=5$ (629 Zeros 17 Uns)				$C_R=10$ (635 Zeros e 11 Uns)			
	Resposta Logística		Resposta Geométrica		Resposta Logística		Resposta Geométrica	
	Estimativa	Erro-padrão	Estimativa	Erro-padrão	Estimativa	Erro-padrão	Estimativa	Erro-padrão
$\hat{\beta}_0$	-8,7674	2,3773	-11,0356	0,4950	-164,44*	22586,988	-19,2901	0,8628
$\hat{\beta}_1$	1,2126*	0,8376	0,0221*	0,0256	3,354*	676,936	-0,0163*	0,0316
$\hat{\beta}_2$	-0,0054*	0,1239	-0,0320*	0,0247	-8,559*	1425,513	-0,0915*	0,0391
$\hat{\beta}_3$	0,0692*	0,1047	-0,0340*	0,0332	-8,028*	1296,690	-0,1184*	0,0399
$\hat{\beta}_4$	-0,3058*	0,1844	-0,2313*	0,0518	-3,087*	2015,613	-0,3704*	0,0742
$\hat{\beta}_5$	0,1123*	0,1264	0,1621	0,0325	9,746*	1573,507	0,3035	0,0416

Valores acompanhados de (\*) não foram significativos para o modelo considerando com 5% de significância.

Observamos por meio da Tabela 6.3 que para todos os valores adotados de  $C_R$ , o erro-padrão das estimativas dos coeficientes segundo o ajuste do modelo de regressão logística com resposta geométrica zero-inflacionada é invariavelmente menor que o erro-padrão das estimativas segundo o ajuste do modelo logístico usual.

A partir de  $C_R = 10$  (635 zeros e 11 uns) todos os coeficientes do modelo logístico usual foram não significativos, produzindo estimativas discrepantes, o que não ocorreu com o modelo de regressão logística com resposta geométrica zero inflacionada. A Tabela 6.3 mostra as estimativas obtidas considerando alguns casos extremos da constante  $C_R$  em que não foi possível o ajuste do modelo logístico usual.

Tabela 6.3. Estimativas obtidas considerando alguns casos extremos de  $C_R$

	$C_R=15$ (639 Zeros e 7 Uns)		$C_R=80$ (645 Zeros e 1 Um)		$C_R=108$ (646 Zeros e 0 Uns)		$C_R=120$ (646 Zeros e 0 Uns)	
	Estimativa	Erro-padrão	Estimativa	Erro-padrão	Estimativa	Erro-padrão	Estimativa	Erro-padrão
$\hat{\beta}_0$	-28,4565	1,2704	-135,8234	6,0423	-182,9073	7,6624	-187,7981	7,8409
$\hat{\beta}_1$	-0,0389*	0,0294	0,1169*	0,2586	0,1895*	0,1746	0,2292*	0,1743
$\hat{\beta}_2$	-0,1550	0,0368	-0,8883	0,1198	-1,2084	0,1171	-1,2284	0,1197
$\hat{\beta}_3$	-0,1755	0,0421	-0,2308*	0,3541	-0,2466*	0,2376	-0,2073*	0,2385
$\hat{\beta}_4$	-0,5197	0,0769	-2,0821	0,1394	-2,7822	0,1445	-2,8522	0,1459
$\hat{\beta}_5$	0,4421	0,0408	1,6688	0,2398	2,2097	0,1120	2,2378	0,1114

Valores acompanhados de (\*) não foram significativos para o modelo considerando com 5%de significância.

A Tabela 6.3 mostra que, mesmo com um quantitativo de 646 fracassos, isto é, vetor de resposta binária composto somente por zeros, foi possível ajustar o modelo de regressão logística com resposta geométrica zero-inflacionada em que algumas variáveis preditoras e o intercepto são significativos para o modelo.

## 7 Discussões e considerações finais

Neste capítulo apresentamos uma análise geral dos resultados obtidos pela metodologia proposta, que trata-se de uma família de modelos de regressão com a distribuição original da variável resposta. Abordamos as vantagens obtidas no uso dessa família de modelos e discutimos alguns aspectos sobre o ajuste dos modelos com e sem os pontos influentes. Discutimos a análise de resíduos sobre as probabilidades de sucessos segundo o modelo logístico usual e a metodologia proposta, tanto no enfoque clássico quanto no enfoque bayesiano. Por fim, apresentamos as perspectivas e possíveis desdobramentos futuros deste trabalho.

### 7.1 Sobre a metodologia proposta e ganhos inferenciais

De acordo com os resultados obtidos até o momento e, considerando todas as classes de distribuições abordadas neste trabalho, verificamos por meio do estudo de simulação que, ao incorporarmos a informação da distribuição de origem no ajuste do modelo de regressão, as estimativas pontuais dos coeficientes de regressão são mais próximas dos valores reais. O erro padrão obtido foi uniformemente menor para todos os parâmetros produzindo estimativas intervalares com menores amplitudes, ou seja, mais precisas. Verificamos também que, à medida que o tamanho da amostra aumenta, as estimativas pontuais de ambos os modelos se aproximam e, no entanto, o erro padrão continua sendo menor para todos os tamanhos de amostra considerados. Nesse sentido, o modelo de regressão com a informação da resposta de origem produz assim estimativas mais precisas que o modelo usual.

### 7.2 Sobre as probabilidades de cobertura dos intervalos de confiança dos parâmetros

De acordo com o estudo de simulação apresentado no Capítulo 3, a metodologia proposta apresentou uma maior cobertura dos intervalos de confiança dos parâmetros, fornecendo valores invariavelmente mais próximos do nominal considerando os casos 90%, 95% e 99% de confiança, para todos os modelos e tamanhos amostrais estudados.

### 7.3 Sobre a análise de resíduos nas probabilidades de sucesso

Monitorando as probabilidades de sucesso percebemos, por meio do estudo de simulação apresentado nesse trabalho, que os dados foram melhor ajustados segundo o modelo de regressão logística com a informação da resposta de origem, pois seus resíduos apresentaram uma variação menor que o modelo logístico usual para qualquer tamanho de amostra. Para pequenas amostras, a quantidade de amostras modeladas segundo o modelo de origem que apresenta um resíduo médio menor que o resíduo médio do modelo usual é maior. À medida que o tamanho da amostra aumenta, essa porcentagem tende a convergir para 50%, o que sugere que, para amostras grandes, o resíduo médio do modelo de origem é o mesmo do modelo logístico usual.

Além disso, verificamos que o desvio dos resíduos segundo o modelo de regressão logística com resposta de origem é, em quase sua totalidade menor que o desvio dos resíduos segundo o modelo logístico usual. Para amostras grandes isso é verdade 100% das vezes.



Esse fato sugere que os modelos com resposta de origem gozam de um melhor ajuste dos dados.

Em síntese, a vantagem do ajuste dos dados segundo o modelo de regressão logística com resposta de origem, para as probabilidades de sucesso, é especialmente verificada nas pequenas amostras em que o desvio de seu resíduo é menor.

## 7.4 Sobre a análise de influência

Verificamos que, para os modelos aproximadamente simétricos, a exclusão dos pontos influentes no ajuste dos dados segundo o modelo logístico com resposta de origem (modelo perturbado) não altera significativamente as estimativas obtidas para os coeficientes de regressão do modelo em comparação ao ajuste considerando todas as observações (modelo completo). Comparando com as estimativas obtidas segundo o ajuste do modelo logístico usual, o modelo de regressão logística com resposta de origem ainda apresenta estimativas pontuais e intervalares mais precisas. O mesmo ocorre com o modelo geométrico com resposta de origem.

## 7.5 Sobre o enfoque bayesiano dos modelos propostos

Os modelos de regressão bayesianos com distribuições *a priori* vagas apresentaram estimativas pontuais dos coeficientes de regressão próximas das estimativas segundo o enfoque clássico. Entretanto, a amplitude dos intervalos de credibilidade foi menor em relação à amplitude dos intervalos de confiança para todos os modelos considerados e para todos os tamanhos de amostra. Em outras palavras, a precisão das estimativas foi maior no enfoque bayesiano. Com relação aos modelos em que há parâmetros perturbadores, tais como os modelos normal, lognormal e normal numa estrutura heteroscedástica, houve a necessidade de aumentar o tamanho das cadeias e o tamanho do salto em relação aos demais modelos, pois apresentaram uma autocorrelação maior nas cadeias geradas, considerando todos os parâmetros.

## 7.6 Perspectivas do trabalho e possíveis desdobramentos

Diante da tese apresentada, propomos:

- Aplicar os modelos desenvolvidos a outros conjuntos de dados reais.
- Estender o estudo de simulação para os demais modelos de regressão com resposta normal, log-normal e exponencial.
- Efetuar análises bayesianas de todas as classes abordadas neste trabalho considerando a inferência bayesiana objetiva.

## 8 Apêndice A: Resultados usados para as funções *score* do modelo logístico com respostas diversas

Apresentamos os resultados das funções derivadas de primeira e segunda ordem dos termos da função de log-verossimilhança dos modelos estudados, usadas para a construção das funções *score*.

