UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

MESTRADO EM MATEMÁTICA E ESTATÍSTICA

Pedro Silvestre da Silva Campos

ESTIMAÇÃO BAYESIANA EM MODELOS DE REGRESSÃO LOGÍSTICA

Orientadora: Profa. Dra. Maria Regina Madruga Tavares

Belém
2007

Pedro Silvestre da Silva Campos

ESTIMAÇÃO BAYESIANA EM MODELOS DE REGRESSÃO LOGÍSTICA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará como requisito parcial para a obtenção do grau de Mestre em Estatística.

Área de Concentração: Inferência Estatística
Orientadora: Profa. Dra. Maria Regina Madruga Tavares

Belém
2007

Pedro Silvestre da Silva Campos

ESTIMAÇÃO BAYESIANA EM MODELOS DE REGRESSÃO LOGÍSTICA

Esta dissertação será julgada para a obtenção do grau de Mestre em Estatística no Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará.

Belém, 30 de Agosto de 2007

Prof. Dr. Marcus Pinto da Costa da Rocha
(Coordenador do Programa de Pós-Graduação em Matemática e Estatística - UFPA)

Banca Examinadora

Profa. Dra. Maria Regina Madruga Tavares
Universidade Federal do Pará
Orientadora

Prof. Dr. Joaquim Carlos Barbosa Queiroz
Universidade Federal do Pará
Examinador

Prof. Dr. Hermínio Simões Gomes
Universidade Federal do Pará
Examinador

Dig e Rico.

Agradecimentos

⋆ À Deus;

⋆ À minha família, em especial aos meus pais, Pedro e Maria, por sempre mostrarem a importância do estudo e acreditarem na minha capacidade;

⋆ À Profa. Regina, minha orientadora e amiga, que em várias situações se comportou como minha mãe, chamando a atenção e mostrando o caminho a seguir, mas sempre incentivando a caminhar sozinho e pela confiança depositada;

⋆ Aos professores Héliton, Joaquim e Silvia que ajudaram, de alguma forma, na minha formação estatística. Além de outros professores como Protázio, Hermínio e Aldo que também contribuíram na minha formação;

⋆ Ao Prof. Aldo Vieira, que de forma implícita, me fez tomar o viés da Estatísitca;

⋆ À Universidade Federal do Pará;

⋆ À Faculdade de Estatística, representado pela pessoa do Prof. Dr. Joaquim Queiroz;

⋆ Ao Curso de Matemática a Distância, representado pela pessoa do Prof. Dr. José Miguel Martins Veloso;

⋆ Ao Programa de Pós-Graduação em Matemática e Estatística (PPGME), representado pela pessoa do Prof. Dr. Marcus Rocha;

⋆ À Secretaria de Educação do Estado do Pará (SEDUC) pelo apoio financeiro destinado à este trabalho;

⋆ À todos os alunos e funcionários do Programa de Pós-Graduação em Matemática e Estatística. Em especial aos alunos e ex-alunos: Edney, Raquel, Ulisses, Leandro, Heleno, Gracildo, Luiz Otávio, Jatene, Eraldo, Agostinho, Irazel, Sebastião e a funcionária Telma;

⋆ Aos amigos e colegas: Janair, Jardel, Silza, Janete, Iza, Ewerton, Silvério, José Antônio, Irene, Eliane e Midori;

⋆ Ao Odermar, Solange, Alexandre e Josy que tem me auxiliado durante esta impleitada;

⋆ Finalmente, à minha Esposa e filhos que tem aturado os momentos de mau humor, isolamento e ausência durante esta impleitada.

“Na crise, Estude!”

Resumo

CAMPOS, P.S.S. Estimação Bayesiana em Modelos de Regressão Logística, 2007. Dissertação de Mestrado (Programa de Pós-Graduação em Matemática e Estatística - UFPA, Belém - PA, Brasil).

Neste trabalho foram apresentados os métodos de Estimação Clássico e Bayesiano dos parâmetros dos modelos de regressão logística, bem como métodos Bayesianos de seleção e método de validação do modelo. A estimação Bayesiana apresentada, baseia-se na proposta de Groenewald e Mokgatlhe [2005], que fazem uso da introdução de variáveis latentes com distribuição uniforme no modelo. O uso de variáveis latentes com distribuições uniformes por Groenewald e Mokgatlhe [2005], tornaram de fácil implementação o processo de simulação das distribuições a posteriori dos parâmetros dos Modelos de Regressão Logística a partir do Amostrador de Gibbs, utilizado para estimar os parâmetros destes modelos em dados reais. Na etapa de seleção do modelo foram utilizados o do Fator de Bayes (FB), BIC e da proposta de Pereira e Stern [1999], o FBST. O ajuste do modelo foi satisfatório nos dados considerados, produzindo erros pequenos nas estimações geradas pelos modelos ajustados.

Palavras-chave: Regressão Logística, Variável Latente, Amostrador de Gibbs.

Abstract

Campos, P.S.S. Bayesian Estimation in Logistic Regression Models, 2007. Dissertation of Master's degree (Graduate Program in Mathematics and Statistics - UFPA, Belém - PA, Brazil).

Methods of Classical and Bayesian estimation of the Logistic Regression models parameters as well as Bayesian methods of selection and validation of the models were presented. The Bayesian estimation presented, is based on the proposal of the Groenewald and Mokgatlhe [2005], that make use of the introduction of latent variables with uniform distribution on the model. The use of latents variables with uniform distributions by Groenewald and Mokgatlhe [2005], turned out to be of easy implementation in the process of simulation of the distribution a posteriori of logistic Regression model parameters derived from Gibbs Sampler, used to estimate these model parameters with real data. Upon the model selection the Bayes Factor (FB), BIC and the FBST (proposed by Pereira and Stern [1999]) were used. The adjustment of the model was satisfactory with the given data, producing small errors of the estimates generated by adjusted model.

Key words: Logistic Regression, Latentes Variables, Gibbs Sampler.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Os modelos estatísticos de regressão são utilizados para predizer resultados de uma variável dependente Y, através do ajuste de uma relação funcional entre Y e um conjunto de variáveis preditoras (independentes). Dentre os diversos modelos de regressão, o mais conhecido e utilizado é o modelo de Regressão Linear, que estabelece uma relação funcional linear entre Y e as preditoras. Para o ajuste deste modelo, é necessário o atendindimento de algumas suposições no processo de estimação.

As suposições que devem ser consideradas nos modelos de Regressão Linear são as seguintes: para cada valor da variável independente, a distribuição da variável dependente deve ser normalmente distribuída; a variância da distribuição da variável dependente deve ser constante para todos os valores da variável independente, ou seja, o modelo é homocedástico; a relação entre a variável dependente e cada variável independente deveria ser linear, e todas as observações deveriam ser independentes. Isto é, o modelo de Regressão Linear assume que há uma relação funcional linear entre a variável dependente e cada variável preditora. Esta relação é dada por:

$$Y_i = \boldsymbol{\beta}\boldsymbol{X_i} + \varepsilon_i; i = 1, 2, 3, \ldots, n$$

tal que, $\boldsymbol{X_i} = (1, X_{i1}, X_{i2}, \ldots, X_{ip})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ e $\varepsilon_i \sim N(0, \mathbf{1}\sigma^2)$. O método clássico usual de estimação dos parâmetros do modelo é o método de Mínimos Quadrados (Casella e Berger [2002], Draper e Smith [1998], Neter *et al.* [1996]).

No entanto, existem situações em que a variável dependente não é quantitativa, mas sim qualitativa, ou seja, não se pode ajustar um modelo de Regressão Linear, haja visto que a condição de normalidade da variável dependente não é satisfeita. Para contornar este problema Nelder e Wedderburn [1972] (*apud* Cordeiro [1986]) propuseram os Modelos Lineares Generalizados (MLG), que foi um marco no tratamento de modelos de regressão

com variáveis respostas não necessariamente normais, pois vieram a unificar a teoria que envolve os modelos de regressão (Agresti [2002]).

Um MLG tem sua estrutura formada por três partes: uma *componente aleatória*, uma *componente sistemática* e uma função monótona diferenciável, dita *função de ligação*, que relaciona as componentes aleatória e sistemática (Fahrmeir e Tutz [2001], Cordeiro [1986]. Os MLG serão descritos no Capítulo 2.

Dentre os Modelos Lineares Generalizados estão os Modelos de Regressão Logística, que são modelos de regressão não linear usados quando a variável resposta é qualitativa, com duas ou mais categorias.

Os modelos de Regressão Logística são úteis em situações em que se deseja estimar a proporção de uma determinada característica ou resultado baseado em valores de um conjunto de variáveis preditoras. É semelhante ao modelo de Regressão Linear, mas a variável dependente é qualitativa (dicotômicas ou policotômicas (nominais ou ordinais)), ou seja, estes modelos servem para modelar respostas categorizadas (Chen *et al.* [1999]). A resposta da função Logística, assim como outras funções de resposta, é usada para descrever a natureza da relação entre a resposta média e um conjunto de variáveis preditoras (Neter *et al.* [1996]).

Neste trabalho serão considerados os modelos de Regressão Logística com funções lineares e não lineares na preditora, com o objetivo de estimar seus parâmetros para dados categorizados a partir do algoritmo proposto por Groenewald e Mokgatlhe [2005] que ajusta modelos de Regressão Logística fazendo uso da metodologia Bayesiana.

A proposta de Groenewald e Mokgatlhe [2005] está baseada no trabalho de Albert e Chib [1993], que utiliza variáveis latentes com distribuição normal no processo de estimação dos parâmetros de um modelo *probit* (no modelo de Regressão Probit a função de ligação é a função acumulada da distribuição normal). Sendo que Groenewald e Mokgatlhe [2005] utilizaram variáveis latentes com distribuição uniforme e função de ligação Logística.

A seleção do modelo será feita com o uso do *Fator de Bayes* (Kass e Raftery [1995]), do BIC (*Bayesian Information Criterion*) e com o procedimento apresentado por Pereira e Stern [1999], o FBST (*Full Bayesian Statistical Test*). O *Fator de Bayes* seleciona o

melhor modelo comparando as probabilidades a posteriori (Kass e Raftery [1995], Berger e Pericchi [1997], Paulino *et al.* [2003]), o BIC faz a comparação entre as verossimilhanças a posteriori (Paulino *et al.* [2003]) e a proposta de Pereira e Stern[1999], que é um procedimento para testar hipóteses paramétricas, que se baseia no cálculo da probabilidade da Região $\boldsymbol{HPD}$ (*Highest Posteriori Density*) tangente ao conjunto que define a hipótese nula. A Evidência Bayesiana em favor da hipótese nula é o complementar da probabilidade da Região $\boldsymbol{HPD}$(Pereira e Stern [1999], Madruga *et al.* [2001]).

1.1 Justificativa e Importância

Os Modelos de Regressão Logística vêm sendo aplicados em larga escala nos meios científicos, por tratar de variáveis qualitativas que muitas vezes surgem como a variável resposta de um conjunto de dados.

Os Modelos de Regressão Logística são modelos que visam criar uma relação funcional entre uma variável aleatória $\boldsymbol{Y}$ (qualitativa) com um vetor $\boldsymbol{X}$ de variáveis preditoras. Várias são as aplicações dos Modelos de Regressão Logística, para exemplificar: nas áreas de Ciências Biológicas e Avaliação Educacional. Nas Ciências Biológicas, um dos casos é modelar dosimetria citogenética (Madruga *et al.* [1994]), em Avaliação Educacional são usados nos modelos da Teoria da Resposta ao Item (TRI) (Andrade *et al.* [2000]).

Neste trabalho, os parâmetros dos Modelos de Regressão Logística serão estimados a partir da Metodologia Bayesiana, que considera os parâmetros de um modelo como variáveis aleatórias, diferentemente da Metodologia Clássica, que considera os parâmetros do modelo como fixos e sem nenhum conhecimento prévio dos mesmos. Segundo O'Hagan [1994] esta é a diferença fundamental entre as duas metodologias.

A Metodologia Bayesiana leva em consideração duas fontes de informação: o conhecimento prévio sobre o processo, representado pela distribuição a priori, em conjunto com as informações trazidas pelos dados, através da função de verossimilhança. Estas duas fontes de infomação são usadas simultaneamente fazendo uso do *Teorema de Bayes*, que é a ferramenta básica de toda Metodologia Bayesiana.

A metodologia de estimação a ser utilizada é recente, pois tem suas bases no trabalho de Albert e Chib [1993] em modelos de regressão *probit* para dados qualitativos, cujos parâmetros foram estimados fazendo uso de variáveis latentes com distribuição normal.

A proposta aqui se faz importante por tratar de metodologia de regressão para dados qualitativos utilizando metodologia Bayesiana, semelhante a utilizada por Albert e Chib [1993], que foi proposta por Groenewald e Mokgatlhe [2005] fazendo uso de variáveis latentes com distribuição uniforme e também por fazer a seleção do modelo utilizando técnicas Bayesianas, que até pouco tempo eram de difícil implementação, devido a falta de recursos computacionais.

1.2 Objetivos

1.2.1 Objetivo Geral

Utilizar a Metodologia proposta por Groenewald e Mokgatlhe [2005] para ajustar Modelos de Regressão Logística em dados reais.

1.2.2 Objetivos Específicos

- Ajustar modelos de Regressão Logística para dados categorizados usando metodologia Bayesiana;
- Implementar computacionalmente a metodologia proposta por Groenewald e Mokgatlhe [2005];
- Comparar os métodos de seleção do modelo ajustado, *Fator de Bayes*, *BIC* e *FBST*;
- Usar dados reais para ilustrar a aplicação do metodologia proposta;

1.3 Estrutura do Trabalho

Este trabalho encontra-se dividido em seis capítulos, a saber:

- Capítulo 1: refere-se à introdução do trabalho, contendo a justificativa e importância do trabalho, objetivo geral e objetivos específicos;
- Capítulo 2: serão apresentados os Modelos Lineares Generalizados e a formalização dos Modelos de Regressão Logística;
- Capítulo 3: serão apresentadas os métodos de Estimação Clássico e Bayesiano dos parâmetros nos Modelos de Regressão Logística;

- Capítulo 4: serão apresentadas técnicas Bayesianas de seleção do modelo, que são o Fator de Bayes, o BIC e a proposta de Pereira-Stern o FBST, bem como a técnica de validação do modelo;
- Capítulo 5: serão apresentadas aplicações em Modelos de Regressão Logística Dicotômica e Policotômica;
- Capítulo 6: serão apresentadas as conclusões e recomendações para trabalhos futuros.