### 8.1 Resultados usados para o modelo logístico com resposta normal

#### 8.1.1 Derivadas de primeira ordem

Apresentamos as funções derivadas de primeira ordem dos termos da função de log-verossimilhança usadas para a construção das funções *score* do modelo logístico com resposta normal. Para todas as derivadas abaixo considere  $\psi_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ , para  $i = 1, 2, \dots, n$ .

$$\begin{aligned} \frac{\partial}{\partial \sigma} \left\{ -\frac{n}{2} \ln(\sigma^2) \right\} &= -\frac{n}{\sigma}. \\ \frac{\partial}{\partial \sigma} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (r_i - C_R)^2 \right\} &= \frac{1}{\sigma^3} \sum_{i=1}^n (r_i - C_R)^2. \\ \frac{\partial}{\partial \sigma} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) \right\} &= -\frac{1}{\sigma^2} \sum_{i=1}^n \psi_i (r_i - C_R). \\ \frac{\partial [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j} &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij}, \quad j = 0, 1, \dots, p. \\ \frac{\partial \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j} &= \frac{x_{ij}}{f\{\Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}}, \quad j = 0, 1, \dots, p. \\ \frac{\partial}{\partial \beta_j} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) \right\} &= \frac{1}{\sigma} \sum_{i=1}^n \frac{(r_i - C_R)}{f\{\psi_i\}} x_{ij}, \quad j = 0, 1, \dots, p. \\ \frac{\partial}{\partial \beta_j} \left\{ -\frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} &= -\sum_{i=1}^n \frac{\psi_i}{f\{\psi_i\}} x_{ij}, \quad j = 0, 1, \dots, p. \end{aligned}$$

#### 8.1.2 Derivadas de segunda ordem

Apresentamos as funções derivadas de segunda ordem dos termos da função de log-verossimilhança usadas para a construção da matriz de variâncias e covariâncias do modelo logístico com resposta normal. Para todas as derivadas abaixo considere  $\psi_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ , para  $i = 1, 2, \dots, n$ .

$$\begin{aligned}
\frac{\partial^2}{\partial \sigma^2} \left\{ -\frac{n}{2} \ln(\sigma^2) \right\} &= \frac{n}{\sigma^2}. \\
\frac{\partial^2}{\partial \sigma^2} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (r_i - C_R)^2 \right\} &= -\frac{3}{\sigma^4} \sum_{i=1}^n (r_i - C_R)^2. \\
\frac{\partial^2}{\partial \sigma^2} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) \right\} &= \frac{2}{\sigma^3} \sum_{i=1}^n (r_i - C_R)^2. \\
\frac{\partial^2 [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j^2} &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}) [1 - \exp(\mathbf{x}_i^T \boldsymbol{\beta})]}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^3} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2 \Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j^2} &= \frac{-df\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}}{d\beta_j} \frac{x_{ij}^2}{[f\{\Phi^{-1}[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}]^2}. \\
\frac{\partial^2}{\partial \beta_j^2} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) \right\} &= -\frac{1}{\sigma} \sum_{i=1}^n \frac{(r_i - C_R)}{[f\{\psi_i\}]^2} \frac{\partial f\{\psi_i\}}{\partial \beta_j} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \beta_j \partial \beta_k} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) \right\} &= -\frac{1}{\sigma} \sum_{i=1}^n \frac{(r_i - C_R)}{[f\{\psi_i\}]^2} \frac{\partial f\{\psi_i\}}{\partial \beta_j} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \beta_j^2} \left\{ -\frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} &= -\frac{1}{2} \sum_{i=1}^n \left\{ \frac{1 - \psi_i \frac{\partial f\{\psi_i\}}{\partial \beta_j}}{[f\{\psi_i\}]^2} \right\} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \beta_j \partial \beta_k} \left\{ -\frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} &= -\frac{1}{2} \sum_{i=1}^n \left\{ \frac{1 - \psi_i \frac{\partial f\{\psi_i\}}{\partial \beta_j}}{[f\{\psi_i\}]^2} \right\} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \sigma \partial \beta_j} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \psi_i (r_i - C_R) \right\} &= -\frac{1}{\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)}{f\{\psi_i\}} x_{ij}, \quad j = 0, 1, \dots, p.
\end{aligned}$$

## 8.2 Resultados usados para o modelo logístico com resposta exponencial

### 8.2.1 Derivadas de primeira ordem

Apresentamos as funções derivadas de primeira ordem dos termos da função de log-verossimilhança usadas para a construção das funções *score* do modelo logístico com resposta exponencial.

$$\begin{aligned}
\frac{\partial \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j} &= \frac{x_{ij}}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \quad j = 0, 1, \dots, p. \\
\frac{\partial \ln \{-\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}}{\partial \beta_j} &= \frac{x_{ij}}{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]}, \quad j = 0, 1, \dots, p.
\end{aligned}$$

### 8.2.2 Funções derivadas de segunda ordem

Apresentamos as funções derivadas de segunda ordem dos termos da função de log-verossimilhança usadas para a construção da matriz de variâncias e covariâncias do modelo logístico com resposta exponencial.

$$\begin{aligned}
\frac{\partial^2 \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j^2} &= -\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2 \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j \partial \beta_k} &= -\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2 \ln \{-\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}}{\partial \beta_j^2} &= \frac{\{1 + \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}}{\{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]\}^2} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2 \ln \{-\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]\}}{\partial \beta_j \partial \beta_k} &= \frac{\{1 + \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}}{\{\ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]\}^2} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p.
\end{aligned}$$

### 8.3 Resultados usados para o modelo logístico com resposta geométrica

#### 8.3.1 Derivadas de primeira ordem

Apresentamos as funções derivadas de primeira ordem dos termos da função de log-verossimilhança usadas para a construção das funções *score* do modelo logístico com resposta geométrica.

$$\begin{aligned}
\frac{\partial \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j} &= \frac{x_{ij}}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \quad j = 0, 1, \dots, p. \\
\frac{\partial \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{\partial \beta_j} &= -\frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{(C_R + 1) [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\}} x_{ij},
\end{aligned}$$

para  $j = 0, 1, \dots, p$ .

#### 8.3.2 Derivadas de segunda ordem

Apresentamos as funções derivadas de segunda ordem dos termos da função de log-verossimilhança usadas para a construção da matriz de variâncias e covariâncias do modelo logístico com resposta geométrica.

$$\begin{aligned}
\frac{\partial^2 \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j^2} &= -\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2 \ln [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j \partial \beta_k} &= -\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2 \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{\partial \beta_j^2} &= \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\} \psi_i}{(C_R + 1)^2 [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2 \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\}^2} x_{ij}^2, \\
\frac{\partial^2 \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}}{\partial \beta_j \partial \beta_k} &= \frac{[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}} \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\} \psi_i}{(C_R + 1)^2 [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2 \left\{1 - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{\frac{1}{C_R+1}}\right\}^2} x_{ij} x_{ik},
\end{aligned}$$

para  $j = 0, 1, \dots, p$ , em que  $\psi_i = (C_R + 1) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]^{C_R+1}$ ,  $i = 1, 2, \dots, n$ .

## 8.4 Resultados usados para o modelo logístico com resposta Poisson

### 8.4.1 Derivadas de primeira ordem

Apresentamos as funções derivadas de primeira ordem dos termos da função de log-verossimilhança usadas para a construção das funções *score* do modelo logístico com resposta Poisson.

$$\begin{aligned} \frac{\partial \ln \{ -\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \}}{\partial \beta_j} &= \frac{g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})}{\ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]} x_{ij}, \quad j = 0, 1, \dots, p. \\ \frac{\partial \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j} &= -[g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})], \quad j = 0, 1, \dots, p. \end{aligned}$$

### 8.4.2 Derivadas de segunda ordem

Apresentamos as funções derivadas de segunda ordem dos termos da função de log-verossimilhança usadas para a construção da matriz de variâncias e covariâncias do modelo logístico com resposta Poisson.

$$\begin{aligned} \frac{\partial^2 \ln \{ -\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \}}{\partial \beta_j^2} &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} \frac{\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) \}}{\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \}^2} x_{ij}^2, \\ \frac{\partial^2 \ln \{ -\ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})] \}}{\partial \beta_j \partial \beta_k} &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} \frac{\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) \}}{\{ \ln [1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})] \}^2} x_{ij} x_{ik}, \\ \frac{\partial^2 \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j^2} &= -\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij}^2, \quad j = 0, 1, \dots, p. \\ \frac{\partial^2 \ln [1 - g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]}{\partial \beta_j \partial \beta_k} &= -\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]^2} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \end{aligned}$$

## 8.5 Resultados usados para o modelo logístico com resposta Normal em uma estrutura de heteroscedasticidade multiplicativa

### 8.5.1 Derivadas de primeira ordem

Apresentamos as funções derivadas de primeira ordem dos termos da função de log-verossimilhança usadas para a construção das funções *score* do modelo logístico com resposta Normal em uma estrutura de heteroscedasticidade multiplicativa. Para todas as derivadas abaixo considere  $\psi_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ , para  $i = 1, 2, \dots, n$ .

$$\begin{aligned}
\frac{\partial}{\partial \sigma} \left\{ -\frac{n}{2} \ln(\sigma^2) \right\} &= -\frac{n}{\sigma}. \\
\frac{\partial}{\partial \sigma} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \right\} &= \frac{1}{\sigma^3} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \\
\frac{\partial}{\partial \sigma} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= -\frac{1}{\sigma^2} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \\
\frac{\partial}{\partial \lambda} \left\{ -\frac{\lambda}{2} \sum_{i=1}^n \ln(x_i) \right\} &= -\frac{1}{2} \sum_{i=1}^n \ln(x_i). \\
\frac{\partial}{\partial \lambda} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \right\} &= \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2 \ln x_i}{x_i^\lambda} \\
\frac{\partial}{\partial \lambda} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= -\frac{1}{2\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R) \ln x_i}{x_i^{\lambda/2}} \\
\frac{\partial}{\partial \beta_j} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= \frac{1}{\sigma} \sum_{i=1}^n \frac{(r_i - C_R)}{x_i^{\lambda/2} f\{\psi_i\}} x_{ij}, \quad j = 0, 1, \dots, p. \\
\frac{\partial}{\partial \beta_j} \left\{ -\frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} &= -\sum_{i=1}^n \frac{\psi_i}{f\{\psi_i\}} x_{ij}, \quad j = 0, 1, \dots, p.
\end{aligned}$$

### 8.5.2 Derivadas de segunda ordem

Apresentamos as funções derivadas de segunda ordem dos termos da função de log-verossimilhança usadas para a construção da matriz de variâncias e covariâncias do modelo logístico com resposta Normal em uma estrutura de heteroscedasticidade multiplicativa. Para todas as derivadas abaixo considere  $\psi_i = \Phi^{-1} [g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})]$ , para  $i = 1, 2, \dots, n$ .