Capítulo 2

Modelos de Regressão Logística

Antes de fazer a formalização dos Modelos de Regressão Logística será feita a descrição dos Modelo Lineares Generalizados, pois os Modelos de Regressão Logística são casos particulares de Modelos Lineares Generalizados para dados categorizados.

2.1 Modelos Lineares Generalizados

Nelder e Wedeburn [1972] (*apud* Cordeiro [1986]) propuseram uma teoria unificadora da modelagem estatística, e deram o nome de Modelos Lineares Generalizados (MLG). Um MLG é formado por três partes: uma *componente aleatória*, composta de uma variável aleatória $\boldsymbol{Y}$, com n observações independentes, pertencente à família exponencial; uma *componente sistemática*, composta por variáveis preditoras e uma função monótona e diferenciável, dita *função de ligação*, que relaciona as componentes aleatórias e sistemática. Estas três partes serão descritas a seguir, segundo Cordeiro [1986]:

- Componente Aleatória
  Considere um vetor $\boldsymbol{y} = (y_1, \ldots, y_n)'$ como realização das variáveis aleatórias $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, independentemente distribuídas com médias $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ e função de probabilidade ou função densidade de probabilidade pertencente à *família exponencial*, ou seja,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = \exp[c(\boldsymbol{\theta})T(\boldsymbol{y}) + d(\boldsymbol{\theta}) + S(\boldsymbol{y})] \tag{2.1}$$

  onde c(.) e d(.) são funções reais de $\boldsymbol{\theta}$; T e S são funções reais de $\boldsymbol{y}$

- Componente Sistemática
  Considere a estrutura linear de um modelo de regressão

$$\boldsymbol{\eta} = \boldsymbol{\beta X_i} \tag{2.2}$$

onde $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ e $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ip})'$ é uma matriz modelo $n \times (p+1)$ $(n > p+1)$ conhecida de posto $p+1$.

A função $\boldsymbol{\eta}$ dos parâmetros desconhecidos $\boldsymbol{\beta}$, que devem ser estimados, chama-se *preditor linear*.

- Função de Ligação

  As componentes aleatórias e sistemática relacionam-se através de uma função $f(.)$, monótona e diferenciável, denominada de *função de ligação* que transforma μ_i em η_i, ou seja,

$$\eta_i = f(\mu_i) \Leftrightarrow \mu_i = f_i^{-1}(\eta_i), i = 1, \ldots, n. \tag{2.3}$$

  São funções de ligação para dados categorizados:

Tabela 2.1 *Funções de Ligação para dados categorizados.*

Nome	Transformação
logit	$\eta_i = \dfrac{\pi}{1-\pi}$
probit	$\eta_i = \Phi^{-1}(\pi)$
complemento log-log	$\eta_i = log(-log(1-\pi))$

Neste trabalho a função de ligação utilizada será a *logit* que dá origem aos Modelos de Regressão Logística.

2.2 Modelos de Regressão Logística

Segundo Hosmer e Lemeshow [2000], a regressão logística busca explicar a relação, através de um modelo, entre uma variável dependente e um conjunto de variáveis independentes, chamadas de covariáveis. Nesta secção será feita a descrição dos modelos logísticos Dicotômico e Policotômicos (Nominal e Ordinal).

2.2.1 Variável Resposta Dicotômica

Seja Y_i uma variável aleatória dicotômica, ou seja, que admite apenas dois valores possíveis (certo ou errado, sim ou não, etc.), tal que seu modelo de probabilidade pode ser representado como

$$Y_i = \begin{cases} 1, & \text{com probabilidade } \pi_i \\ 0, & \text{com probabilidade } 1-\pi_i. \end{cases}$$

Suponha que a probabilidade π_i associada aos valores de Y_i dependa de um vetor de covariáveis $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, ..., X_{ip})'$.

Usando a transformação *logit* em π_i com uma estrutura linear em $\boldsymbol{X}_i$, tem-se

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\beta X_i}, \tag{2.4}$$

tal que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ é o vetor de parâmetros.

De (2.4), tem-se que:

$$\pi_i = \frac{\exp(\boldsymbol{\beta X_i})}{1 + \exp(\boldsymbol{\beta X_i})}. \tag{2.5}$$

A partir de (2.5) é fácil concluir que π_i tem função de distribuição acumulada de uma variável aleatória logística.

2.2.2 Variável Resposta Policotômica

Seja Y_i uma variável aleatória que assume valores em r $(r > 2)$ categorias possíveis. Nestas condições diz-se que Y_i tem distribuição multinomial com r categorias e vetor de parâmetros $\boldsymbol{\pi_{ij}} = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir})$ (ver Fahrmeir e Tutz [2001]). Inicialmente, será considerada a variável *policotômica nominal*, ou seja, aquela em que as r categorias não são classificadas segundo uma ordem, sendo a probabilidade da i-ésima observação pertencer à categoria j denotada por

$$\pi_{ij} = P\left(Y_i = j\right), \qquad j = 1, 2, ..., r.$$

Suponha que a probabilidade π_i associada aos valores de Y_i dependa de um vetor de covariáveis $\boldsymbol{X_i} = (1, X_{i1}, X_{i2}, ..., X_{ip})$, ou seja,

$$Y_i|\boldsymbol{X_i} \sim Multinomial(\pi_{ij}, N)$$

com $\sum_{j=1}^{r} \pi_{ij} = 1$.

Usando a transformação *logit* de π_{ij} com uma estrutura linear em $\boldsymbol{X_i}$, tem-se

$$logit(\pi_{ij}) = log\left(\frac{\pi_{ij}}{\pi_{ir}}\right) = \boldsymbol{\beta_j X_i}, \qquad j = 1, 2, \ldots, r-1. \tag{2.6}$$

De (2.6) resulta que

$$\frac{\pi_{ij}}{\pi_{ir}} = \exp\left(\boldsymbol{\beta_j X_i}\right), \tag{2.7}$$

com

$$\pi_{i1} + \pi_{i2} + \pi_{i3} + ... + \pi_{ir} = 1 \tag{2.8}$$

e dividindo (2.8) por $\pi_{ir} > 0$, tem-se que

$$\frac{1}{\pi_{ir}} = 1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta_s X_i}). \tag{2.9}$$

De (2.7) e (2.8) segue que:

$$\pi_{ij} = \frac{\exp(\boldsymbol{\beta_j X})}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta_s X_i})} \tag{2.10}$$

Há casos em que as categorias obedecem a uma ordem natural, e nestes casos a variável policotômica é dita Ordinal.

Seja Y_i uma variável aleatória ordenada em r categorias e π_{ij} a probabilidade da i-ésima observação pertencer a categoria j. A probabilidade acumulada das categorias, ou seja, a probabilidade de um indivíduo i pertencer a categoria menor ou igual a j será denotada por η_{ij} e dada por:

$$\eta_{ij} = \sum_{k=1}^{j} \pi_{ik} = P(Y_i \leq j), \qquad j = 1, \ldots, r. \tag{2.11}$$

O modelo de Regressão Logística Ordinal supõe que os logits cumulativos podem ser representados como funções lineares paralelas de variáveis independentes, ou seja, para cada logit cumulativo os parâmetros do modelo são os mesmos, à exceção do intercepto (Fahrmeir e Tutz [2001], Agresti [2002]). Sendo assim:

$$logit(\eta_{ij}) = \log\left(\frac{\eta_{ij}}{1 - \eta_{ij}}\right) = \log\left[\frac{P(Y_i \leq j)}{1 - P(Y_i \leq J)}\right] = \alpha_j + \boldsymbol{\beta X_i}. \tag{2.12}$$

Segue de (2.12) que:

$$\eta_{ij} = \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})} \tag{2.13}$$

tal que $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ é o vetor de parâmetros da parte linear, $\boldsymbol{X_i} = (X_{i1}, X_{i2}, ..., X_{ip})'$ é o vetor de covariadas regressoras e $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_r)$ é o vetor de interceptos, tal que $-\infty = \alpha_0 < \alpha_1 < ... < \alpha_r = \infty$.

Segundo McCullagh [1980] (*apud* Fahrmeir e Tutz [2001]) o modelo logístico cumulativo também é chamado de modelo da proporção de chances (*proportional odds*).

Capítulo 3

Estimação dos Parâmetros

Em experimentos envolvendo análise estatística de dados é comum a necessidade de se fazer inferências (estimação ou testes de hipóteses) sobre algum parâmetro de interesse θ, associado ao modelo de probabilidades da variável em estudo e assumindo valores no Espaço Paramétrico $\boldsymbol{\Theta}$. A inferência estatística paramétrica pode ser feita sob uma das seguintes abordagens: *Inferência Clássica* ou *Inferência Bayesiana*.

Na Inferência Clássica, o parâmetro é considerado uma quantidade fixa e desconhecida. Os resultados são obtidos a partir da distribuição conjunta da amostra observada de tamanho n (dados), $\boldsymbol{x} = (x_1, ..., x_n)$, e representada pela função de verossimilhança $L(\boldsymbol{\theta; x})$. A informação trazida pela amostra sobre o parâmetro desconhecido é representada por alguma função dos dados, denominada "estatística"e, com base na sua distribuição amostral (Lehmann [1959]), são avaliadas as propriedades dos estimadores.

Na Inferência Bayesiana, o parâmetro é considerado uma variável aleatória desconhecida. Neste caso, o grau de incerteza sobre o valor de θ é representado por um modelo de probabilidade definido em $\boldsymbol{\Theta}$, denominado de *Distribuição a Priori* e representado pela função (de densidade ou de probabilidade) $\pi(\theta)$. Este grau de incerteza inicial é atualizado usando a informação trazida pela amostra, através da função de verossimilhança. Assim, o grau de incerteza atualizado passa a ser representado por um novo modelo de probabilidade, denominado de *Distribuição a Posteriori*, e representado pela função $\pi(\boldsymbol{\theta|x})$. Essa atualização é feita utilizando-se o Teorema de Bayes (Mood *et al.* [1974]):

$$\pi(\boldsymbol{\theta|x}) = \frac{L(\boldsymbol{\theta; x})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta; x})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Toda a inferência Bayesiana, seja obtenção de estimadores ou Teste de Hipóteses, é obtida a partir da Distribuição Posterior (Bernardo e Smith [1994]), sendo esta na maioria das vezes difícil de ser determinada analiticamente.

Os métodos de Monte Carlo para Cadeias de Markov (MCMC) em especial o "Gibbs

Sampler", será utilizado para simular amostras da distribuição a posteriori, que serão usadas para obtenção dos estimadores e realização dos testes de hipóteses apropriados (Gilks *et al.* [1996]). A implementação desses métodos dar-se-á através da elaboração de rotinas computacionais nos softwares MATLAB e no pacote WINBUGS que já possui este algoritmo implementado. No desenvolvimento deste trabalho realizou-se um levantamento bibliográfico sobre o tema em estudo, visando a atualização das informações técnicas.

3.1 Método de Estimação Clássico

O método usual para estimação clássica dos modelos de Regressão Logística é o método de Máxima Verosimilhança que, em geral, já está implementado em pacotes estatísticos (Hosmer e Lemeshow [2001]). Este método baseia-se na determinação dos valores que maximizam a função de verossimilhança. A função de verossimilhança expressa a probabilidade dos dados observados como função dos parâmetros desconhecidos.

O método de estimação por Máxima Verossimilhança tem por objetivo determinar o valor do parâmetro $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, que maximiza a função de verossimilhança $L(\boldsymbol{\theta}; \boldsymbol{Y})$. Segundo Casella e Berger [2002] a determinação do vetor de estimadores $\hat{\boldsymbol{\theta}}$ que maximiza a *função de verossimilhança* é equivalente a maximizar $\log L(\boldsymbol{\theta}; \boldsymbol{Y})$. Assim, uma condição necessária para maximizar $l = \log L(\boldsymbol{\theta}; \boldsymbol{Y})$, é dada pelas equações:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{3.1}$$

são conhecidas como *equações de máxima verossimilhança* e suas soluções são os estimadores de máxima verossimilhança.

As equações de máxima verossimilhança quase nunca admitem soluções analíticas, sendo que nos casos em que não é possível se encontrar analiticamente os estimadores de máxima verossimilhança, faz-se uso de métodos numéricos como *Newton-Raphson* ou *Escore de Fisher* (Fahrmeir e Tutz [2000]).

É feita a seguir a estimação para os Modelos Logísticos Dicotômico e Policotômicos, segundo a Metodologia Clássica.

3.1.1 Variável Resposta Dicotômica

Sejam Y_i v.a's i.i.d (variáveis aleatórias independentes e identicamente distribuídas) e seja $\boldsymbol{X_i}$ o vetor de covariadas, tal que, a distribuição de $Y_i|\boldsymbol{X}_i$ tenha distribuição $Bernoulli(\pi_i)$, ou seja,

$$Y_i|\boldsymbol{X}_i \sim Bernoulli(\pi_i)$$

a função de verossimilhança é dada por:

$$\begin{aligned} L(\boldsymbol{\beta}; Y_i|X_i) &= \prod_{i=1}^{n} [\pi_i]^{y_i} [1-\pi_i]^{1-y_i} \\ &= \prod_{i=1}^{n} \left[\frac{\pi_i}{1-\pi_i}\right]^{y_i} . [1-\pi_i] . \end{aligned} \tag{3.2}$$

Mas pela transformação *logit*, tem-se que

$$\frac{\pi_i}{1-\pi_i} = \exp(\boldsymbol{\beta X_i}) \tag{3.3}$$

e

$$1-\pi_i = \frac{1}{1+\exp(\boldsymbol{\beta X_i})}, \tag{3.4}$$

e substituindo (3.3) e (3.4) em (3.2), tem-se que:

$$L(\boldsymbol{\beta}; \boldsymbol{Y_i}|\boldsymbol{X_i}) = \prod_{i=1}^{n} \exp(y_i \boldsymbol{\beta X_i}) [1+\exp(\boldsymbol{\beta X_i})]^{-1} . \tag{3.5}$$

O logaritmo da verossimilhança em (3.5) será dado por:

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta; Y_i}|\boldsymbol{X_i})) = \sum_{i=1}^{n} [y_i \boldsymbol{\beta X_i} - \log(1+\exp \boldsymbol{\beta X_i})] . \tag{3.6}$$

Derivando (3.6) em relação ao t-ésimo parâmetro, tem-se:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_t} = \sum_{i=1}^{n} X_{it} (y_i - \pi_i) . \tag{3.7}$$

Os valores dos estimadores, são os valores que satisfazem $\dfrac{\partial l(\boldsymbol{\beta})}{\partial \beta_t} = \mathbf{0}$ acima.

3.1.2 Variável Resposta Policotômica

i) Multinomial Nominal

Para estimação dos parâmetros do modelo de Regressão Logística Multinomial Nominal, via método de Máxima Verossimilhança, faz-se necessário o uso de variáveis indicadoras, que são introduzidas apenas para facilitar a descrição da função de verossimilhança, mas não são usadas no modelo Multinomial de Regressão Logística (Hosmer e Lemeshow [2001], Agresti [2002]).

A variável indicadora será denotada por:

$$k_{ij} = \begin{cases} 1, & \text{se a observação i pertence à categoria j} \\ 0, & \text{outros casos} \end{cases}$$

note que, $\sum_{j=1}^{r} k_{ij} = 1$.

Usando esta notação e sabendo que a probabilidade da i-ésima observação pertencer a j-ésima categoria é dada por

$$P(Y_i = j) = \pi_{ij} = \frac{\exp(H_{ij})}{1 + \sum_{j=1}^{r-1} \exp(H_{ij})}, \qquad j = 1, 2, \ldots, r \tag{3.8}$$

sendo $H_{ij} = \boldsymbol{\beta_j X_i}$, com $\boldsymbol{\beta_j} = (\beta_{oj}, \beta_{1j}, ..., \beta_{pj})$ e $\boldsymbol{X_i} = (1, X_{i1}, X_{i2}, ..., X_{ip})'$, a função de verossimilhança será dada por:

$$L(\boldsymbol{\beta; Y_i | X_i}) = \prod_{i=1}^{n} \left[\pi_{i1}^{k_{i1}} . \pi_{i2}^{k_{i2}} . \pi_{i3}^{k_{i3}} . \cdots . \pi_{ir}^{k_{ir}}\right]. \tag{3.9}$$

Determinando $\log\left[L(\boldsymbol{\beta; Y_i | X_i})\right] = l(\boldsymbol{\beta})$, e sabendo que $\sum_{j=1}^{r} k_{ij} = 1$, tem-se que:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \log\left[\prod_{i=1}^{n} \left(\pi_{i1}^{k_{i1}} . \pi_{i2}^{k_{i2}} . \pi_{i3}^{k_{i3}} . \cdots . \pi_{ir}^{k_{ir}}\right)\right] \\ &= \sum_{i=1}^{n} \left[\log \pi_{i1}^{k_{i1}} + \log \pi_{i2}^{k_{i2}} + \log \pi_{i3}^{k_{i3}} + \cdots + \log \pi_{ir}^{k_{ir}}\right] \\ &= \sum_{i=1}^{n} \left[k_{i1} \log \pi_{i1} + k_{i2} \log \pi_{i2} + k_{i3} \log \pi_{i3} + \cdots + k_{ir} \log \pi_{ir}\right]. \end{aligned} \tag{3.10}$$

Usando a definição de π_{ij} em (3.10), $j = 1, 2, 3, ..., r$ e usando o fato que $\sum_{j=1}^{r} \pi_{ij} = 1$,

tem-se que:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[\sum_{s=1}^{r-1} k_{is}\log\pi_{is} + \left(1-\sum_{s=1}^{r-1}k_{is}\right)\log\left(1-\sum_{s=1}^{r-1}\pi_{is}\right)\right] \\ &= \sum_{i=1}^{n}\left[\sum_{s=1}^{r-1} k_{is}\log\pi_{is} - \sum_{s=1}^{r-1}k_{is}\log\left(1-\sum_{s=1}^{r-1}\pi_{is}\right) + \log\left(1\sum_{s=1}^{r-1}\pi_{is}\right)\right] \\ &= \sum_{i=1}^{n}\left[\sum_{s=1}^{r-1} k_{is}\log\left(\frac{\pi_{is}}{1-\sum_{s=1}^{r-1}\pi_{is}}\right) + \log\left(1-\sum_{s=1}^{r-1}\pi_{is}\right)\right] \end{aligned}$$

mas $\pi_{ir} = 1-\sum_{s=1}^{r-1}\pi_{is} = \dfrac{1}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}$, logo

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[\sum_{s=1}^{r-1} k_{is}\log\left(\frac{\pi_{is}}{\pi_{ir}}\right) + \log(\pi_{ir})\right] \\ &= \sum_{i=1}^{n}\left[\sum_{s=1}^{r-1} k_{is}\boldsymbol{\beta_s X_i} - \log\left(1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})\right)\right] \end{aligned} \tag{3.11}$$

Derivando $l(\boldsymbol{\beta})$ em relação a cada categoria e a cada parâmetro desta categoria, tem-se de forma geral:

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{jt}} &= \sum_{i=1}^{n}\left[k_{ij}X_{it} - \frac{X_{it}\exp(\boldsymbol{\beta_j X})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}\right] \\ &= \sum_{i=1}^{n} X_{it}\left[k_{ij} - \frac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}\right] \\ &= \sum_{i=1}^{n} X_{it}\left(k_{ij}-\pi_{ij}\right), \end{aligned} \tag{3.12}$$

tal que $j = 1, 2, \ldots, r-1$ e $t = 0, 1, \ldots, p$.

Os valores que resolvem $\dfrac{\partial l(\boldsymbol{\beta})}{\partial \beta_{jt}} = \mathbf{0}$ não podem ser determinados de forma analítica, mas sim por métodos numéricos de Newton-Raphson ou Escore de Fischer, e assim pode-se obter os estimadores de Máxima Verossimilhança (ver Hosmer e Lemeshow [2000] e Agresti [2002]).

ii) Multinomial Ordinal

O caso de estimação dos parâmetros do modelo de regressão logística multinomial ordinal pelo método de Máxima Verossimilhaça é semelhante à metodologia empregada para estimar os parâmetros do modelo multinomial nominal.

No entanto, deve-se observar na estimação do modelo multinomial ordinal que

$$\gamma_{ij} = P(y_i \leq j) = \pi_{i1} + \pi_{i2} + \ldots + \pi_{ij} = \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}, \tag{3.13}$$

onde $\pi_{ij} = P(y_i \leq j)$, e segue que

$$\pi_{ij} = \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta X_i})} \tag{3.14}$$

Logo a função de verossimilhança será dada por,

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{Y}) = \prod_{i=1}^{n} \left[\pi_{i1}^{k_{i1}}.\pi_{i2}^{k_{i2}}.\pi_{i3}^{k_{i3}}.\cdots.\pi_{ir}^{k_{ir}}\right] \tag{3.15}$$

onde k_{ij} é a variável indicadora usada no modelo multinomial nominal, ou seja,

$$k_{ij} = \begin{cases} 1, & \text{se a observação i pertence à categoria j} \\ 0. & \text{outros casos} \end{cases}$$

O log da verossimilhaça é dado por:

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) \;=\; \sum_{i=1}^{n}\sum_{s=1}^{r} k_{is} \log \pi_{is} \tag{3.16}$$

Derivando o $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ da função de verossimilhança em relação ao intercepto na categoria j:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_j} &= \sum_{i=1}^{n} \left\{ k_{ij} \left[\frac{\left(\frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}\right)' - \left(\frac{\exp(\alpha_{j-1} + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta X_i})}\right)'}{\frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta X_i})}} \right] \right. \\
&\quad \left. + k_{ij+1} \left[\frac{\left(\frac{\exp(\alpha_{j+1} + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_{j+1} + \boldsymbol{\beta X_i})}\right)' - \left(\frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}\right)'}{\frac{\exp(\alpha_{j+1} + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_{j+1} + \boldsymbol{\beta X_i})} - \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}} \right] \right\} \\
&= \sum_{i=1}^{n} \left[k_{ij} \frac{\gamma_{ij}(1 - \gamma_{ij})}{\gamma_{ij} - \gamma_{ij-1}} + k_{ij+1} \frac{\gamma_{ij}(1 - \gamma_{ij})}{\gamma_{ij} - \gamma_{ij+1}} \right]
\end{aligned} \tag{3.17}$$

e derivando $l(\boldsymbol{\beta}, \boldsymbol{\alpha})$ em relação a cada β_t do vetor de parâmetros $\boldsymbol{\beta}$, tem-se

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \beta_t} &= \sum_{i=1}^{n}\sum_{s=1}^{r}\left\{ k_{is}\left[\frac{\left(\frac{\exp(\alpha_s+\boldsymbol{\beta X_i})}{1+\exp(\alpha_s+\boldsymbol{\beta X_i})}\right)' - \left(\frac{\exp(\alpha_{s-1}+\boldsymbol{\beta X_i})}{1+\exp(\alpha_{s-1}+\boldsymbol{\beta X_i})}\right)'}{\frac{\exp(\alpha_s+\boldsymbol{\beta X_i})}{1+\exp(\alpha_s+\boldsymbol{\beta X_i})} - \frac{\exp(\alpha_{s-1}+\boldsymbol{\beta X_i})}{1+\exp(\alpha_{s-1}+\boldsymbol{\beta X_i})}}\right]\right. \\ &\quad \left. + k_{is+1}\left[\frac{\left(\frac{\exp(\alpha_{s+1}+\boldsymbol{\beta X_i})}{1+\exp(\alpha_{s+1}+\boldsymbol{\beta X_i})}\right)' - \left(\frac{\exp(\alpha_s+\boldsymbol{\beta X_i})}{1+\exp(\alpha_s+\boldsymbol{\beta X_i})}\right)'}{\frac{\exp(\alpha_{s+1}+\boldsymbol{\beta X_i})}{1+\exp(\alpha_{s+1}+\boldsymbol{\beta X_i})} - \frac{\exp(\alpha_s+\boldsymbol{\beta X_i})}{1+\exp(\alpha_s+\boldsymbol{\beta X_i})}}\right]\right\} \\ &= \sum_{i=1}^{n}\sum_{s=1}^{r}\left[k_{is}X_{it}(1-\gamma_{is}-\gamma_{is-1}) + k_{is+1}X_{it}(1-\gamma_{is+1}-\gamma_{is})\right]. \end{aligned} \tag{3.18}$$

Com isso, as equações de verossimilhança serão dadas por:

$$\sum_{i=1}^{n}\left[k_{ij}\frac{\gamma_{ij}(1-\gamma_{ij})}{\gamma_{ij}-\gamma_{ij-1}} + k_{ij+1}\frac{\gamma_{ij}(1-\gamma_{ij})}{\gamma_{ij}-\gamma_{ij+1}}\right] = \mathbf{0} \tag{3.19}$$

e

$$\sum_{i=1}^{n}\sum_{s=1}^{r}\left[k_{is}X_{it}(1-\gamma_{is}-\gamma_{is-1}) + k_{is+1}X_{it}(1-\gamma_{is+1}-\gamma_{is})\right] = \mathbf{0}. \tag{3.20}$$

Sendo que para o modelo multinomial ordinal, o que difere para cada categoria são os interceptos ($\alpha's$). Portanto, para cada categoria os interceptos serão estimados pela solução das equações de verossimilhança (3.19) e todos os outros parâmetros serão estimados por (3.20), através do uso de métodos numéricos. McCullagh [1980], Walker e Duncan [1967] usaram Escore de Fischer (*apud* Agresti [2002]).

3.2 Método de Estimação Bayesiano

3.2.1 Elemento básico da Inferência Bayesiana

O elemento básico da Inferência Bayesiana é o *Teorema de Bayes*. O *Teorema de Bayes* faz uso da informação trazida pelos dados resumidos na *função de verossimilhança* e da informação prévia sobre o modelo resumida na *distribuição a priori*, sendo que o *Teorema de Bayes* unifica estas duas fontes de informação e as resume por meio da *distribuição a posteriori*.

Seja $\boldsymbol{X}$ uma variável aleatória com função de probabilidade (f.p.) ou função densidade de probabilidade (f.d.p.) de $\boldsymbol{X}$ denotada por $f(\boldsymbol{X}|\theta)$, tal que $\boldsymbol{\theta}$ é o vetor de

parâmetros associado à variável aleatória $\boldsymbol{X}$. O grau de incerteza sobre o valor de $\boldsymbol{\theta}$ é representado por um modelo de probabilidade definido no espaço paramétrico $\boldsymbol{\Theta}$, denominado de *Distribuição a Priori* e representado pela função (de densidade ou de probabilidade) $\pi(\boldsymbol{\theta})$. Este grau de incerteza inicial é atualizado usando a informação trazida pela amostra, através da função de verossimilhança. Assim, o grau de incerteza atualizado passa a ser representado por um novo modelo de probabilidade, denominado de *Distribuição a Posteriori*, e representado pela função $\pi(\boldsymbol{\theta}|\boldsymbol{X})$. Essa atualização é feita utilizando-se o Teorema de Bayes (Mood *et al.* [1974]):

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{L(\boldsymbol{\theta};\boldsymbol{X})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta};\boldsymbol{X})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

A distribuição *a priori* $\pi(\boldsymbol{\theta})$ que representa (probabilisticamente) o conhecimento prévio acerca do vetor de parâmetros $\boldsymbol{\theta}$, pode ser especificada de várias formas (ver Paulino *et al.* [2003] e O'Hagan [1994]). Aqui serão apresentadas os tipos de distribuições a prioir mais utilizadas: a *Distribuição a priori Conjugada* e a *Distribuição a priori Não-informativa*.

Nas Distribuições *a priori* Conjugadas, a distribuição de probabilidade *a priori* e a distribuição de probabilidade *a posteriori* pertencem a mesma classe de distribuições de probabilidade, na chamada *Classe de Distribuições Conjugadas*, envolvendo apenas uma mudança nos hiperparâmetros (parâmetros indexadores da classe de distribuições *a priori*). Diz-se que a distribuição *a priori* é *conjugada* para a distribuição de probabilidade que originou os dados amostrais.

Gamermam [1996] sugere prudência ao utilizar prioris conjugadas, por estas nem sempre representam adequadamente o conhecimento prévio do parâmetro.

Nas Distribuições a priori Não-informativas são usadas quando não há conhecimento prévio acerca do vetor de parâmetros $\boldsymbol{\theta}$. Com isso é atribuída uma opinião probabilística "vaga" para o vetor de parâmetros $\boldsymbol{\theta}$. A distribuição assume que todo valor de $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ ocorre com igual probabilidade em todo $\boldsymbol{\Theta}$.