$$\begin{aligned}
\frac{\partial^2}{\partial \sigma^2} \left\{ -\frac{n}{2} \ln(\sigma^2) \right\} &= \frac{n}{\sigma^2}. \\
\frac{\partial^2}{\partial \sigma^2} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \right\} &= -\frac{3}{\sigma^4} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \\
\frac{\partial^2}{\partial \sigma^2} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= \frac{2}{\sigma^3} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)^2}{x_i^{\lambda/2}} \\
\frac{\partial^2}{\partial \lambda^2} \left\{ -\frac{\lambda}{2} \sum_{i=1}^n \ln(x_i) \right\} &= 0. \\
\frac{\partial^2}{\partial \lambda^2} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \right\} &= -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2 [\ln(x_i)]^2}{x_i^\lambda} \\
\frac{\partial^2}{\partial \lambda^2} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= \frac{1}{4\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)^2 [\ln(x_i)]^2}{x_i^{\lambda/2}} \\
\frac{\partial^2}{\partial \beta_j^2} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= -\frac{1}{\sigma} \sum_{i=1}^n \frac{(r_i - C_R)}{[f\{\psi_i\}]^2} \frac{\partial f\{\psi_i\}}{\partial \beta_j} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \beta_j \partial \beta_k} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{\lambda/2} \right\} &= -\frac{1}{\sigma} \sum_{i=1}^n \frac{(r_i - C_R)}{[f\{\psi_i\}]^2} \frac{\partial f\{\psi_i\}}{\partial \beta_j} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \beta_j^2} \left\{ -\frac{1}{2} \sum_{i=1}^n \psi_i^2 \right\} &= -\frac{1}{2} \sum_{i=1}^n \left\{ \frac{1 - \psi_i \frac{\partial f\{\psi_i\}}{\partial \beta_j}}{[f\{\psi_i\}]^2} \right\} x_{ij}^2, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \sigma \partial \lambda} \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2}{x_i^\lambda} \right\} &= \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)^2 \ln(x_i)}{x_i^\lambda} \\
\frac{\partial^2}{\partial \sigma \partial \lambda} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{\lambda/2} \right\} &= -\frac{1}{\sigma^3} \sum_{i=1}^n \frac{(r_i - C_R) \psi_i (\ln x_i)}{x_i^{\lambda/2}}. \\
\frac{\partial^2}{\partial \sigma \partial \beta_j} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= \frac{1}{\sigma^2} \sum_{i=1}^n \frac{(r_i - C_R)}{[f\{\psi_i\}]^2} \frac{\partial f\{\psi_i\}}{\partial \beta_j} x_{ij} x_{ik}, \quad j = 0, 1, \dots, p. \\
\frac{\partial^2}{\partial \lambda \partial \beta_j} \left\{ \frac{1}{\sigma} \sum_{i=1}^n \frac{\psi_i (r_i - C_R)}{x_i^{\lambda/2}} \right\} &= -\frac{1}{2\sigma} \sum_{i=1}^n \frac{(r_i - C) \ln(x_i)}{x_i^{\lambda/2} f\{\psi_i\}} x_{ij}, \quad j = 0, 1, \dots, p.
\end{aligned}$$

# 9 Apêndice B: Histogramas das estimativas bayesianas

## 9.1 Modelo logístico com resposta normal

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão e para o parâmetro perturbador  $\sigma^2$ , do modelo logístico com resposta normal, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

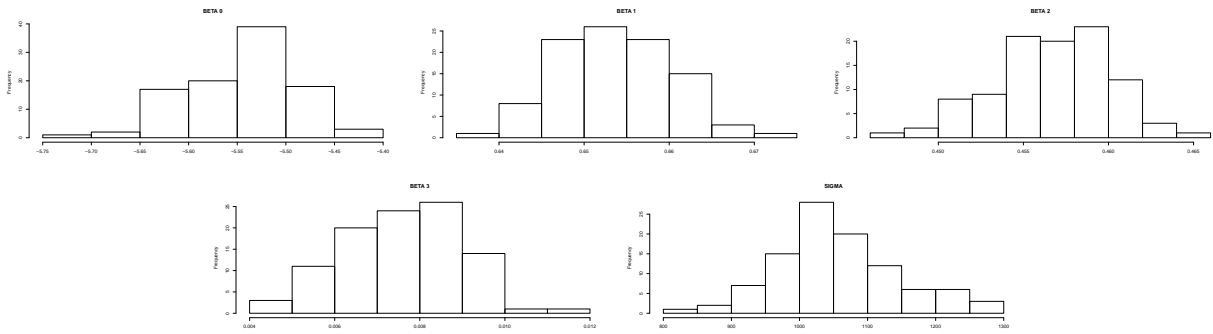


Figura 9.1: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 50$ .

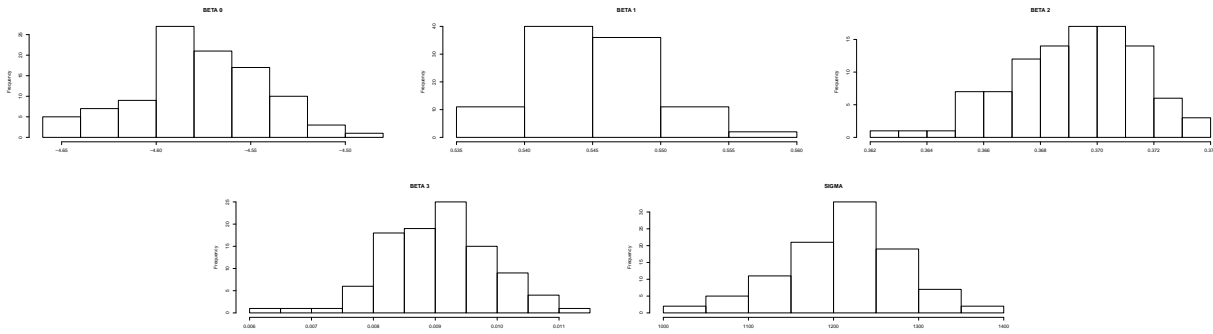


Figura 9.2: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 100$ .

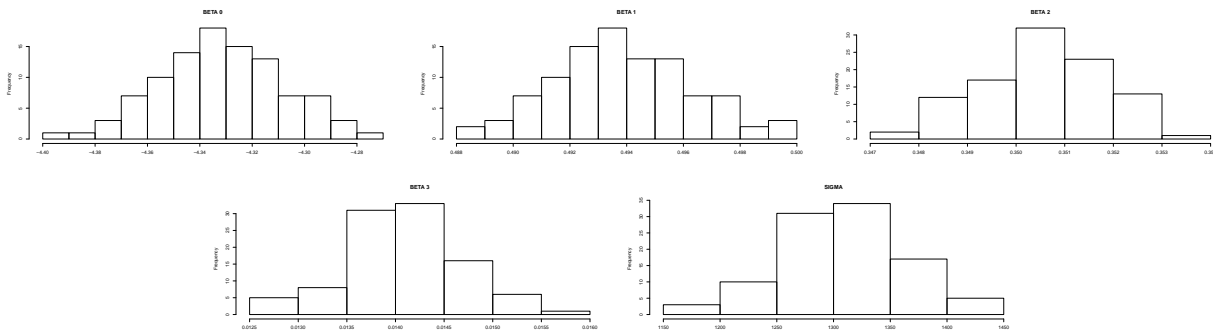


Figura 9.3: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 200$ .



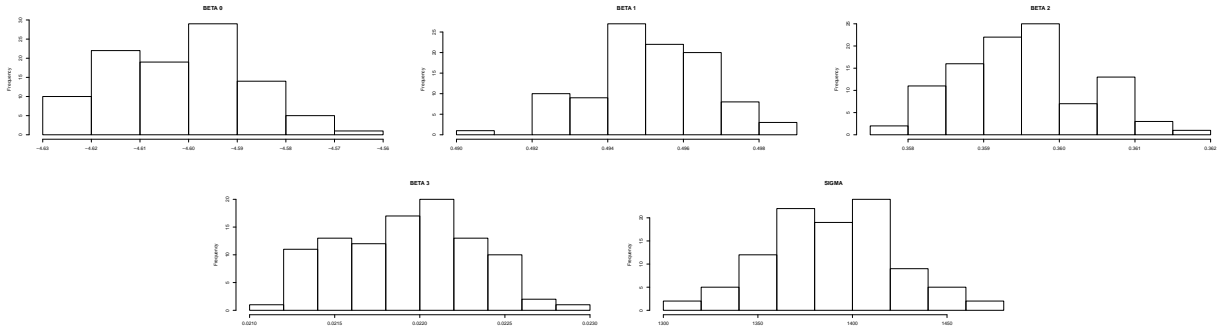


Figura 9.4: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 500$ .

## 9.2 Modelo logístico com resposta lognormal

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão e para o parâmetro perturbador  $\sigma^2$ , do modelo logístico com resposta lognormal, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

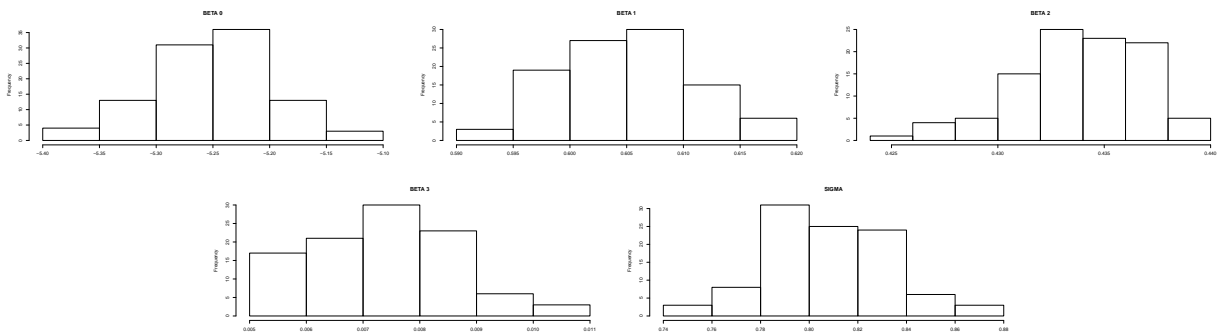


Figura 9.5: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 50$ .

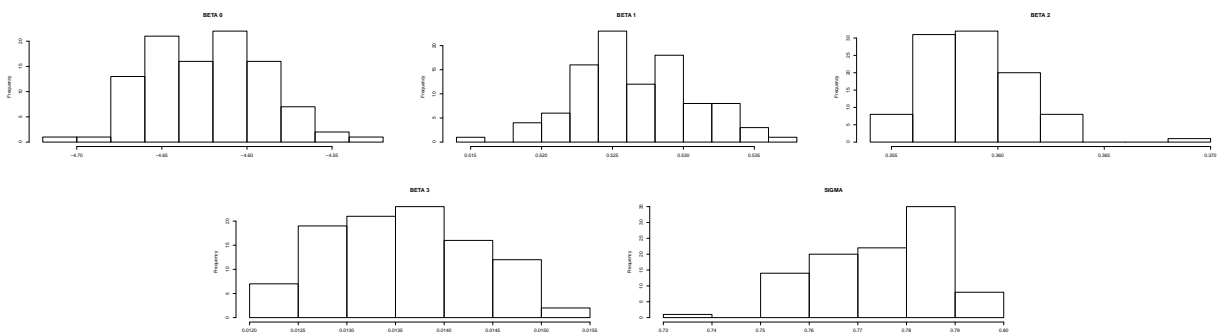


Figura 9.6: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 100$ .

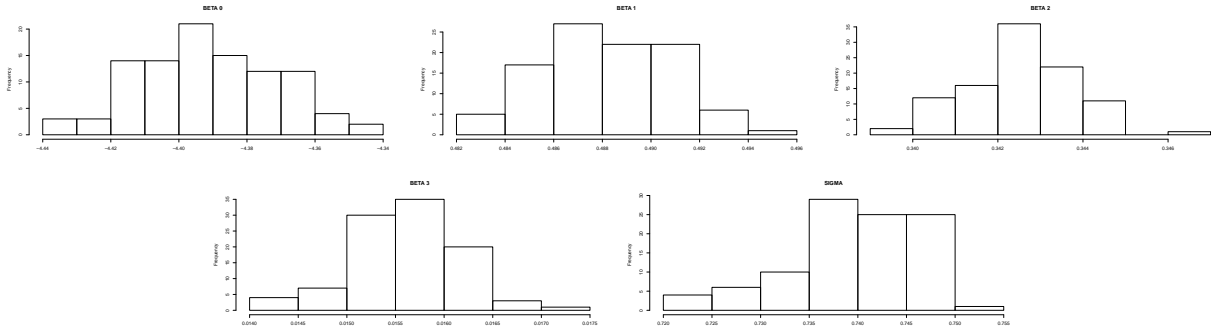


Figura 9.7: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 200$ .

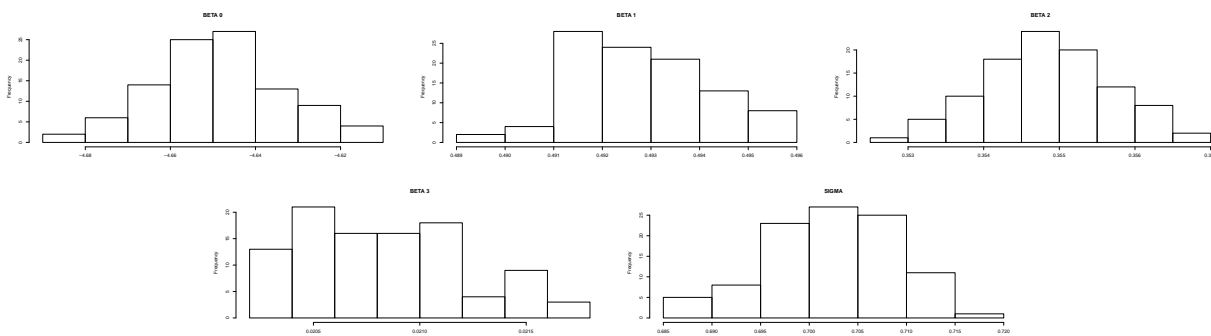


Figura 9.8: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 500$ .

### 9.3 Modelo logístico com resposta exponencial

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão do modelo logístico com resposta exponencial, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

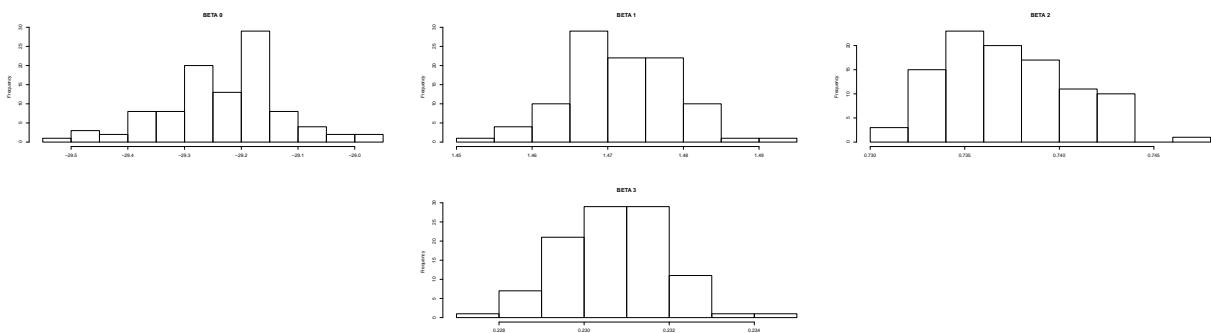


Figura 9.9: Histograma das estimativas dos coeficientes de regressão,  $n = 50$ .

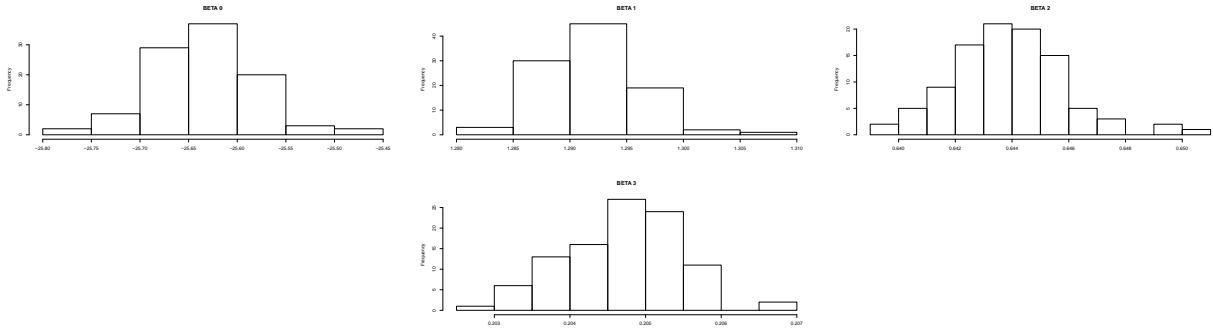


Figura 9.10: Histograma das estimativas dos coeficientes de regressão,  $n = 100$ .

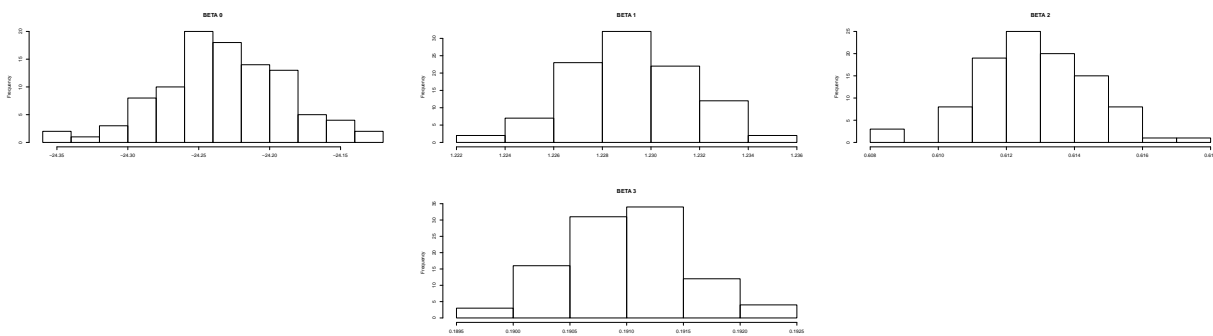


Figura 9.11: Histograma das estimativas dos coeficientes de regressão,  $n = 200$ .

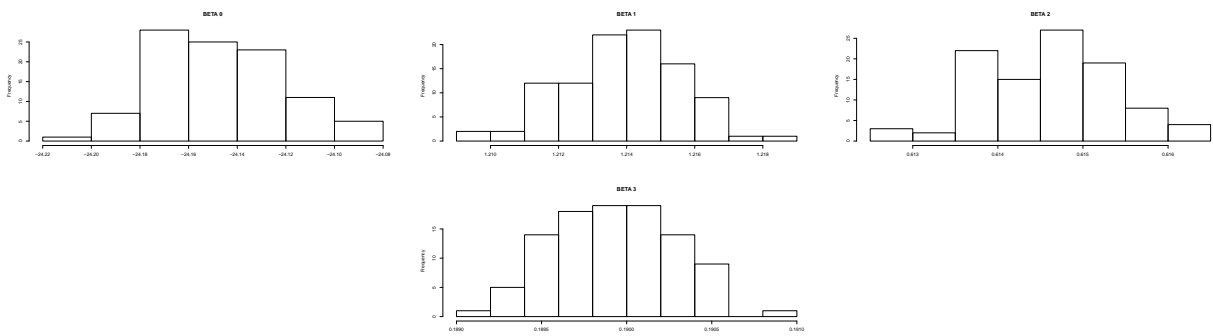


Figura 9.12: Histograma das estimativas dos coeficientes de regressão,  $n = 500$ .

## 9.4 Modelo logístico com resposta geométrica

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão do modelo logístico com resposta geométrica, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

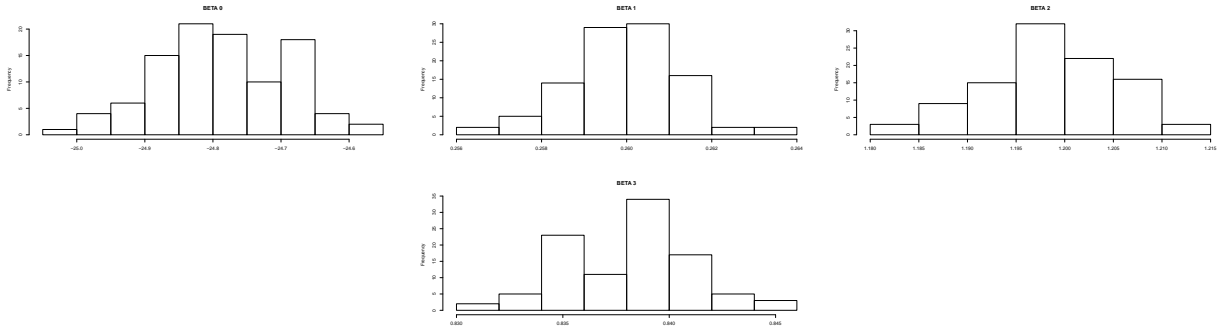


Figura 9.13: Histograma das estimativas dos coeficientes de regressão,  $n = 50$ .

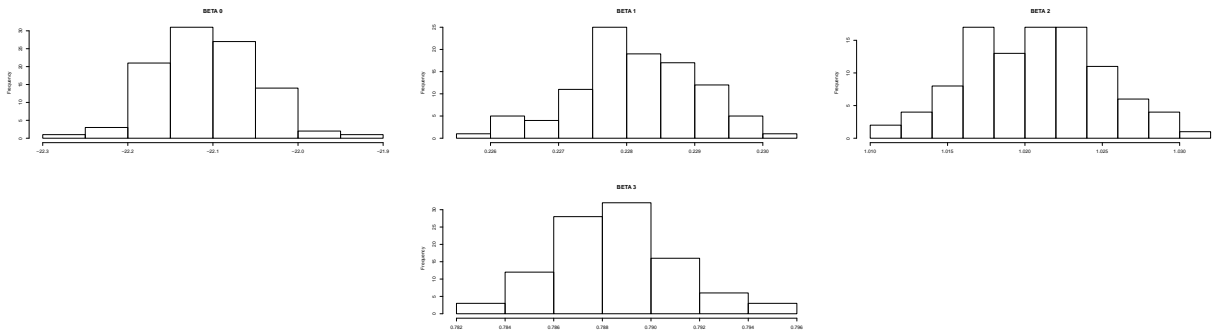


Figura 9.14: Histograma das estimativas dos coeficientes de regressão,  $n = 100$ .