A distribuição de probabilidade *a posteriori* de $\boldsymbol{\theta}$ pode ser apresentada levando-se em consideração apenas a parte que depende de $\boldsymbol{\theta}$, chamada de núcleo, e com isso $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ será proporcional à informação trazida pelos dados e ao conhecimento prévio do modelo

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) \propto L(\boldsymbol{\theta};\boldsymbol{X})\pi(\boldsymbol{\theta})$$

ou seja, $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ será apresentada a menos de $\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta};\boldsymbol{X})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$.

3.2.2 Aspectos Computacionais

O recente desenvolvimento dos Métodos de Monte Carlo tem eliminado muitas das dificuldades historicamente associadas com a análise Bayesiana de modelos não-lineares. A dificuldade está na grande maioria das vezes na determinação da distribuição à posteriori, $\pi(\boldsymbol{\theta}|\boldsymbol{X})$, objeto da Inferência Bayesiana, por esta envolver integrais geralmente intratáveis analiticamente.

Os métodos computacionais de Monte Carlo via Cadeias de Markov (MCMC) têm sido largamente utilizados em *Inferência Bayesiana*, pois possibilitam simular grandes amostras de uma determinada densidade à posteriori cuja determinação analítica é difícil de ser obtida. A implementação dos métodos MCMC só foi possível devido ao grande avanço tecnológico dos computadores, cada vez mais robustos e acessíveis, e devido ao trabalho de Gelfand e Smith [1990] (*apud* Paulino *et al.* [2003]). Segundo O'Hagan [1994] a necessidade computacional é essencial para o cálculo de características da distribuição a posteriori e, conforme Chib [1995], os trabalhos de Gelfand e Smith[1990] revolucionaram a *Inferência Bayesiana* no que toca à simulação da distribuição a posteriori, fazendo uso dos métodos MCMC.

A idéia básica dos métodos MCMC é construir uma cadeia de Markov com distribuição de equilíbrio dada pela distribuição posterior $\pi(\boldsymbol{\theta}|\boldsymbol{X})$, as cadeias de Markov utilizadas são as *ergódica*. Uma cadeia de Markov é Ergódica, se cada estado pode ser atingido a partir de qualquer outro com um número finito de iterações (*irredutível*), as probabilidades de transição de um estado para outro são invariantes (*homogênea*) e não possui estados absorventes (*aperiódica*) (ver Bernardo e Smith [1994] e Paulino *et al.* [2003]).

Conforme Bernardo e Smith [1994], O'Hagan [1994] e Paulino *et al.* [2003], após um número sufucientemente grande de iterações t, a cadeia converge para uma distribuição de equilíbrio (a posteriori), que pode ser usada para fazer inferências no modelo em estudo, isto é,

$$\boldsymbol{\theta}^{(t)} \to \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{X}) \quad \text{e} \quad \frac{1}{t}\sum_{i=1}^{t} g\left(\boldsymbol{\theta}^{(i)}\right) \to E_{\boldsymbol{\theta}|\boldsymbol{X}}\left(g(\boldsymbol{\theta})\right),$$

para $t \to \infty$.

Dentre os métodos MCMC está o Amostrador de Gibbs, proposto por Geman e Geman [1984] em uma aplicação de reconstrução de imagens, e difundido por Gelfand e Smith [1990] na simulação de distribuições à posteriori (Paulino *et al.* [2003]).

O Amostrador de Gibbs é um método de amostragem iterativo de uma cadeia de Markov, cuja transição de um estado a outro é feita a partir das distribuições condicionais completas (a partir de um vetor de parâmetros $\boldsymbol{\theta}$ à posteriori, define-se a condicional completa de um sub-vetor paramétrico genérico $\boldsymbol{\theta}$ como a distribuição deste, dado todos os outros parâmetros e os dados, que será denotado por $p\left(\theta_i|\boldsymbol{\theta}_{(-\theta_i)}, Y\right)$. A atualização feita pelo amostrador de Gibbs (Geman e Geman [1984]; Gelfand e Smith [1990]) é um caso particular do algoritmo de Metropolis-Hastings (Paulino *et al.* [2003]).

O algoritmo de Metropolis-Hasting gera o valor de uma distribuição proposta $q(.)$ e o aceita com uma dada probabilidade, nestas condições é garantida a convergência para a distribuição a posteriori $\pi(\boldsymbol{\theta}|\boldsymbol{X})$. Considere que a cadeia esteja no estado θ^t e um valor θ^* seja gerado da distribuição $q(.|\theta^t)$, o valor gerado será aceito com probabilidade,

$$\xi(\theta^t, \theta^*) = \min\left(1, \frac{\pi(\theta^*)q(\theta^t|\theta^*)}{\pi(\theta^t)q(\theta^*|\theta^t)}\right)$$

tal que π é o núcleo da distribuição a posteriori $\pi(\boldsymbol{\theta}|\boldsymbol{X})$.

O algoritmo de Metropolis-Hasting é estruturado da seguinte maneira:

Passo 1: Inicialize as iterações com $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_n^{(0)}\right)$ e tome $j = 1$;

Passo 2: Gere $\boldsymbol{\theta}^{(j)} = \left(\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_n^{(j)}\right)$ a partir da distribuição de $q(.|\theta^t)$;

Passo 3: Calcule a probabilidade de aceitação $\xi(\theta^t, \theta^{(j)})$;

Passo 4: Se o valor $\theta^{(j)}$ for aceito, faça $j = j + 1$ e $\theta^t = \theta^{(j)}$, caso contrário a cadeia permanece em θ^t e o processo reinicia a iteração a partir do passo 2.

Já o Amostrador de Gibbs é estruturado da seguinte forma: dado o conjunto de valores iniciais para os parâmetros em estudo, $\left(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_n^{(0)}\right)$, as amostras para cada parâmetro serão calculadas a partir dos seguintes passos:

Passo 1: $\theta_1^{(k)} \sim p_1\left(\theta_1|\theta_2^{(k-1)}, \theta_3^{(k-1)}, ..., \theta_n^{(k-1)}, Y\right)$

Passo 2: $\theta_2^{(k)} \sim p_2\left(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, ..., \theta_n^{(k-1)}, Y\right)$

Passo 3: $\theta_3^{(k)} \sim p_3\left(\theta_3|\theta_1^{(k)}, \theta_2^{(k)}, ..., \theta_n^{(k-1)}, Y\right)$

$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$

Passo n: $\theta_n^{(k)} \sim p_n\left(\theta_n|\theta_1^{(k)}, \theta_2^{(k)}, ..., \theta_{n-1}^{(k)}, Y\right)$

onde $p_n(\theta_i|\boldsymbol{\theta_{(-i)}})$ é a distribuição condicional completa.

Repita os passos 1, 2, 3, ..., n para k=1, 2, 3, ...

O Amostrador de Gibbs será usado neste trabalho por ser um bom algoritmo para a geração de amostras cuja função analítica é difícil de ser obtida, e também por já estar implementado em alguns pacotes de Inferência Bayesiana, como por exemplo, o Winbugs e o WinLim. Neste trabalho utilizou-se o Winbugs e o Matlab para estimação dos parâmetros de modelos de regressão logística dicotômica e policotômica a partir do algoritmo proposto por Groenewald e Mokgatlhe [2005].

3.2.3 Algoritmo de Groenewald e Mokgatlhe

Albert e Chib [1993] introduziram um processo de simulação baseado na aproximação do cálculo da distribuição à posteriori exata do vetor de parâmetros $\boldsymbol{\beta}$ do modelo de Regressão Logística, usando a função de ligação *probit*. Esta aproximação é baseada no *data augmentation*(Tanner e Wong [1987]), usando variável latente normalmente distribuida.

A proposta de Albert e Chib [1993] utiliza um modelo probit binário, tal que a variável observada $\boldsymbol{Y}$ é associada à uma variável latente $\boldsymbol{z}$ como sendo

$$z_i = \begin{cases} z_i < 0, & \text{se } y_i = 0 \\ z_i > 0, & \text{se } y_i = 1 \end{cases}$$

e eles mostram que $z_i|\boldsymbol{\beta}, \boldsymbol{\sigma^2}$ tem distribuição normal truncada.

A técnica de Groenewald e Mokgatlhe [2005] está baseada na proposta de aproximação da distribuição à posteriori de Albert e Chib [1993] para o *data augmentation* usando como função de ligação a *logit* e variável latente com distribuição uniforme para implementação do Amostrador de Gibbs (Geman e Geman [1984]; Gelfand e Smith [1990]).

A estimação para os modelos logísticos com variáveis dependentes dicotômicas e policotômicas (nominais e ordinais), utilizando a proposta de Groenewald e Mokgatlhe [2005], é feita a seguir.

i) Variável Resposta Dicotômica

As variáveis dicotômicas podem ser ajustadas ao modelo logístico para estimar a proporção de sucessos dado um conjunto de covariáveis, conforme (2.5),

$$\pi_i = \frac{\exp(\boldsymbol{\beta X_i})}{1 + \exp(\boldsymbol{\beta X_i})}.$$

É fácil ver que π_i é a distribuição acumulada de uma variável aleatória com distribuição logística:

$$\pi_i = \int_{-\infty}^{\boldsymbol{\beta X_i}} \frac{\exp(z)}{(1+\exp(z))^2} dz$$

A partir desta característica, Groenewald e Mokgatlhe [2005] usaram uma variável aleatória U com distribuição Uniforme(0,1) como variável latente, e usando o fato que

$$P(U \leq x) = \int_0^x du = x$$

tem-se que

$$\pi_i = \int_{-\infty}^{\boldsymbol{\beta X_i}} \frac{\exp(z)}{(1+\exp(z))^2} dz = P\left(U < \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}\right)$$

Sendo assim, a probabilidade de sucesso π_i estará relacionada à variável latente U, ao vetor de parâmetros $\boldsymbol{\beta}$ e às covariáveis $\boldsymbol{X}$.

A distribuição conjunta a posteriori de $\boldsymbol{\beta}$ e $\boldsymbol{u}$, dado $\boldsymbol{Y}$, poderá ser apresentada por

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{Y}\right) \propto \pi(\boldsymbol{\beta}) L(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{Y}) \tag{3.21}$$

tal que $\pi(\boldsymbol{\beta})$ é densidade da distribuição a priori de $\boldsymbol{\beta}$ e $L(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y})$ é a função de verossimilhança conjunta de $\boldsymbol{\beta}$ e $\boldsymbol{u}$, dado $\boldsymbol{Y}$. Segundo Groenewald e Mokgatlhe [2005] e Albert e Chib [1993], a função de verossimilhança conjunta é dada por

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{Y}) &= \prod_{i=1}^{n} \left[I\left(u_i \leq \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp \boldsymbol{\beta X_i}}\right) I(y_i = 1) \right. \\ &\quad \left. + I\left(u_i > \frac{\exp(\boldsymbol{\beta X})}{1+\exp \boldsymbol{\beta X_i}}\right) I(y_i = 0) \right] I(0 \leq u_i \leq 1) \end{aligned} \tag{3.22}$$

tal que $I(X \in A)$ é a função indicadora, que assume o valor 1 se $X \in A$ e 0 caso contrário.

Escrevendo a função de verossimilhança desta forma nota-se duas condições, descritas a seguir:

1) Se $y_i = 1$, segue que

$$I\left(u_i \leq \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}\right) I(0 < u_i < 1) = 1$$

implica que u_i terá 0 como limite inferior e $\dfrac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}$ como limite superior, ou seja,

$$u_i|\boldsymbol{\beta}, \boldsymbol{Y} \sim Uniforme\left(0, \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}\right)$$

que é a distribuição da variável latente caso $y_i = 1$.

Fazendo uso do fato que

$$I\left(u_i \leq \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp \boldsymbol{\beta X_i}}\right) = 1$$

segue que,

$$\boldsymbol{\beta X_i} > log\left(\frac{u_i}{1-u_i}\right)$$

onde $\boldsymbol{\beta X_i} = \sum_{k=0}^{p} \beta_k X_{ik}$ tem-se,

$$\beta_j \geq \frac{1}{x_{ij}}\left[\log\left(\frac{u_i}{1-u_i}\right) - \sum_{k\neq j}^{p} \beta_k X_{ik}\right]. \tag{3.23}$$

A desigualdade é garantida para todo i desde que $y_i = 1$ e $x_{ij} > 0$ ou $y_i = 0$ e $x_{ij} < 0$.

2) Se $y_i = 0$, segue que

$$I\left(u_i > \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}\right) I(0 < u_i < 1) = 1$$

implica que u_i terá $\dfrac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}$ como limite inferior e 1 como limite superior, ou seja,

$$u_i|\boldsymbol{\beta, Y} \sim Uniforme\left(\frac{\exp(\boldsymbol{\beta X_i})}{1+\exp(\boldsymbol{\beta X_i})}; 1\right)$$

que é a distribuição da variável latente caso $y_i = 0$.

Fazendo uso do fato que,

$$I\left(u_i > \frac{\exp(\boldsymbol{\beta X_i})}{1+\exp \boldsymbol{\beta X_i}}\right) = 1$$

segue que,

$$\boldsymbol{\beta X_i} < log\left(\frac{u_i}{1-u_i}\right)$$

onde $\boldsymbol{\beta X_i} = \sum_{k=0}^{p} \beta_k X_{ik}$ tem-se,

$$\beta_j < \frac{1}{x_{ij}}\left[\log\left(\frac{u_i}{1-u_i}\right) - \sum_{k\neq j}^{p} \beta_k X_{ik}\right]. \tag{3.24}$$

A desigualdade é garantida para todo i desde que $y_i = 0$ e $x_{ij} > 0$ ou $y_i = 1$ e $x_{ij} < 0$.

Com isso, são gerados dois conjuntos, definidos a seguir:

$$A_k = \{i : ((Y_i = 1) \cap (x_{ik} > 0)) \cup ((Y_i = 0) \cap (x_{ik} < 0))\}$$

e

$$B_k = \{i : ((Y_i = 0) \cap (x_{ik} > 0)) \cup ((Y_i = 1) \cap (x_{ik} < 0))\}.$$

Assumindo a priori para $\boldsymbol{\beta}$,$\pi(\boldsymbol{\beta}) \propto 1$, para $\boldsymbol{\beta}$, tem-se que para um determinado β_k:

$$\beta_k | \boldsymbol{\beta}_{(-k)}, u, y \sim Uniforme\,(a_k, b_k)$$

tal que,

$$a_k = \max_{i \in A_k} \left\{ \frac{1}{x_{ik}} \left[log\left(\frac{u_i}{1-u_i}\right) - \sum_{j \neq k}^{p} \beta_j x_{ij} \right] \right\}$$

e

$$b_k = \min_{i \in B_k} \left\{ \frac{1}{x_{ik}} \left[log\left(\frac{u_i}{1-u_i}\right) - \sum_{j \neq k}^{p} \beta_j x_{ij} \right] \right\}.$$

ii) Variável Resposta Policotômica

Para uma variável Y_{ij} policotômica nominal, ou seja, com distribuição Multinomial com r categorias, foi visto que o modelo Logístico Multinomial Nominal é dado por:

$$\pi_{ij} = \frac{\exp(\boldsymbol{\beta_j X_i})}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta_s X_i})}, \qquad j = 1, 2, \ldots, r.$$

Então, usando argumento similar ao caso dicotômico, tem-se que

$$\pi_{ij} = \frac{\exp(\boldsymbol{\beta_j X_i})}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta_s X}_i)} = P\left(U < \frac{\exp(\boldsymbol{\beta_j X_i})}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta_s X_i})} \right),$$

onde $U \sim Uniforme(0,1)$ e $U = \{u_{ij}\}$ de dimensão $n \times (r-1)$.