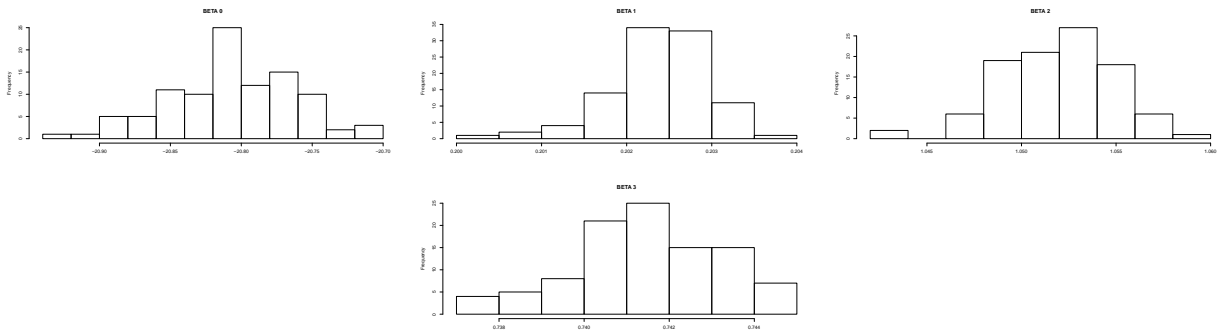


Figura 9.15: Histograma das estimativas dos coeficientes de regressão,  $n = 200$ .

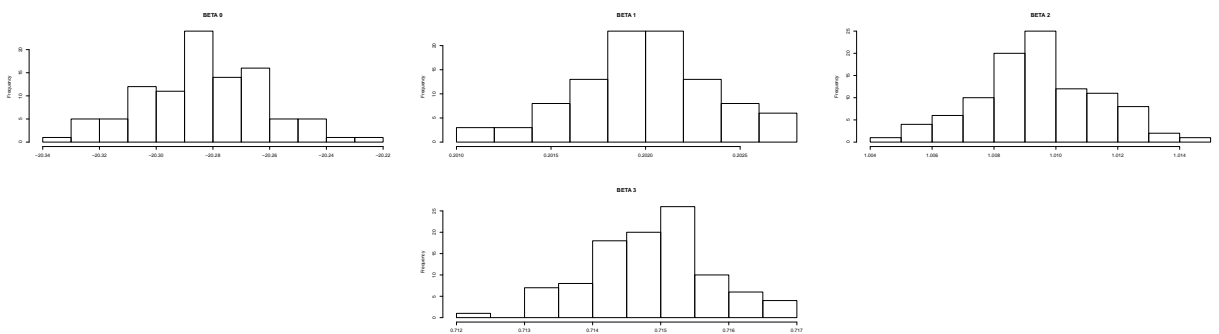


Figura 9.16: Histograma das estimativas dos coeficientes de regressão,  $n = 500$ .

## 9.5 Modelo logístico com resposta Poisson

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão do modelo logístico com resposta Poisson, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

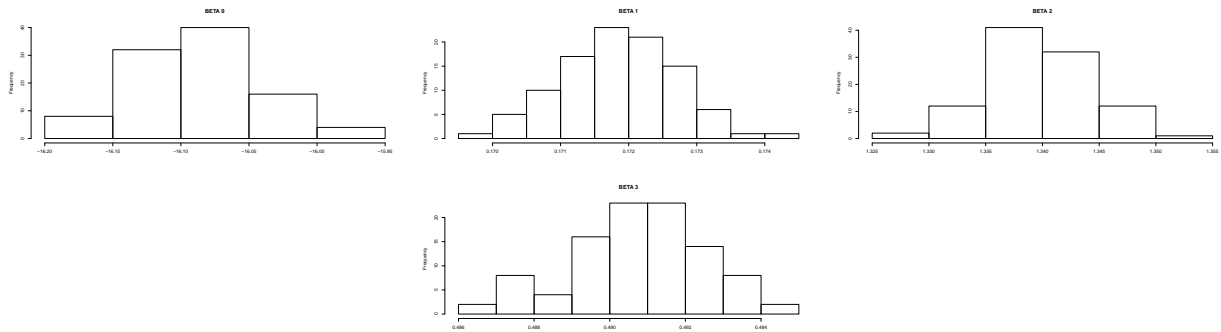


Figura 9.17: Histograma das estimativas dos coeficientes de regressão,  $n = 50$ .

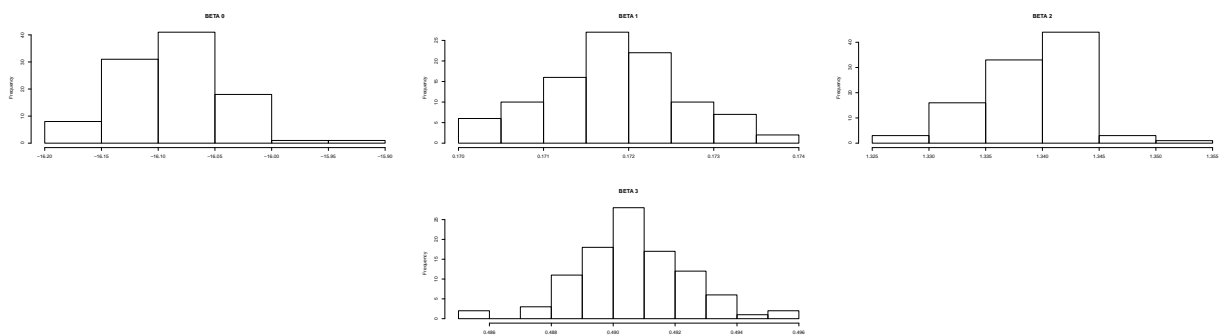


Figura 9.18: Histograma das estimativas dos coeficientes de regressão,  $n = 100$ .

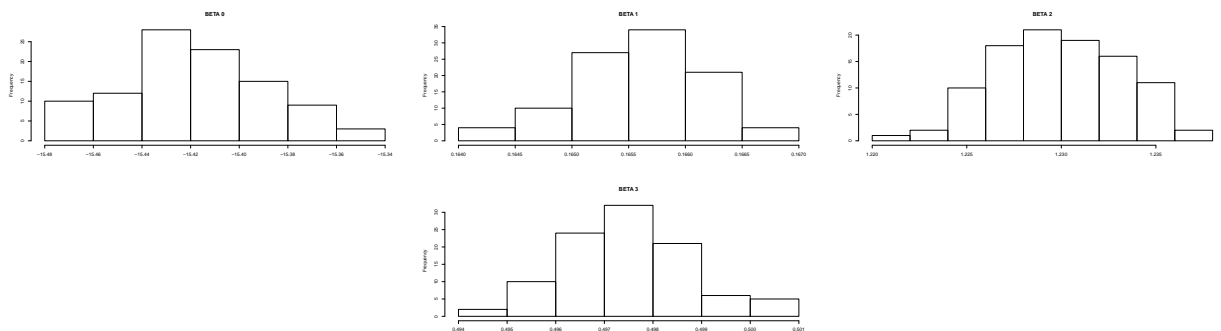


Figura 9.19: Histograma das estimativas dos coeficientes de regressão,  $n = 200$ .

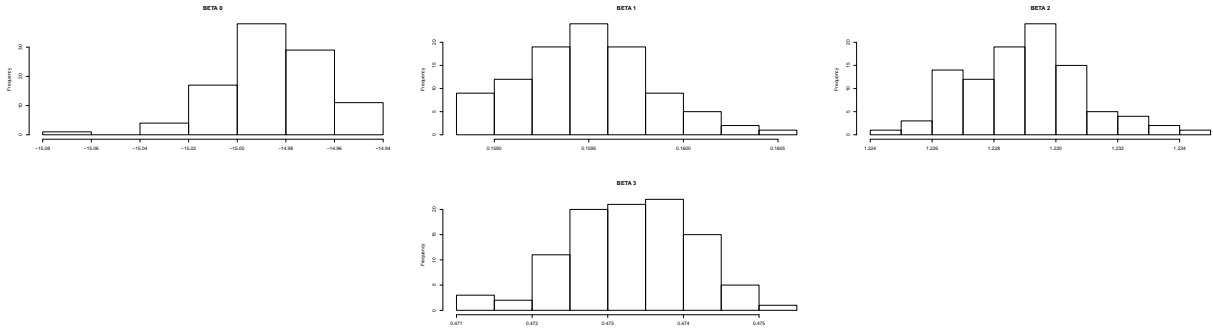


Figura 9.20: Histograma das estimativas dos coeficientes de regressão,  $n = 500$ .

## 9.6 Modelo geométrico com resposta normal

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão e para o parâmetro perturbador  $\sigma^2$ , do modelo geométrico com resposta normal, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

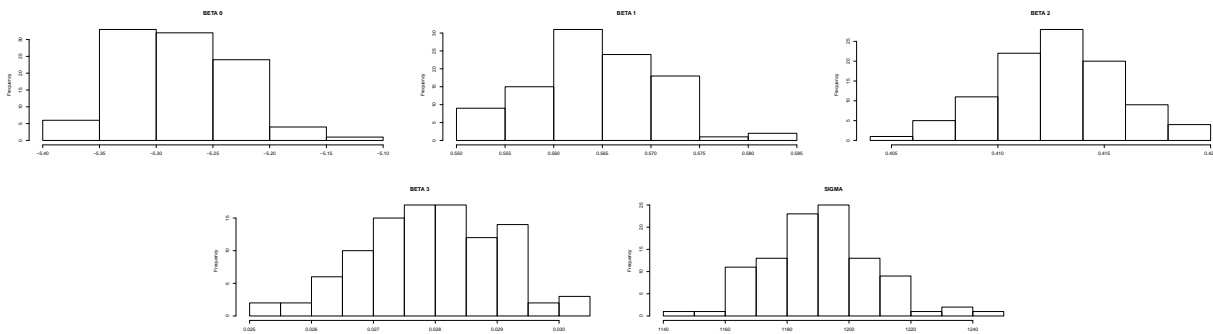


Figura 9.21: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 50$ .

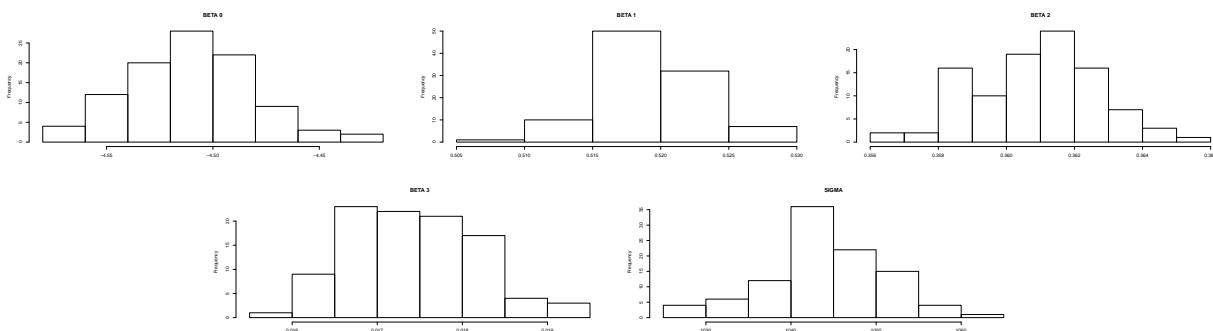


Figura 9.22: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 100$ .

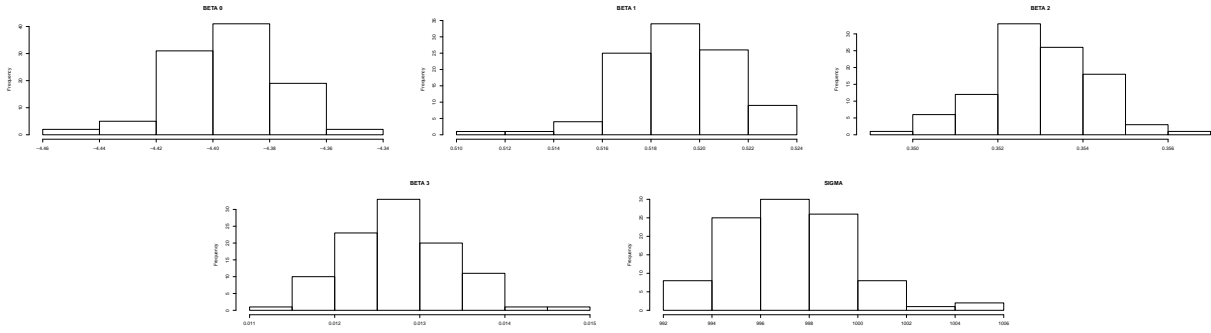


Figura 9.23: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 200$ .

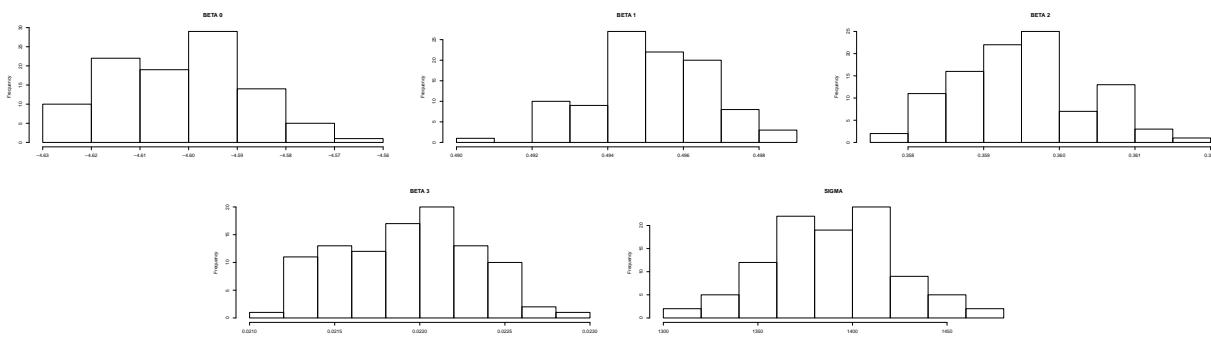


Figura 9.24: Histograma das estimativas dos coeficientes de regressão e de  $\sigma^2$ ,  $n = 500$ .

## 9.7 Modelo geométrico com resposta exponencial

As figuras a seguir mostram os histogramas das estimativas bayesianas dos coeficientes de regressão do modelo geométrico com resposta exponencial, considerando as 100 amostras de tamanho  $n$  geradas no capítulo 5.

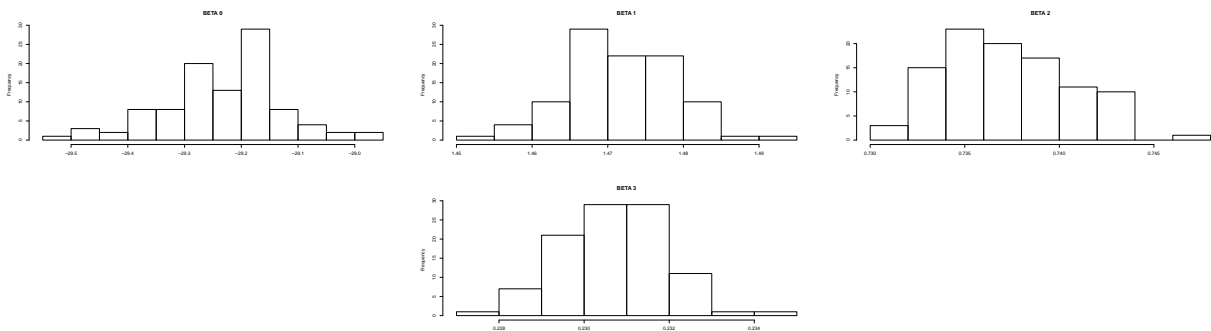


Figura 9.25: Histograma das estimativas dos coeficientes de regressão,  $n = 50$ .

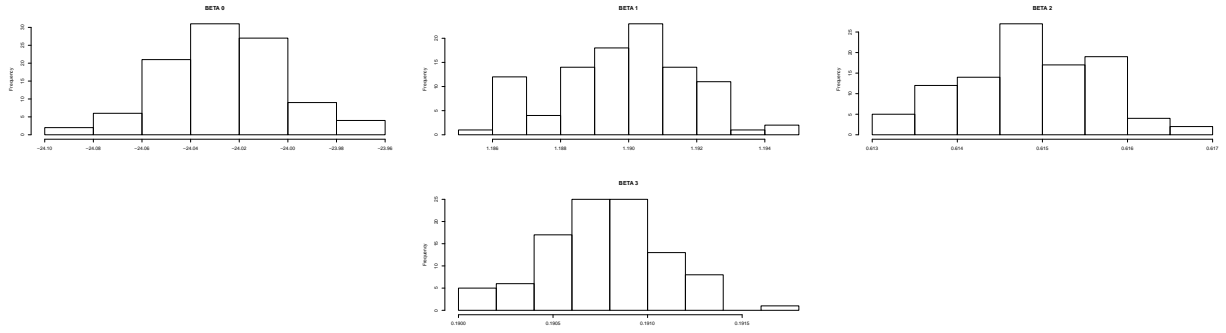


Figura 9.26: Histograma das estimativas dos coeficientes de regressão,  $n = 100$ .

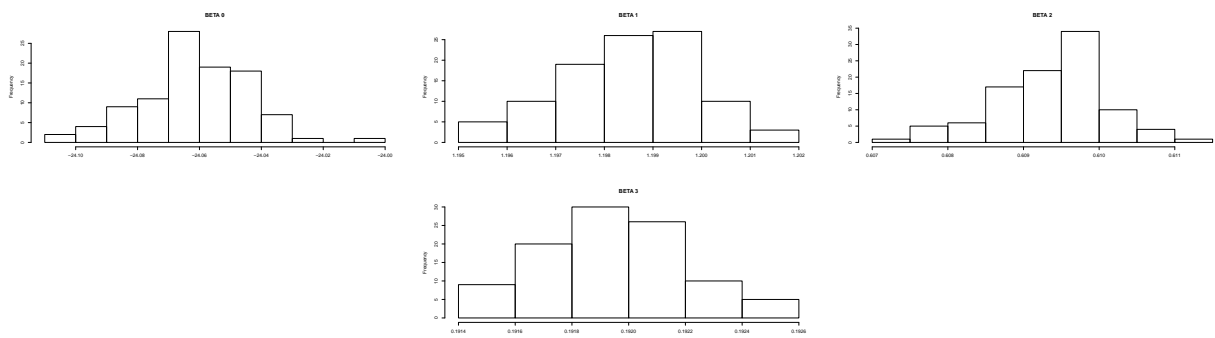


Figura 9.27: Histograma das estimativas dos coeficientes de regressão,  $n = 200$ .

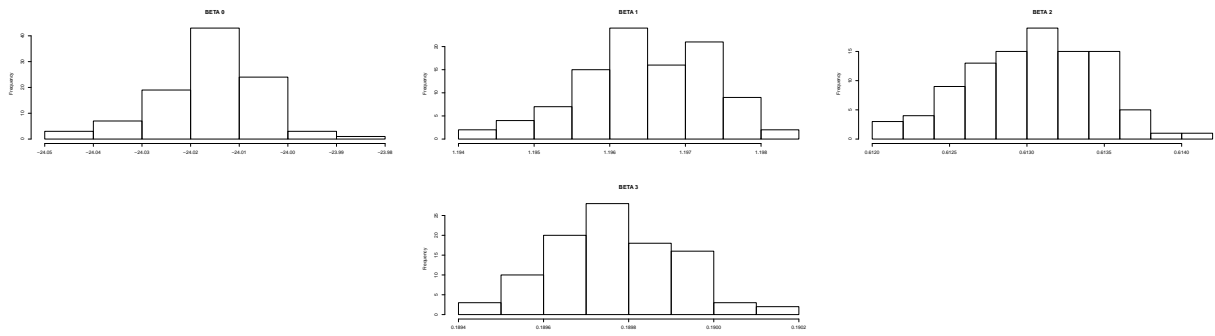


Figura 9.28: Histograma das estimativas dos coeficientes de regressão,  $n = 500$ .