A distribuição conjunta a posteriori de $\boldsymbol{\beta, U|Y}$ será dada por:

$$\pi(\boldsymbol{\beta, U|Y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \sum_{j=1}^{r-1} \left[I\left(u_{ij} < \frac{\exp(\boldsymbol{\beta_j X_i})}{1 + \sum_{s=1}^{r-1} \exp(\boldsymbol{\beta_s X_i})} \right) I(y_i = j) \right] I(0 \leq u_{ij} \leq 1)$$

tal que para um dado $y_i = j$, segue que a variável latente $0 \leq u_{ij} \leq 1$ terá 0 como limite inferior e $\dfrac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}$ como limite superior, isto é

$$u_{ij}|\boldsymbol{\beta, Y} \sim Uniforme\left(0; \frac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}\right)$$

e para $y_i \neq j$ a variável latente terá $\dfrac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}$ como limite inferior e 1 como limite superior, isto é,

$$u_{ij}|\boldsymbol{\beta, Y} \sim Uniforme\left(\frac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}; 1\right).$$

Se $y_i = j$, da distribuição conjunta de $\pi(\boldsymbol{\beta, U|Y})$, segue que:

$$I\left(u_{ij} \leq \frac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}\right) I(0 \leq u_{ij} \leq 1) = 1$$

e tem-se que

$$u_{ij} \leq \frac{\exp(\boldsymbol{\beta_j X_i})}{1+\sum_{s=1}^{r-1}\exp(\boldsymbol{\beta_s X_i})}$$

logo,

$$\boldsymbol{\beta_j X_i} \geq \log\left[\frac{u_{ij}}{1-u_{ij}}\left(1+\sum_{s\neq j}^{r-1}\exp(\boldsymbol{\beta_s X_i})\right)\right]$$

e sendo $\boldsymbol{\beta_j X} = \sum_{k=0}^{p}\beta_{jk}X_{ik}$ segue que

$$\beta_{jt} \geq \frac{1}{x_{it}}\left\{\log\left[\frac{u_{ij}}{1-u_{ij}}\left(1+\sum_{s\neq j}^{p}\exp(\boldsymbol{\beta_s X_i})\right)\right] - \sum_{k\neq t}^{p}\beta_{jk}x_{ik}\right\} \tag{3.25}$$

de forma análoga, se $y_i \neq j$ tem-se

$$\beta_{jt} < \frac{1}{x_{it}}\left\{\log\left[\frac{u_{ij}}{1-u_{ij}}\left(1+\sum_{s\neq j}^{p}\exp(\boldsymbol{\beta_s X_i})\right)\right] - \sum_{k\neq t}^{p}\beta_{jk}x_{ik}\right\} \tag{3.26}$$

Para simplificar a notação constrói-se o conjunto Λ_{ijt} como sendo o conjunto que contém todos os elementos gerados por (3.25) e (3.26), dado por

$$\Lambda_{ijt} = \frac{1}{x_{ik}}\left\{log\left[\frac{u_{ij}}{1-u_{ij}}\left(1+\sum_{s\neq j}^{r-1}\exp\left(\sum_{k=0}^{p}\beta_{sk}x_{ik}\right)\right)\right] - \sum_{k\neq t}^{p}\beta_{jk}x_{ik}\right\}$$

A desigualdade é garantida, quando para todo i desde que $y_i = j$ e $x_{it} > 0$ ou $y_i \neq j$ e $x_{it} < 0$.

Com isso, serão gerados dois conjuntos, definidos a seguir:

$$A_{jt} = \{i : ((Y_i = j) \cap (x_{it} > 0)) \cup ((Y_i \neq j) \cap (x_{it} < 0))\}$$

e

$$B_{jt} = \{i : ((Y_i = j) \cap (x_{ik} < 0)) \cup ((Y_i \neq j) \cap (x_{it} > 0))\}$$

Sendo assim, a distribuição condicional de β_{jt} será dada por

$$\beta_{jt} \sim Uniforme(a_{jt}, b_{jt})$$

onde

$$a_{jt} = \max_{i \in A_{jt}} \Lambda_{ijt} \text{ e } b_{jt} = \min_{i \in B_{jt}} \Lambda_{ijt}.$$

No caso Ordinal, o modelo logístico é ajustado para a probabilidade acumulada na categoria. Assim para a categoria j tem-se:

$$\eta_{ij} = \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}.$$

Então, ao ser introduzida a variável latente U, uniformemente distribuída em [0, 1], tem-se que

$$\eta_{ij} = P\left(U_i < \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}\right) = \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}$$

Nestas condições segue que a distribuição conjunta de $\boldsymbol{\alpha, \beta, u|y}$ é dada por

$$\pi(\boldsymbol{\alpha, \beta, u|y}) \propto \pi(\boldsymbol{\alpha, \beta}) \prod_{i=1}^{n} \left\{ \sum_{j=1}^{r-1} I(y_i = j) I(\eta_{ij-1} < u_i \leq \eta_{ij}) \right\} I(0 \leq u_i \leq 1)$$

A partir da distribuição conjunta de $\boldsymbol{\alpha, \beta, u|Y}$ tem-se que a variável latente u_i será denotada por

$$\boldsymbol{u_i, \alpha, \beta}|y_i = j \sim Uniforme(\eta_{ij-1}; \eta_{ij}), \qquad i = 1, 2, \ldots, n.$$

Segue que

$$\frac{\exp(\alpha_{j-1} + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta X_i})} \leq u_i \leq \frac{\exp(\alpha_j + \boldsymbol{\beta X_i})}{1 + \exp(\alpha_j + \boldsymbol{\beta X_i})}$$

e sabendo que $\boldsymbol{\beta X_i} = \sum_{s=1}^{p} \beta_s x_{is}$ conclui-se que

$$\beta_t \leq \frac{1}{x_{it}} \left[\log\left(\frac{u_i}{1-u_i}\right) - \alpha_{j-1} - \sum_{s \neq t}^{p} \beta_s x_{is} \right]$$

e

$$\beta_t \geq \frac{1}{x_{it}} \left[\log\left(\frac{u_i}{1-u_i}\right) - \alpha_{j} - \sum_{s \neq t}^{p} \beta_s x_{is} \right]$$

Assim, como nos modelos dicotômico e policotômico nominal serão formados dois conjuntos, A_j e B_j, a partir da distribuição conjunta de $\boldsymbol{\alpha, \beta, u|y}$, denotado por:

$$H_{ijt} = \frac{1}{x_{it}} \left[\log\left(\frac{u_i}{1-u_i}\right) - \alpha_{j} - \sum_{s \neq t}^{p} \beta_s x_{is} \right]$$

sendo assim $A_j = \{i : y_i = j\}$ e $\beta_{(t)}|\boldsymbol{\beta}_{(-t)}, \boldsymbol{\alpha}, u, y \sim U\left(a_t, b_t\right)$, $t = 1, 2, 3, ..., p$ com $a_t < b_t$, tal que

$$a_t = \max_j \left\{ \max_{i \in A} \left[\min\left(H_{ij-1t}, H_{ijt}\right) \right] \right\} \quad \text{e} \quad b_t = \min_j \left\{ \min_{i \in A} \left[\max\left(H_{ij-1t}, H_{ijt}\right) \right] \right\}.$$

Já para os interceptos, temos que a condição $u_i \leq \eta_{ij}$ para todo $i \in A_j$ e $u_i > \eta_{ij}$ para todo $i \in A_{j+1}$ nos dará $\alpha_{j-1} < \alpha_j < \alpha_{j+1}$, tal que a distribuição condicional de $\alpha_j|\boldsymbol{\alpha}_{(-j)}, \boldsymbol{\beta, u, y} \sim Uniforme(c_j, d_j)$, $j = 1, 2, 3, ..., r-1$ sendo,

$$c_j = \max_{i \in A_{j+1}} \left[\max\left(\log \frac{u_i}{1-u_i} - \boldsymbol{\beta X}, \alpha_{j-1} \right) \right]$$

e

$$d_j = \min_{i \in A_j} \left[\min\left(\log \frac{u_i}{1-u_i} - \boldsymbol{\beta X}, \alpha_{j+1} \right) \right].$$

3.3 Interpretação dos Parâmetros

A interpretação dos parâmetros estimados nos modelos de Regressão Logística é diferenciada dos modelos usuais de Regressão. Estes parâmetros são interpretados fazendo uso da razão de chances(OR) (*odds ratio*) para cada categoria.

A razão de chances é um número não negativo, sendo tomado $OR = 1$ como base

para comparação. Se $OR = 1$ indica que a variável resposta e a preditora não estão associadas, se $OR > 1$ indica que a probabilidade de pertencer a uma dada categoria frente ao nível de referência é grande e se $OR < 1$ a razão de chances indica que o sucesso de uma dada categoria frente ao nível de referência é pequeno.

Para o modelo de Regressão Logística Dicotômico, a chance de sucesso é dada por

$$\frac{P(Y_i = 1|X_{ip})}{P(Y_i = 0|X_{ip})} = \frac{\pi_i}{1 - \pi_i} = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}),$$

ao acrescentar uma unidade ao nível de uma dada preditora X_{ip} e tomando todas as outras covariadas como constantes, tem-se que a chance de sucesso será dada por

$$\frac{P(Y_i = 1|X_{ip} + 1)}{P(Y_i = 0|X_{ip} + 1)} = \frac{\pi_i}{1 - \pi_i} = \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p (X_{ip} + 1)),$$

a razão entre as chances acrescentadas de uma unidade na preditora X_{ip} e chance em X_{ip} será dada por

$$\frac{\dfrac{P(Y_i = 1|X_{ip} + 1)}{P(Y_i = 0|X_{ip} + 1)}}{\dfrac{P(Y_i = 1|X_{ip})}{P(Y_i = 0|X_{ip})}} = \exp(\beta_p). \tag{3.27}$$

A relação exponencial indica que para o incremento de uma unidade em X_{ip}, a chance é multiplicada por $\exp(\beta_p)$.

No modelo de Regressão Logística Multinomial Nominal com $r (r > 2)$ categorias, sendo a r-ésima categoria tomada como nível de referência, a razão de chances da j-ésima categoria em relação à r-ésima categoria, para o valor da covariada $X_{it} = a$ versus $X_{it} = b$ é dada por

$$OR_j(a, b) = \frac{\dfrac{P(Y = j|X_{it} = a)}{P(Y = r|X_{it} = a)}}{\dfrac{P(Y = j|X_{it} = b)}{P(Y = r|X_{it} = b)}}. \tag{3.28}$$

Os parâmetros no modelo de Regressão Logística Multinomial Ordinal são interpretados de forma similar ao modelo dicotômico, sendo que cada componente do vetor de parâmetros $\hat{\boldsymbol{\beta}}$ descreve o efeito da covariada relacionada ao parâmetro na categoria j. Como o modelo de Regressão Logística Multinomial Ordinal assume que o efeito é idêntico para todas as $r - 1$ categorias, então a razão de chances utiliza as probabilidades

acumuladas nas categorias e seus respectivos complementos, e será dada por:

$$OR_j(a,b) = \frac{\dfrac{P(Y \leq j|X_{it} = a)}{P(Y > j|X_{it} = a)}}{\dfrac{P(Y \leq j|X_{it} = b)}{P(Y > j|X_{it} = b)}}. \tag{3.29}$$

Capítulo 4

Seleção e Validação do Modelo

4.1 Seleção

A comparação de modelos serve para a escolha do modelo que melhor descreve o fenômeno em estudo. Assim, deve-se considerar todos os possíveis modelos e adotar o mais adequado, levando em consideração os aspectos computacionais, o principio da parcimônia, os resíduos gerados por este modelo e a previsão. Para isso, deve-se adotar testes estatísticos e verificar quais parâmetros e variáveis são realmente significativos para o modelo.

Neste trabalho a seleção do modelo será feita através do uso de métodos bayesianos, o $\boldsymbol{FB_{ij}}$ *(Fator de Bayes)*, o $\boldsymbol{BIC}$ (*Bayesian Information Criterion*) e com a proposta de Pereira e Stern [1999], o $\boldsymbol{FBST}$ *(Full Bayesian Significance Test).*

4.1.1 Fator de Bayes

Segundo Kass e Raftery [1995], o Fator de Bayes é um critério baseado na comparação das verossimilhanças marginais.

Definição 1. *Sejam duas hipóteses H_0 e H_1, correspondentes aos modelos, M_0 e M_1, respectivamente. Para os dados $\boldsymbol{Y}$, o Fator de Bayes a favor de H_0 é dado como a razão de chances da posteriori para a priori.*

$$FB_{01}(\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|M_0)}{p(\boldsymbol{Y}|M_1)}$$

onde,

$$p(\boldsymbol{Y}|M_k) = \int_{\Theta_k} L(\boldsymbol{\theta_k}; \boldsymbol{Y}, M_k)\pi(\boldsymbol{\theta_k})d\theta_k, \qquad k = 0, 1 \tag{4.1}$$

é a verossimilhança marginal do modelo M_k.

Na maioria das vezes $p(\boldsymbol{Y}|M_k)$ é muito difícil de ser calculada (Paulino *et al.* [2003]),

sendo necessário adotar métodos numéricos para sua resolução, como por exemplo, Métodos de Monte Carlo.

A verossimilhança marginal neste trabalho será determinada segundo Chib [1995], esta faz uso da definição da identidade básica da verossimilhança marginal

$$P(\boldsymbol{Y}|M_k) = \frac{L(\boldsymbol{\theta^*; Y})\pi(\boldsymbol{\theta^*})}{\pi(\boldsymbol{\theta^*|Y})}$$

tal que: $L(\boldsymbol{\theta^*; Y})$ é a verossimilhança do modelo em $\boldsymbol{\theta^*}$, $\pi(\boldsymbol{\theta^*})$ é a distribuição a priori em $\boldsymbol{\theta^*}$ e $\pi(\boldsymbol{\theta^*|Y})$ é a distribuição a posteriori em $\boldsymbol{\theta^*}$, sendo $\boldsymbol{\theta^*}$ o vetor de parâmetros estimados.

A verossimilhança do modelo e a priori são determimadas facilmente, dado $\boldsymbol{\theta^*}$. A posteriori do modelo será determinada reescrevendo-a de forma ordenada, chamada de "*posteriori ordenada*", $\pi(\boldsymbol{\theta^*|Y})$, escrita com as densidades condicionais completas, $\pi(\theta_j|\boldsymbol{\theta_{(-j)}, Y})$, $j = 0, 1, 2, \ldots, p_t$. Sendo $\boldsymbol{\theta^*} = (\beta_0^*, \beta_1^*, \ldots, \beta_{p_t}^*)$, tem-se:

$$\pi(\boldsymbol{\theta^*|Y}) = \pi(\beta_0^*|\boldsymbol{Y}).\pi(\beta_1^*|\beta_0^*, \boldsymbol{Y}) \ldots \pi(\beta_{p_t}^*|\beta_{p_t-1}^*, \ldots, \beta_0^*, \boldsymbol{Y}). \tag{4.2}$$

Cada um dos fatores de $\pi(\boldsymbol{\theta^*|Y})$ poderá ser determinado por

$$\pi(\theta_r^*|\boldsymbol{Y}, \boldsymbol{\theta_{s(s<r)}}) = \frac{1}{M}\sum_{j=1}^{M} \pi\left(\theta_r^*|\boldsymbol{Y}, \boldsymbol{\theta_{(-r)}^{(j)}}, u^{(j)}\right).$$

Conforme Kass e Raftery [1995] o Fator de Bayes sofre influência das prioris adotadas, sendo sugerido a adoção de prioris proprias.

O cálculo do Fator de Bayes neste trabalho será feito segundo a proposta de Groenewald e Mokgatlhe [2005] que utilizaram prioris logísticas com média zero e parâmetro de escala σ, para cada parâmetro, ou seja,

$$\pi(\beta_t|\sigma) = \frac{\exp\left(\dfrac{\beta_t}{\sigma}\right)}{\sigma\left[1 + \exp\left(\dfrac{\beta_t}{\sigma}\right)\right]^2} \tag{4.3}$$

e priori

$$\pi(\sigma) \propto \frac{1}{\sigma}. \tag{4.4}$$

para o parâmetro de escala.

Na secção 3.2, foi visto que cada parâmetro tem distribuição uniforme dada por

$$\beta_t|\boldsymbol{\beta}_{(-t)}, \boldsymbol{u}, \boldsymbol{Y} \sim Uniforme(a_t, b_t) \tag{4.5}$$

com $\pi(\beta_t) \propto 1$. Para utilização das prioris (4.3) e (4.4) é necessário que se faça uma transformação conveniente na função densidade de probabilidade de β_t.

Tendo que,

$$a_t < \beta_t < b_t \tag{4.6}$$

e $\sigma > 0$. Dividindo (4.5) por $\sigma > 0$, segue que:

$$\frac{a_t}{\sigma} < \frac{\beta_t}{\sigma} < \frac{b_t}{\sigma}. \tag{4.7}$$

Aplicando a função exponencial em (4.6), a desigualdade continua válida, ou seja,

$$\exp\left(\frac{a_t}{\sigma}\right) < \exp\left(\frac{\beta_t}{\sigma}\right) < \exp\left(\frac{b_t}{\sigma}\right) \tag{4.8}$$

Como a função exponencial é maior que zero e fazendo uso de propriedades de desigualdades para números maiores que zero, pode-se ter

$$\frac{\exp\left(\frac{a_t}{\sigma}\right)}{1+\exp\left(\frac{a_t}{\sigma}\right)} < \frac{\exp\left(\frac{\beta_t}{\sigma}\right)}{1+\exp\left(\frac{\beta_t}{\sigma}\right)} < \frac{\exp\left(\frac{b_t}{\sigma}\right)}{1+\exp\left(\frac{b_t}{\sigma}\right)} \tag{4.9}$$

denotando

$$v_t = \frac{\exp\left(\frac{\beta_t}{\sigma}\right)}{1+\exp\left(\frac{\beta_t}{\sigma}\right)} \tag{4.10}$$

segue que

$$\frac{\exp\left(\frac{a_t}{\sigma}\right)}{1+\exp\left(\frac{a_t}{\sigma}\right)} < v_t < \frac{\exp\left(\frac{b_t}{\sigma}\right)}{1+\exp\left(\frac{b_t}{\sigma}\right)}. \tag{4.11}$$

Sendo assim, tem-se que:

$$v_t \sim Uniforme\left(\frac{\exp\left(\frac{a_t}{\sigma}\right)}{1+\exp\left(\frac{a_t}{\sigma}\right)}; \frac{\exp\left(\frac{b_t}{\sigma}\right)}{1+\exp\left(\frac{b_t}{\sigma}\right)}\right). \tag{4.12}$$

De (4.9) tem-se que

$$\beta_t = -\sigma \log\left(\frac{1-v_t}{v_t}\right). \tag{4.13}$$

A função densidade de cada β_t é facilmente determinada fazendo uso de transformação de variáveis aleatórias (ver James [1981]), a posteriori de cada parâmetro β_t será dada por

$$f(\beta_t|\boldsymbol{\beta}_{(-t)}, \boldsymbol{u}, \boldsymbol{Y}) = \left[\frac{(1+\exp(a_t/\sigma))(1+\exp(b_t/\sigma))}{\exp(b_t/\sigma)-\exp(a_t/\sigma)}\right] \frac{\exp(\beta_t/\sigma)}{\sigma\left(1+\exp(\beta_t/\sigma)\right)^2} \tag{4.14}$$

e cada um dos fatores em (4.2) poderá ser estimado determinando a média de (4.14).

O Fator de Bayes, FB_{01}, é interpretado frequentemente como a vantagem do modelo M_0 contra M_1, trazida pelos dados (Berger e Pericchi [1997]), sendo escolhido o modelo que apresentar maior valor de FB_{01} entre os pares de modelos concorrentes. Em Kass e Raftery [1995] é sugerida a interpretação do Fator de Bayes por meio do $\log(FB_{01})$, descrita na Tabela 4.1 seguir:

Tabela 4.1 *Interpretação do Fator de Bayes*

$2\log(FB_{ij})$	FB_{ij}	Evidência contra H_j
0 a 2	1 a 3	Inconclusiva
2 a 6	3 a 20	Significativa
6 a 10	20 a 150	Forte
> 10	> 150	Decisiva

Fonte: Kass e Raftery [1995]

4.1.2 BIC

O BIC (*Bayesian Information Criterion*), Critério de Informação Bayesiano, faz a comparação entre as verossimilhanças a posteriori levando em consideração a complexidade do modelo no critério de seleção (Paulino *et al.* [2003]).

Definição 2. *Sejam duas hipóteses H_0 e H_1, correspondentes aos modelos, M_0 e M_1, respectivamente. Dado os dados $\boldsymbol{Y}$, o $\boldsymbol{BIC}$ a favor de M_1 é dado por*

$$\Delta BIC = -2\log\left[\frac{\sup_{M_0} L(\boldsymbol{\theta_0}; \boldsymbol{Y}, M_0)}{\sup_{M_1} L(\boldsymbol{\theta_1}; \boldsymbol{Y}, M_1)}\right] - (p_0 - p_1)\log n$$

onde $\boldsymbol{\theta_0}$ e $\boldsymbol{\theta_1}$ são, respectivamente, os vetores de parâmetros dos modelos M_o e M_1, n é o tamanho da amostra e p_i, $i = 0, 1$ é o número de parâmetros de cada modelo.

Segundo Paulino *et al.* [2003], Schwarz [1978] mostrou que, para grandes amostras, ΔBIC aproxima satisfatoriamente $-2\log BF_{01}$.

Carlin e Louis [2000] (*apud* Paulino *et al.* [2003]) sugerem a modificação do ΔBIC, calculando para cada modelo M_i em competição

$$\hat{BIC} = 2E[\log L(\boldsymbol{\theta_i}; \boldsymbol{Y}, M_i)] - p_i \log n,$$

e o modelo escolhido será o que apresentar maior valor de $\hat{BIC}$.

4.1.3 FBST

O **FBST (*Full Bayesian Significance Test*)** é um teste de significância Bayesiano proposto por Pereira e Stern [1999], que se baseia no cálculo da probabilidade da Região $\boldsymbol{HPD}$ (*Highest Posteriori Density*) tangente ao conjunto que define a hipótese nula. A Evidência Bayesiana em favor da hipótese nula é o complementar da probabilidade da Região $\boldsymbol{HPD}$(Pereira e Stern [1999], Madruga *et al.* [2001]), as regiões $\boldsymbol{HPD}$ são interpretadas como regiões fixadas que contém o parâmetro aleatório com determinada probabilidade. A definição do FBST é dada a seguir:

Definição 3. *Seja $\pi(\theta|\boldsymbol{X})$ uma densidade posterior de θ, dada a amostra $\boldsymbol{X}$, e considere o conjunto $T(\boldsymbol{X})$ definido no espaço paramétrico $\boldsymbol{\Theta}$, com $T(\boldsymbol{X}) = \{\theta \in \boldsymbol{\Theta} : \pi(\theta|\boldsymbol{X}) > \sup_{\boldsymbol{\Theta}_0} \pi(\theta|\boldsymbol{X})\}$. A medida de evidência Bayesiana de Pereira-Stern é definida como*

$$EV(\boldsymbol{\Theta}_0; \boldsymbol{X}) = 1 - P(\theta \in T(\boldsymbol{X})|\boldsymbol{X}) \tag{4.15}$$

e um teste (procedimento) de Pereira-Stern é aceitar H_0 sempre que a $EV(\boldsymbol{\Theta}_0; \boldsymbol{X})$ for "grande".

Segundo Pereira e Stern [1999], um valor grande de $EV(\boldsymbol{\Theta_0}, X)$ significa que o subconjunto $\boldsymbol{\Theta_0}$ cai em uma região do espaço paramétrico de alta probabilidade, ou seja, os dados favorecem a hipótese nula; sendo assim um valor pequeno de $EV(\boldsymbol{\Theta_0}, X)$ indica que $\boldsymbol{\Theta_0}$ está em uma região do espaço paramétrico de baixa probabilidade posterior, logo os dados não trazem evidências a favor da hipótese nula.

Assim, como no Fator de Bayes, as duas hipóteses H_0 e H_1, correspondem aos modelos M_0 e M_1, respectivamente. A medida de evidência Bayesiana do procedimento FBST em favor de $M_1(H_1)$ é o complementar da medida de evidência Bayesiana do procedimento FBST em favor de $M_0(H_0)$, ou seja, $EV(\boldsymbol{\Theta}_0; \boldsymbol{X}) = 1 - EV(\boldsymbol{\Theta}_1; \boldsymbol{X})$.

O FBST considera igualmente a hipótese alternativa frente a hipótese nula, de modo que aumentando o tamanho da amostra não somos levados a rejeitar a hipótese, mas sim a convergir para a decisão correta (rejeitar ou aceitar).

Para o cálculo da $EV(\boldsymbol{\Theta}_0; \boldsymbol{X})$, caso não seja possível analiticamente, utiliza-se a aproximação por Método de Monte Carlo, isto é,

$$EV(\boldsymbol{\Theta}_0; \boldsymbol{X}) \approx 1 - \frac{1}{M} \sum_{j=1}^{M} h(\theta_j) \tag{4.16}$$

com,

$$h(\theta) = I(\theta \in T(\boldsymbol{X}))$$

e

$$T(\boldsymbol{X}) = \{\theta \in \boldsymbol{\Theta} : \pi(\theta|\boldsymbol{X}) > sup_{\boldsymbol{\Theta}_0} \pi(\theta|\boldsymbol{X})\}$$

A distribuição a posteriori do modelo será determinda levando em consideração prioris *não informativas* para os parâmetros, $\pi(\boldsymbol{\beta}) \propto 1$. Logo a distribuição a posteriori será dada por

$$\pi(\boldsymbol{\beta}; \boldsymbol{Y}) \propto L(\boldsymbol{\beta}; \boldsymbol{Y}) \pi(\boldsymbol{\beta}) \propto L(\boldsymbol{\beta}; \boldsymbol{Y}).$$

4.2 Validação

A validação de um modelo visa garantir que os resultados gerados por este modelo sejam significantes na amostra, significantes no sentido de haver uma proporção alta de acerto deste modelo na classificação de suas estimativas. Neste trabalho será utilizada a *validação cruzada* definida a seguir, segundo Hair *et al.* [2005].

Definição 4. *A Validação cruzada divide a amostra em duas partes: a amostra de estimação, usada na estimação dos parâmetros nos Modelos de Regressão Logística e a amostra de validação, usada para verificar a correspondência entre as estimativas geradas pelo modelo e a amostra de validação.*

Para validar o modelo deve-se decidir a qual categoria pertence a proporção estimada pelo modelo, que será feita com o uso de uma idéia puramente estatística, baseada nas probabilidades de classificação incorreta. A idéia é minimizar as probabilidades de classificação incorreta, ou seja, minimizar a chance de classificar uma observação como sendo de

uma dada categoria sendo que é de outra, ou vice-versa. Para a tomada de decisão sobre a pertinência da categoria constrói-se uma tabela que estima qual será o comportamento do modelo ajustado caso se adote determinada proporção como decisão de pertencer a uma categoria, chamada de *ponto de corte*.

Em resumo, a validação é feita para verificar a concordância entre as estimativas geradas pelo modelo ajustado e a amostra de validação, fazendo uso do ponto de corte.

Capítulo 5

Aplicações

5.1 Regressão Logística Dicotômica

5.1.1 Aplicação 1: Besouros expostos ao CS_2

Um conjunto de dados clássico de modelos de dose-resposta encontra-se em Bliss [1935] (*apud* Paulino *et al.* [2003]), e baseia-se no comportamento de besouros adultos face a exposição à dissulfeto de carbono (CS_2) durante 5 horas. A curva de dose-resposta da mortalidade dos besouros foi formada a partir de 8 dosagens, e os respectivos dados encontram-se na Tabela 5.1, onde as três colunas correspondem, respectivamente, ao número de besouros observados (n_i), ao número de besouros mortos r_i e ao log de cada dosagem de CS_2, $i = 1, 2, ..., 8$.