# 10 Apêndice C: Programas desenvolvidos

## Modelo logístico com resposta normal heteroscedástica multiplicativa

```
Sem título
#####
##### GERAÇÃO DOS DADOS - TRÊS VARIÁVEIS DE ENTRADA
#####
set.seed(2010)
x1 <- x2 <- x3 <- mi <- mi.vetor <- vetor.y <- vetor.yc <- yc <- y <- quantil <- resposta.logistica <- sigma2i <- sequencia <- numeric()
n.obs <- 50 # tamanho n do conjunto de dados
n.par.usual <- 4 # número de coeficientes beta
n.par.origem <- 6 # número de parâmetros do modelo logístico com resposta de origem
sigma2 <- 4000 # Variância de Y
sigma <- sqrt(sigma2)
C <- 10000 # Ponto de Corte
lambda <- 3
n.magico <- 10000
#####
# VALORES VERDADEIROS DOS PARÂMETROS (COEFICIENTES)
#####
beta.0 <- -32
beta.1 <- 0.77
beta.2 <- 2.20
beta.3 <- 1.00
#####
### AMOSTRA TOTAL
#####
for( i in 1 : n.obs ) {
  x1[i] <- rnorm(1, 10, 2.0) # renda mensal em SM
  x2[i] <- rnorm(1, 5.0, 1.0) # número de eletro-eletrônicos
  x3[i] <- rnorm(1, 15, 2.0) # idade do indivíduo
  sigma2i[i] <- sigma2*x1[i]^lambda
  quantil[i] <- qnorm(exp(beta.0+beta.1*x1[i]+beta.2*x2[i]+beta.3*x3[i])/
    (1+exp(beta.0+beta.1*x1[i]+beta.2*x2[i]+beta.3*x3[i])))
}
for( i in 1 : n.obs ) {
  mi[i] <- (sqrt(sigma2)*(x1[i]^(lambda/2))*quantil[i])+C
  mi.vetor <- rep(mi, n.magico)
  yc <- round(rnorm(length(mi.vetor), mi.vetor, sqrt(sigma2i)),0)
  y <- ifelse(yc > C, y <- 1, y <- 0)
  matriz.mi <- matrix(mi.vetor, nrow = n.obs, ncol = n.magico)
  matriz.yc <- matrix(yc, nrow = n.obs, ncol = n.magico)
  matriz.y <- matrix(c(y), nrow = n.obs, ncol = n.magico)
  vetor.y <- matriz.y
  estimativas.usuais <- matrix(NA, nrow = n.magico, ncol = n.par.usual)
  matriz.erro.padrao.usual <- matrix(NA, nrow = n.magico, ncol = n.par.usual)
  matriz.X <- matrix(NA, nrow = n.obs, ncol = n.par.usual)
  matriz.X[,1] <- 1
  matriz.X[,2] <- x1
  matriz.X[,3] <- x2
  matriz.X[,4] <- x3
#####
### ACURÁCIA DO MODELO USUAL E PROBABILIDADE DE COBERTURA
#####
cobertura.beta0.usual.90 <- cobertura.beta1.usual.90 <- cobertura.beta2.usual.90 <- cobertura.beta3.usual.90 <- numeric()
cobertura.beta0.usual.95 <- cobertura.beta1.usual.95 <- cobertura.beta2.usual.95 <- cobertura.beta3.usual.95 <- numeric()
cobertura.beta0.usual.99 <- cobertura.beta1.usual.99 <- cobertura.beta2.usual.99 <- cobertura.beta3.usual.99 <- numeric()
IC.usual.90 <- matrix(NA, nrow = n.magico, ncol = 8)
IC.usual.95 <- matrix(NA, nrow = n.magico, ncol = 8)
IC.usual.99 <- matrix(NA, nrow = n.magico, ncol = 8)
pi.estimado.usual <- matrix(NA, nrow = n.magico, ncol = n.obs)
for( i in 1 : n.magico ) {
  estimativas.usuais[i,] <- (glm(matriz.y[,i] ~ x1 + x2 + x3, family = binomial)$coefficients)
  matriz.erro.padrao.usual[i,] <- summary(glm(matriz.y[,i] ~ x1 + x2 + x3, family = binomial))$coefficients[,2]
  pi.estimado.usual[i,] <- exp(matriz.X*estimativas.usuais[i,])/(1+exp(matriz.X*estimativas.usuais[i,]))
  sequencia[i] <- ifelse(min(pi.estimado.usual[i,]) > 0 & max(pi.estimado.usual[i,]) < 1, sequencia[i] <- i, sequencia[i]
<- NA)
  IC.usual.90[i,] <- round(confint(glm(matriz.y[,i] ~ x1 + x2 + x3, family = binomial),level = 0.90),4)
  IC.usual.95[i,] <- round(confint(glm(matriz.y[,i] ~ x1 + x2 + x3, family = binomial),level = 0.95),4)
  IC.usual.99[i,] <- round(confint(glm(matriz.y[,i] ~ x1 + x2 + x3, family = binomial),level = 0.99),4)
}
cat("\n", "Amostra", i, "Estimativas Modelo Logístico Usual:", round(estimativas.usuais[i,],4))
#####
##### MAXIMIZAÇÃO RESPOSTA NORMAL - ALGORÍTMO BFGS
#####
### ACURÁCIA DO MODELO DE ORIGEM E PROBABILIDADE DE COBERTURA
#####
beta0.estimado.usual <- round(mean(estimativas.usuais[sequencia,1], na.rm = TRUE),4) # valor inicial para beta 0 usando o BFGS
beta1.estimado.usual <- round(mean(estimativas.usuais[sequencia,2], na.rm = TRUE),4) # valor inicial para beta 1 usando o BFGS
beta2.estimado.usual <- round(mean(estimativas.usuais[sequencia,3], na.rm = TRUE),4) # valor inicial para beta 2 usando o BFGS
beta3.estimado.usual <- round(mean(estimativas.usuais[sequencia,4], na.rm = TRUE),4) # valor inicial para beta 3 usando o BFGS

cobertura.beta0.origem.90 <- cobertura.beta1.origem.90 <- cobertura.beta2.origem.90 <- cobertura.beta3.origem.90 <- cobertura.sigma.origem.90
<- cobertura.lambda.origem.90 <- numeric()
cobertura.beta0.origem.95 <- cobertura.beta1.origem.95 <- cobertura.beta2.origem.95 <- cobertura.beta3.origem.95 <- cobertura.sigma.origem.95
<- cobertura.lambda.origem.95 <- numeric()
cobertura.beta0.origem.99 <- cobertura.beta1.origem.99 <- cobertura.beta2.origem.99 <- cobertura.beta3.origem.99 <- cobertura.sigma.origem.99
<- cobertura.lambda.origem.99 <- numeric()
beta <- matrix(c(beta0.estimado.usual, beta1.estimado.usual, beta2.estimado.usual, beta3.estimado.usual, sqrt(sigma2), 3),
nrow = n.par.origem, ncol = 1) #chutes iniciais para os betas
p <- numeric()
estimativas.BFGS <- matrix(NA, nrow = n.magico, ncol = n.par.origem)
matriz.erro.padrao.origem <- matrix(NA, nrow = n.magico, ncol = n.par.origem)
p[1] <- beta[1,] # valor inicial para beta0
p[2] <- beta[2,] # valor inicial para beta1
p[3] <- beta[3,] # valor inicial para beta2
p[4] <- beta[4,] # valor inicial para beta3
p[5] <- beta[5,] # valor inicial para sigma
p[6] <- beta[6,] # valor inicial para lambda
for(i in 1:n.magico) {
  lv <- function(p) {
    ((n.obs/2)*log(p[5]^2))+
    ((p[6]/2)*sum(log(x1)))+
    ((1/(2*p[5]^2))*sum(((matriz.yc[,i]-p[5]*(x1^(p[6]/2))*qnorm((exp(p[1]+p[2]*x1+p[3]*x2+p[4]*x3))/(1+exp(p[1]+p[2]*x1+p[3]*x2+p[4]*x3)))+C))^2)/(x1*p[6]))
  }
  estimativas.BFGS[i,] <- c(optim(p=c(p), lv, gr = NULL, method = c("BFGS"), hessian = TRUE)$par)
  matriz.erro.padrao.origem[i,] <- c(sqrt(diag(abs(solve(optim(p=c(p), lv, gr = NULL, method = c("BFGS"), hessian = TRUE)$hessian))))))
}
cat("\n", "Amostra", i, "Estimativas Modelo Origem:", round(estimativas.BFGS[i,1],4), round(estimativas.BFGS[i,2],4),
round(estimativas.BFGS[i,3],4),
round(estimativas.BFGS[i,4],4), round(estimativas.BFGS[i,5],0), round(estimativas.BFGS[i,6],4))
}
```