Tabela 5.1 *Mortalidade de Besouros*

n_i	r_i	$\log(Dose_i)$	t_i
59	6	1,6907	5,42
60	13	1,7242	5,61
62	18	1,7552	5,78
56	28	1,7842	5,95
63	52	1,8113	6,12
59	53	1,8369	6,28
62	61	1,8610	6,43
60	60	1,8839	6,58

Fonte: Bliss [1935](*apud* Paulino *et al.* [2003])

Nesta aplicação será ajustado um modelo logístico para estimar proporção de besouros mortos como função da dose de CS_2 sofrida. Será usada a metodologia de estimação dos parâmetros desenvolvida na Secção 3.2 para os dados de Bliss [1935], sendo que não houve a necessidade de ser aplicada o logaritmo da dose, utilizado em Paulino *et al.* [2003], sendo utilizado os valores aproximadamente reais das doses, obtidos através de uma transformação exponencial do log da dose, isto é, $t_i = \exp(\log(Dose_i))$. A transformação foi

necessária devido ao modelo ajustado com a metodologia adotada ao utilizar a covariada como apresentado em Bliss [1935] não ser satisfatória, produzindo erros muito grandes. O modelo ajustado é apresentado a seguir:

$$\hat{\pi}_i = \frac{\exp(-34,3367 + 5,8321t_i)}{1 + \exp(-34,3367 + 5,8321t_i)}. \tag{5.1}$$

A razão de chances do parâmetro da covariada e intervalos de credibilidade (ver Paulino *et al.* [2003]) são dados na Tabela 5.2 a seguir:

Tabela 5.2 *Razão de Chance(OR) e Intervalo de Credibilidade de 95%*

Parâmetros	IC(θ,95%)	OR
α	$(-35,160;\ -33,512)$	–
β	(5,828; 5,837)	350,2335

O incremento de uma unidade na dose indica que a chance de um besouro ser morto quando exposto ao CS_2 aumenta em torno de 350 vezes, dando indicios que o CS_2 é eficaz no controle da população de besouros.

A Figura 5.1 mostra o gráfico do modelo de dose-resposta ajustado para os dados de Bliss [1935].

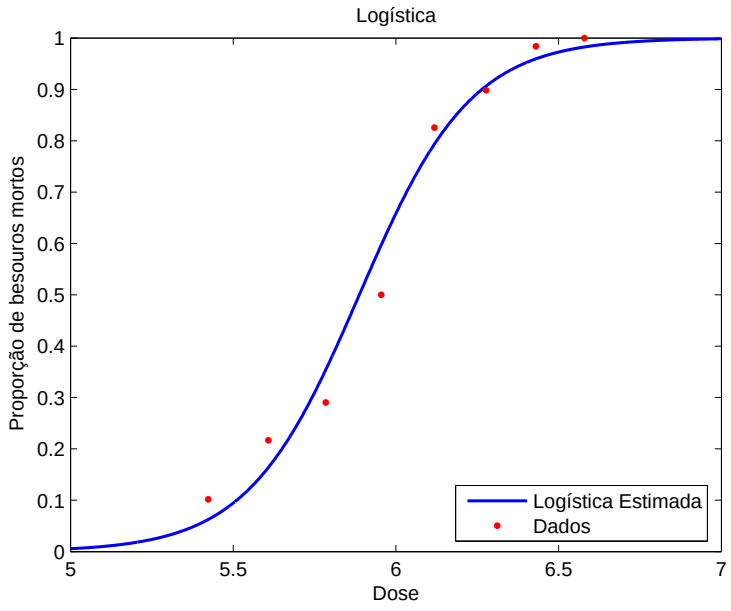

Figura 5.1 *Proporção de Besouros mortos expostos à CS_2*

As Figuras 5.2 apresenta aproximadamente a densidade a posteriori conjunta dos parâmetros e indica que a mesma está condensada em uma pequena região, indicando sua pouca variabilidade e reduzindo de forma significativa a pouca informação inicial, representada pela distribuição a priori não-informativa, este indício é reforçado pela amplitude dos intervalos de credibilidade dos parâmetros dispostos na Tabela 5.2.

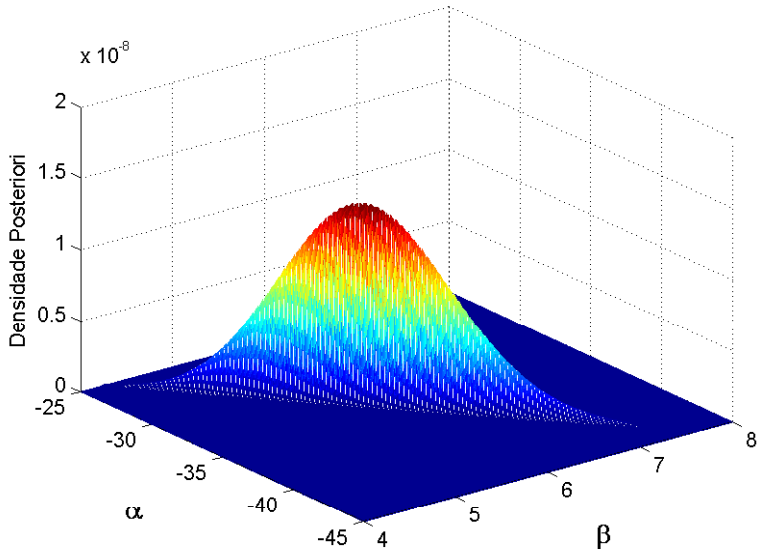

Figura 5.2 *Densidade Posteriori Conjunta dos Parâmetros*

A convergência para a distribuição de equilíbrio dos parâmetros, a posteriori, foi lenta e necessitou de um número alto de iterações, sendo as cadeias de α e β apresentadas nas Figuras 5.3 e 5.4, respectivamente.

Foi ajustado o modelo de regressão logística, segundo a metodologia clássica, para os dados de Bliss [1935], desenvolvida na Secção 3.1.1, e obteve-se

$$\hat{\pi}_i = \frac{\exp(-60.7175 + 34.2703\log(Dose_i))}{1 + \exp(-60.7175 + 34.2703\log(Dose_i))}. \tag{5.2}$$

O gráfico do modelo de regressão logística clássico ajustado é apresentado na Figura 5.5, indicando haver um bom ajuste inicial para o $\log(Dose_i)$, mas sugerindo o uso da função de ligação *complemento* $\log - \log$, devido a maior concentração dos dados em um dos extremos da sigmóide.

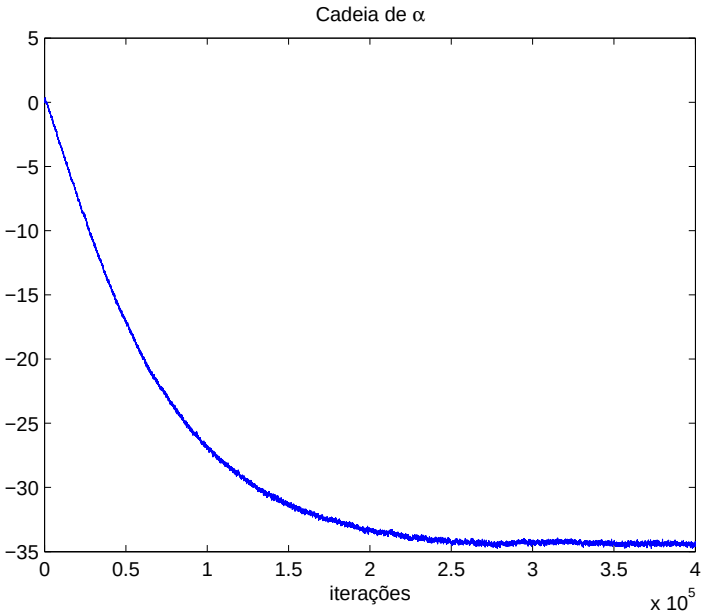

Figura 5.3 *Convergência da cadeia para distribuição de equilíbrio de* α

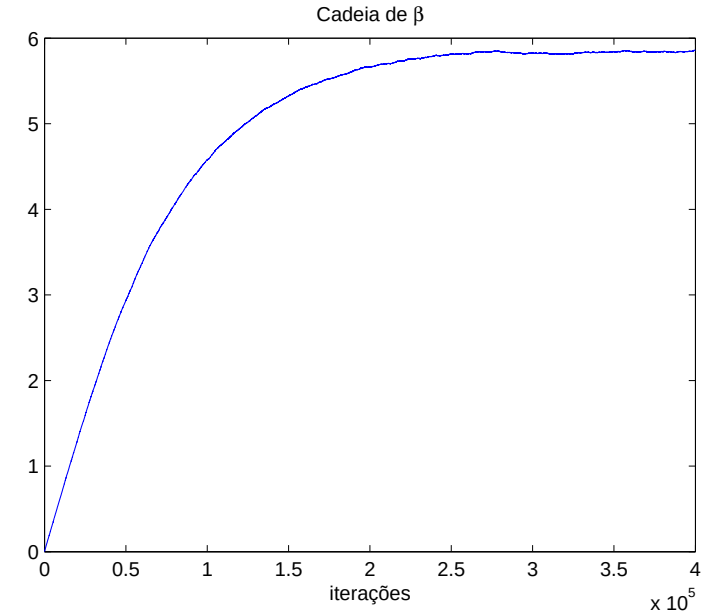

Figura 5.4 *Convergência da cadeia para distribuição de equilíbrio de* β

5.1.2 Aplicação 2: Falência de Empresas

Nesta aplicação são usados os dados de Johnson [1987], que foram coletados de 21 empresas, aproximadamente dois anos antes de suas falências, e de outras 25 empresas que não faliram no mesmo período. As variáveis observadas foram: X_1 (fluxo de caixa/total de débitos); X_2 (rendimento da empresa/total de patrimônio); X_3 (patrimônio atual/total

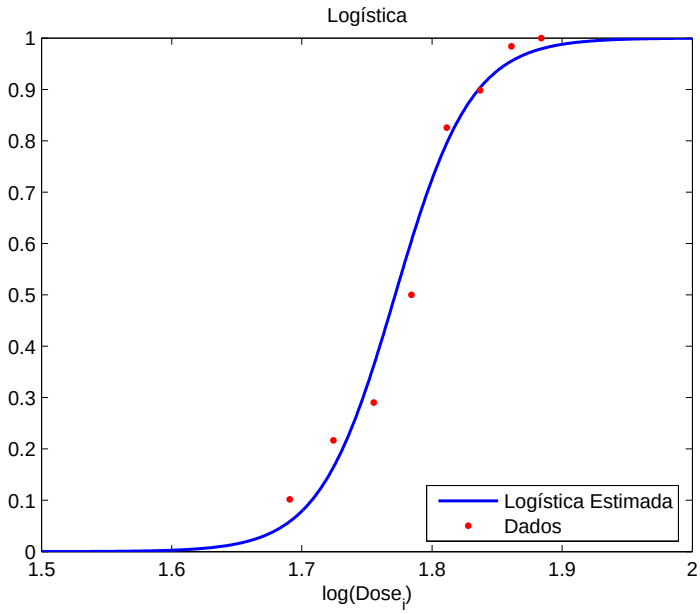

Figura 5.5 *Estimação Clássica da Proporção de Besouros mortos expostos à CS_2*

de débito), X_4 (patrimônio atual/rendimento das vendas) e Y(0 se a empresa faliu e 1 se a empresa não faliu).

Foram ajustados os modelos mais simples, sem combinação ou transformação das covariadas, que utilizam as covariadas mais o intercepto a partir da metodologia desenvolvida na Secção 3.2. Sendo que a seleção foi feita comparando 14 modelos frente ao modelo completo. Os valores de $2\log(FB_{ij})$, BIC e $FBST$ estão dispostos na Tabela 5.2.

Os critérios de seleção adotados sugerem que o modelo selecionado é o que contém a covariada X_3 (patrimônio atual/total de débito) mais o intercepto, quando comparados com o modelo completo. Devido apresentarem maiores valores no FB e BIC, o FBST seleciona o modelo quando seu valor é grande. O modelo é apresentado a seguir

$$\hat{\pi}_i = \frac{\exp(-7,5646 + 4,1221X_{3i})}{1 + \exp(-7,5646 + 4,1221X_{3i})}. \tag{5.3}$$

Devido ao uso de distribuições logísticas, com uso dos parâmetros estimados, percebe-se no Fator de Bayes uma certa influência na seleção do modelo que utiliza esta metodologia, já que o Fator de Bayes é influenciado pelas prioris adotadas (Kass e Raftery [1995]). O BIC, por depender da dimensão do espaço paramétrico e do tamanho da amostra, não apresentou influência na seleção do modelo nesta aplicação devido à dimensão do espaço paramétrico ser relativamente pequeno ($\leq$ 5) e o tamanho da amostra de estimação também ser pequena (n=40), já o FBST foi coerente com a seleção feita pelo FB

Tabela 5.3 *Seleção do Modelo*

$Modelos$	$2\log(FB_{ij})$	BIC	$FBST$
X_1	19,2014	$-360,4486$	0,0000
X_2	7,9181	$-54,0047$	0,0000
$\boldsymbol{X_3}$	$\mathbf{21,0373}$	$\mathbf{-35,3312}$	$\mathbf{0,4677}$
X_4	15,4131	$-61,8432$	0,0000
X_1, X_2	17,2656	$-50,4516$	0,0000
X_1, X_3	12,3253	$-37,6263$	0,9730
X_1, X_4	13,6188	$-49,7530$	0,0000
X_2, X_3	3,4773	$-39,3019$	0,3685
X_2, X_4	$-3,8175$	$-56,8475$	0,0000
X_3, X_4	14,2759	$-38,8062$	0,6078
X_1, X_2, X_3	6,6920	$-42,5436$	0,5912
X_1, X_2, X_4	7,2780	$-53,0974$	0,0000
X_1, X_3, X_4	4,8980	$-41,5744$	0,9197
X_2, X_3, X_4	$-3,0016$	$-43,0074$	0,3822
X_1, X_2, X_3, X_4	$-$	$-46,4832$	$-$

e BIC, dependendo somente da distribuição a posteriori e selecionando dentre os modelos, o melhor, mas sendo feita esta seleção em conjunto com outras técnicas de seleção como em Pereira e Stern [2001].

A Tabela 5.4 apresenta os intervalos de credibilidade de 95% e a razão de chances (OR) de β segundo a metodologia Bayesiana adotada.