# Modelo logístico bayesiano com resposta geométrica

```
Sem título
#####
##### GERAÇÃO DOS DADOS - TRÊS VARIÁVEIS DE ENTRADA
#####
set.seed(2014)
x1 <- x2 <- x3 <- ginv <- y.geometrico <- y.logistico <- prob <- ponto.corte <- numeric()
pi.estimado <- conf <- sequencia <- resposta.logistica <- percent <- numeric()
n.obs <- 500 # tamanho n do conjunto de dados
n.par.usual <- 4 # número de parâmetros do modelo logístico usual
n.par.origem <- n.par.usual # número de parâmetros do modelo logístico com resposta de origem
Cr <- 0 # Ponto de Corte
n.magico <- 100 # número de geração de amostras
### VALORES VERDADEIROS DOS PARÂMETROS (COEFICIENTES)
#####
beta.0 <- -20
beta.1 <- 0.2
beta.2 <- 1.0
beta.3 <- 0.7
### GERAÇÃO DAS AMOSTRAS - PROPRIEDADES FREQUENTISTAS
#####
for( i in 1 : n.obs ) {
  x1[i] <- round(rnorm(1, 35, 5),0) # Idade do cliente (em anos)
  x2[i] <- round(rnorm(1, 05, 1),0) # Renda salarial do cliente (em salários mínimos)
  x3[i] <- round(rnorm(1, 10, 2),0) # Número de eletro-eletrônicos na residência
  ginv[i] <- exp(beta.0 + beta.1*x1[i] + beta.2*x2[i] + beta.3*x3[i])/
    (1+exp(beta.0 + beta.1*x1[i] + beta.2*x2[i] + beta.3*x3[i]))
  prob[i] <- 1 - (ginv[i]^(1/(Cr+1)))
}
prob.vetor <- rep(prob, n.magico)
r <- round(rgeom(length(prob.vetor), prob),0)
y <- ifelse(r > Cr, y <- 1, y <- 0)
matriz.r <- matrix(c(r),nrow = n.obs, ncol = n.magico)
matriz.y <- matrix(c(y),nrow = n.obs, ncol = n.magico)
matriz.X <- matrix(NA, nrow = n.obs, ncol = n.par.usual)
matriz.X[,1] <- x1
matriz.X[,2] <- x2
matriz.X[,3] <- x3
matriz.X[,4] <- x3
beta0.C1 <- beta1.C1 <- beta2.C1 <- beta3.C1 <- numeric()
#####
loop <- 120000 # Tamanho da cadeia
burnin <- 20000 # Trecho inicial a ser descartado
salto <- 50 # Salto da cadeia
for( i in 1 : n.magico ) {
  for( j in 2 : loop ) {
##### CADEIA 1 #####
#####
beta0.C1[i] <- beta0.estimativa.usual #beta[1] # Chute inicial para o parâmetro beta 0
beta1.C1[i] <- beta1.estimativa.usual #beta[2] # Chute inicial para o parâmetro beta 1
beta2.C1[i] <- beta2.estimativa.usual #beta[3] # Chute inicial para o parâmetro beta 2
beta3.C1[i] <- beta3.estimativa.usual #beta[4] # Chute inicial para o parâmetro beta 3
hiperparametro.a0 <- 0 # Hiperparâmetro : média da priori de beta 0
hiperparametro.a1 <- 0 # Hiperparâmetro : média da priori de beta 1
hiperparametro.a2 <- 0 # Hiperparâmetro : média da priori de beta 2
hiperparametro.a3 <- 0 # Hiperparâmetro : média da priori de beta 3
hiperparametro.b0 <- 100 # Hiperparâmetro : desvio-padrão da priori de beta 0
hiperparametro.b1 <- 100 # Hiperparâmetro : desvio-padrão da priori de beta 1
hiperparametro.b2 <- 100 # Hiperparâmetro : desvio-padrão da priori de beta 2
hiperparametro.b3 <- 100 # Hiperparâmetro : desvio-padrão da priori de beta 3
### CONDICIONAL DE BETA 0 #####
teste.b0 <- rnorm(1, hiperparametro.a0, hiperparametro.b0)
B1 <- sum(log(1-(exp(beta0.teste+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.teste+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.teste+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))
+((beta0.teste/(Cr+1))*sum(matriz.r[,i])))
B2 <- sum(log(1-(exp(beta0.C1[j-1]+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.C1[j-1]+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j-1]+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))
+((beta0.C1[j-1]/(Cr+1))*sum(matriz.r[,i])))
testeB <- exp(B1-B2)
u <- runif(1)
#ifelse(u < min(1, testeB), beta0.C1[j] <- beta0.teste, beta0.C1[j] <- beta0.C1[j-1])
#ifelse(u < min(1, testeB), beta0.C1[j] <- beta0.teste, beta0.C1[j] <- teste.b0)
### CONDICIONAL DE BETA 1 #####
teste.b1 <- rnorm(1, hiperparametro.a1, hiperparametro.b1)
C1 <- sum(log(1-(exp(beta0.C1[j]+beta1.teste*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.C1[j]+beta1.teste*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j]+beta1.teste*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))
+((beta1.teste/(Cr+1))*sum(matriz.r[,i]*x1)))
C2 <- sum(log(1-(exp(beta0.C1[j]+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.C1[j]+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j]+beta1.C1[j-1]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))
+((beta1.C1[j-1]/(Cr+1))*sum(matriz.r[,i]*x1)))
testeC <- exp(C1-C2)
u <- runif(1)
#ifelse(u < min(1, testeC), beta1.C1[j] <- beta1.teste, beta1.C1[j] <- beta1.C1[j-1])
#ifelse(u < min(1, testeC), beta1.C1[j] <- beta1.teste, beta1.C1[j] <- teste.b1)
### CONDICIONAL DE BETA 2 #####
teste.b2 <- rnorm(1, hiperparametro.a2, hiperparametro.b2)
D1 <- sum(log(1-(exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.teste*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.teste*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.teste*x2+beta3.C1[j-1]*x3))
+((beta2.teste/(Cr+1))*sum(matriz.r[,i]*x2)))
D2 <- sum(log(1-(exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j-1]*x2+beta3.C1[j-1]*x3))
+((beta2.C1[j-1]/(Cr+1))*sum(matriz.r[,i]*x2)))
testeD <- exp(D1-D2)
u <- runif(1)
#ifelse(u < min(1, testeD), beta2.C1[j] <- beta2.teste, beta2.C1[j] <- beta2.C1[j-1])
#ifelse(u < min(1, testeD), beta2.C1[j] <- beta2.teste, beta2.C1[j] <- teste.b2)
### CONDICIONAL DE BETA 3 #####
teste.b3 <- rnorm(1, hiperparametro.a3, hiperparametro.b3)
E1 <- sum(log(1-(exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j]*x2+beta3.teste*x3)/
(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j]*x2+beta3.teste*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j]*x2+beta3.teste*x3))
+((beta3.teste/(Cr+1))*sum(matriz.r[,i]*x3)))
E2 <- sum(log(1-(exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j]*x2+beta3.C1[j-1]*x3)/
(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j]*x2+beta3.C1[j-1]*x3))))^(1/(Cr+1))))
- sum((matriz.r[,i]/(Cr+1))*log(1+exp(beta0.C1[j]+beta1.C1[j]*x1+beta2.C1[j]*x2+beta3.C1[j-1]*x3))
+((beta3.C1[j-1]/(Cr+1))*sum(matriz.r[,i]*x3)))
testeE <- exp(E1-E2)
u <- runif(1)
#ifelse(u < min(1, testeE), beta3.C1[j] <- beta3.teste, beta3.C1[j] <- beta3.C1[j-1])
#ifelse(u < min(1, testeE), beta3.C1[j] <- beta3.teste, beta3.C1[j] <- teste.b3) }
#####
```

## 11 Referências Bibliográficas

- ARAÚJO, A. R. Regressão logística com resposta contínua. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2002.
- Best, N. G., Cowles, M. K., and Vines, K. (1995), "CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, University of Cambridge MRC Biostatistics Unit.
- BOX, G. E. P.; COX, D. R. An analysis of transformation (with discussion). *J. R. Stat. Soc. Ser. B*, v.26, p.211-252, 1964.
- BROYDEN, C. G. The convergence of a class of double-rank minimization algorithms. Parts I and II, *J. Inst. Math. Appl.*, Malden, p.76-90 e p.222-231, 1970.
- COOK, R. D. Detection of influential observations in linear regression. *Technometrics*, v.19, p.15-18, 1977.
- CORDEIRO, G. M. *Modelos Lineares Generalizados*. VII Simpósio Brasileiro de Probabilidade e Estatística. UNICAMP. Campinas, São Paulo, 1986.
- COX, D. R., SNELL, E. J. *Analysis of Binary Data*. London: Chapman & Hall, 1989.
- CHIB, S., GREENBERG, E. Understanding the Metropolis-Hastings algorithm. *American Statistical Association*, v.49, p.327-35, 1995.
- DEMÉTRIO, C. G. B. *Modelos Lineares Generalizados em Experimentação Agrônômica*. 46 Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBRAS) e 9 Simpósio de Estatística Aplicada à Experimentação Agrônômica (SEAGRO), ESALQ/USP. Piracicaba, São Paulo, 2001.
- DOBSON, A.J.; BARNETT, A.G. *Introduction to Generalized Linear Models*. 3rd ed, Boca Raton, FL: Chapman and Hall/CRC, 2008.
- FLETCHER, R. A new approach to variable metric algorithms. *Computacional Journal*, Oxford, v.13, p.317-322, 1970.
- GOLDFARB, D. A family of variable metric methods derived by variational means. *Math. Comp.*, v.26, p.23-26, 1970.
- GUPTA, P. L., GUPTA, R. C. e TRIPATHI, R. C. Inflated Modified Power Series Distributions with Applications, *Communications in Statistics - Theory and Methods*, v. 24, p. 2355-2374, 1995.
- HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, v.57, p.97-109, 1970.
- HOSMER, D. W., LEMESHOW, S. *Applied Logistic Regression*. John Wiley, New York, 2000.
- KENDALL, M. G.; STUART, A. *The Advanced Theory of Statistics - Distribution Theory*. New York: Hafner, 3rd ed., 1969.
- KLEINBAUM, D. G., KLEIN, M. *Logistic Regression: a self-learning text*. New York: Springer-Verlag, 2002.
- LINNET, K., BRANDT, E. Assessing diagnostic tests once an optimal cutoff point has been selected. *Clin. Chem.*, v.32, p.1341-6, 1986.

- McCULLAGH, P., NELDER, J.A. *Generalized Linear Models*. 2nd edition, Chapman and Hall: London, 1989.
- MENDES, C. C. Modelos para Dados de Contagem e Aplicações. Dissertação (Mestrado), Programa Pós-Graduação em Estatística - Universidade de Campinas, Campinas 2007.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H., TELLER, E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*. v.21, p.1087-1091, 1953.
- MURAT, M. e SZYNAL, D. Non-Zero Inflated Modified Power Series Distributions, *Communications in Statistics - Theory and Methods*, v. 27, p. 3047-3064, 1998.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society A*, 135, 3, p.370-84, 1972.
- PARAÍBA, L. C., QUEIROZ, S. C. N., MAIA, A. H. N., FERRACINI, V. L. Bioconcentration factor estimates of polycyclic aromatic hydrocarbons in grains of corn plants cultivated in soils treated with sewage sludge. *Science of the Total Environment* 408, 3270–3276, 2010.
- PAULA, G. A. *Modelos de Regressão com Apoio Computacional*. São Paulo: IME/USP. 2002.
- PRESS, W. H. *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. Cambridge: Cambridge University Press, 1992.
- RAO, C. R. *Linear Statistical Inference and its Applications*. John Wiley, New York, 1973.
- RONCHETTI, E., HERITIER, S., MORABIA, A. *Robust Binary Regression with Continuous Outcomes*. Genève: Cahiers du Département d'Econométrie, Université de Genève, 21p, 1997.
- SHANNO, D, F. Conditioning of quasi-Newton methods for function minimization. *Math. Comp.*, v.24, p. 647-657, 1970.
- SUISSA, S. Binary methods for continuous outcomes: a parametric alternative. *Journal of Clinical Epidemiology*, **44**, 241-248, 1991.
- SUISSA, S., BLAIS, L. Binary regression with continuous outcomes. *Statistics in Medicine*, **14**, 247-255, 1995.
- THOMPSON, R., BAKER, R. J. Composite link functions in generalized linear models. *Applied Statistics*, **30**, 125-131. 1981.