Tabela 5.4 *Razão de Chance(OR) e Intervalo de Credibilidade de 95%*

Parâmetros	IC(θ,95%)	OR
α	$(-7,801;\ -7,328)$	$-$
β	$(3,983;\ 4,261)$	61,68

A metodologia de estimação Clássica gerou a estimativa dos parâmetros dispostos na Tabela 5.5, selecionando o mesmo modelo da metodologia Bayeisana adotada, no entanto com intervalo de confiança (Casella e Berger [2002]) de amplitude bem maior que o da metodologia Bayesiana e sendo o modelo Clássico menos influenciado pelo incremento de uma unidade na covariada X_3 (patrimônio atual/total de débito) que o modelo Bayesiano adotado, que é mais sensível ao acréscimo de uma unidade na covariada selecionada.

Tabela 5.5 *OR e IC segundo a Metodologia Clássica*

Parâmetros	Estimativas	p-value	IC(θ,95%)	OR
α	-6,7382	0,002	–	–
β	3,6777	0,001	(1,449; 5,906)	36,55

A validação para o modelo ajustado segundo a metodologia da Secção 3.2 com a adoção do modelo apenas com a covariada X_3 (patrimônio atual/total de débito), obteve com o ponto de corte que gerou a maior proporção de concordância com os dados, a Tabela 5.6 apresenta a proporcão de acertos com os pontos de corte para a estimação da proporção no modelo, sendo adotado o ponto de corte de 45% estando em concordância com os dados cerca de $83,33\%$ dos dados utilizados na validação do modelo, seis elementos dos dados de Johnson [1987] formam a amostra de validação.

Tabela 5.6 *Ponto de Corte e proporção de acerto na validação do modelo*

Ponto de Corte	Proporção de acerto
40%	0,833
45%	**0, 833**
50%	0,667
55%	0,667
60%	0,500

Portanto, para os dados de Johnson [1987] a variável selecionada foi X_3 (patrimônio atual/total de débito), tanto pela metodologia Bayesiana adotada quanto pela metodologia Clássica, sendo a variável dentre as utilizadas que está influenciando na falência ou não de uma empresa. Sendo assim, se o patrimônio da empresa é maior que o seu débito, esta empresa tem chances de continuar atuando no mercado, tal que o incremento de uma unidade em X_3 aumenta em torno de 62 vezes as chances da empresa não ir a falência.

5.2 Regressão Logística Multinomial Nominal

5.2.1 Aplicação 3: Dosimetria Citogenética

Madruga *et al.* [1994] propôs um modelo de Regressão Logística Multinomial para dados de dose-resposta de um experimento em dosimetria citogenética. O modelo de regressão logística proposto, usa um modelo linear inverso para a transformação *log-odds*

da freqüência de aberrações, onde a presença de micronucleos (MN) indica células com aberrações após a radiação.

Sendo y_{ij} a freqüência de células com j MN ($j = 0, 1$) e y_{i2} a freqüência de células com 2 ou mais MN no i-ésimo nível de dose (D_i). Os modelos ajustados estimarão a proporção de aberrações nas 3 categorias citadas, com *zero MN*, com *um MN* e com *dois ou mais MN*, ou seja, π_{i0}, π_{i1} e π_{i2}, respectivamente. O modelo ajustado por Madruga *et al.* [1994] (com $\pi_{i0} = 1 - \pi_{i1} - \pi_{i2}$) é dado por:

$$\pi_{ij} = \frac{exp(H_j)}{1 + exp(H_1) + exp(H_2)}, j = 1, 2$$

tal que

$$H_j = -\left(\beta_{0j} + \frac{\beta_{1j}}{\beta_{2j} + D_i}\right), j = 1, 2.$$

O trabalho de Madruga *et al.* [1994] sofre críticas por fazer uso dos dados duas vezes para estimar os parâmetros do modelo de Regressão Logística. Nesta aplicação é feito o uso da proposta desenvolvida por Groenewald e Mokgatlhe [2005] para estimar os parâmetros do Modelo de Regressão Logística Multinomial, no entanto usando um modelo linear para a transformação *log-odds* da freqüência de aberrações e outro adotando-se estruturas diferentes na preditora, um preditor linear e outro quadrático, chamado de linear-quadrático. Foi considerada a transformação $x_i = \sqrt{d_i}$, devido os dados não se ajustarem adequadamente a metodologia desenvolvida na Secção 3.2, com d_i representando o i-ésimo valor da dose e tomando o nível com *zero MN* (y_{i0}) como referência. Os dados estão dispostos na Tabela 5.7.

O modelo com estrutura linear (L), ajustado foi

$$\hat{\pi}_{i1} = \frac{\exp(-3,5709 + 0,1649x_i)}{1 + \exp(-3,5709 + 0,1649x_i) + \exp(-6,5386 + 0,2876x_i)}$$

e

$$\hat{\pi}_{i2} = \frac{\exp(-6,5386 + 0,2876x_i)}{1 + \exp(-3,5709 + 0,1649x_i) + \exp(-6,5386 + 0,2876x_i)}$$

sendo que $\hat{\pi}_{i0} = 1 - \hat{\pi}_{i1} - \hat{\pi}_{i2}$, gerando os erros empíricos dispostos na Tabela 5.5.

Os gráficos de dose-resposta dos níveis de *zero MN*, *um MN* e *dois MN* para o modelo com estruturas lineares são dadas, respectivamente, nas Figuras 5.6, 5.7 e 5.8.

Tabela 5.7 *Freqüência de aberrações*

i	*Dose* (CGy)	y_{i0}	y_{i1}	y_{i2}	n_i
1	5	481	17	2	500
2	10	477	19	4	500
3	25	471	24	5	500
4	50	450	44	6	500
5	100	431	59	10	500
6	200	339	140	21	500
7	300	304	132	64	500
8	400	240	189	72	501
9	500	174	197	129	500
10	600	122	173	211	506

Fonte: Madruga *et al.* [1994]

Tabela 5.8 *Erro Empírico do modelo L*

i	e_{i0}	e_{i1}	e_{i2}
1	-0,0036	0,0050	-0,0014
2	-0,0025	0,0071	-0,0046
3	-0,0076	0,0119	-0,0043
4	0,0080	-0,0060	-0,0020
5	-0,0088	0,0069	0,0019
6	0,0497	-0,0692	0,0195
7	-0,0198	0,0238	-0,0040
8	-0,0279	-0,0338	0,0617
9	-0,0170	-0,0222	0,0392
10	-0,0062	0,0334	-0,0272

No modelo com diferentes estruturas nas preditoras, foi adotado um preditor linear para o nível de *um MN* e outro preditor quadrático para o nível de *dois MN*, os modelos ajustados são

$$\hat{\pi}_{i1} = \frac{\exp(-3,6392 + 0,16858x_i)}{1 + \exp(-3,6392 + 0,16858x_i) + \exp(-4,5016 + 0,008598x_i^2)}$$

e

$$\hat{\pi}_{i2} = \frac{\exp(-4,5016 + 0,008598x_i^2)}{1 + \exp(-3,6392 + 0,16858x_i) + \exp(-4,5016 + 0,008598x_i^2)}$$

sendo que $\hat{\pi}_{i0} = 1 - \hat{\pi}_{i1} - \hat{\pi}_{i2}$.

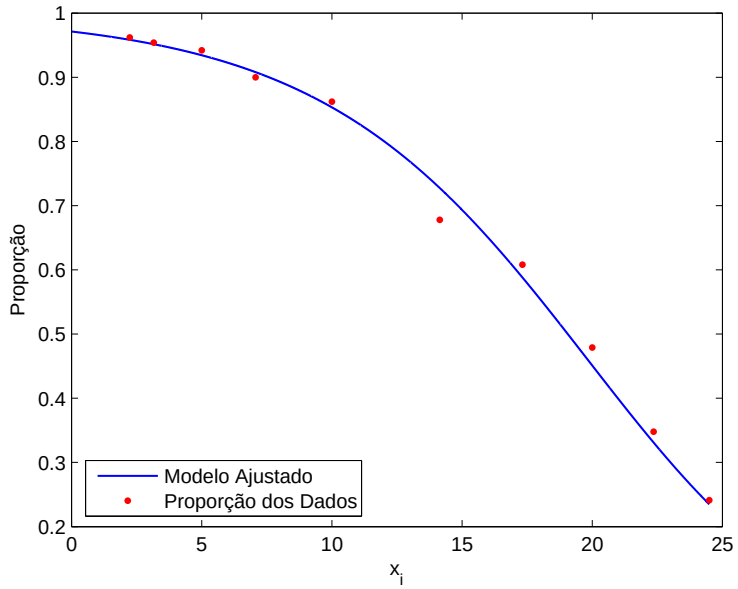

Figura 5.6 *Freqüência de células com zero MN do modelo L*

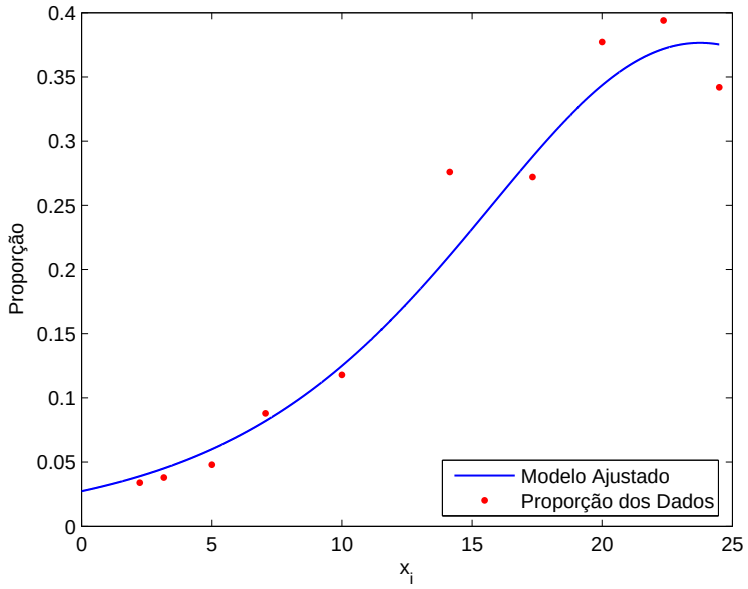

Figura 5.7 *Freqüência de células com um MN do modelo L*

A idéia de utilizar estruturas diferentes nas preditoras, foi para tentar contornar o problema que há na categoria com *um MN*. O problema é devido ao crescimento inicial da freqüência de aberrações e depois do declínio desta freqüência, segundo Madruga *et al.* [1994] o comportamento da freqüência de um MN é esperado, sendo justificado pelo aumento do nível da dose implicar em uma queda do número de células com apenas um

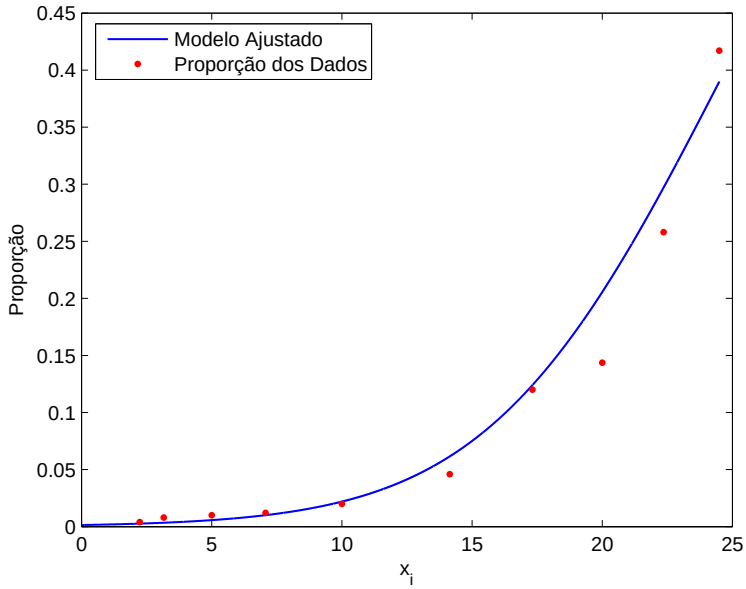

Figura 5.8 *Freqüência de células com dois MN do modelo L*

MN. O modelo ajustado para as três categorias gerou os erros empíricos dispostos na Tabela 5.8.

Tabela 5.9 *Erro Empírico LQ*

i	e_{i0}	e_{i1}	e_{i2}
1	-0,0095	0,0025	0,0070
2	-0,0078	0,0044	0,0034
3	-0,0116	0,0088	0,0028
4	0,0061	-0,0096	0,0034
5	-0,0058	0,0034	0,0024
6	0,0644	-0,0684	0,0040
7	0,0042	0,0342	-0,0384
8	-0,0053	-0,0147	0,0200
9	-0,0097	-0,0086	0,0183
10	-0,0219	0,0160	0,0059

Os gráficos do modelo de dose-resposta ajustado para as categorias de *zero MN*, *um MN* e *dois ou mais MN* do modelo LQ são apresentados, respectivamente, nas Figuras 5.9, 5.10 e 5.11.

A Tabela 5.10 apresenta o erro quadrático médio empírico associado às estimativas

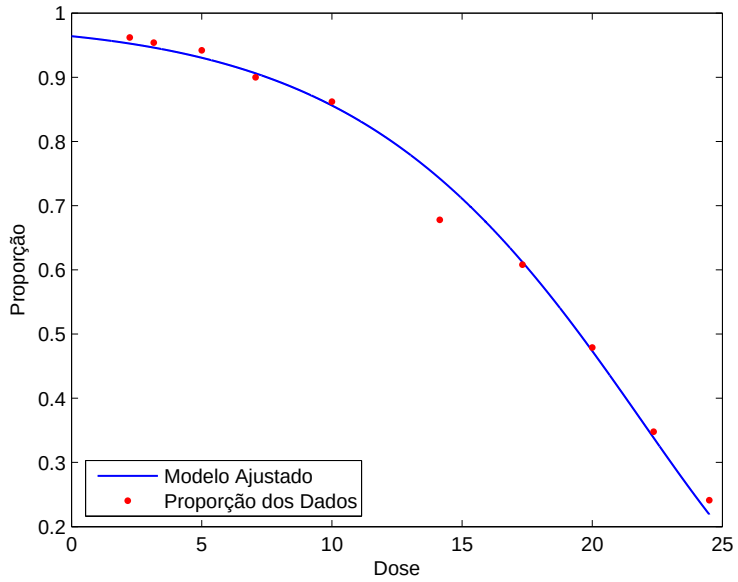

Figura 5.9 *Freqüência de células com zero MN do modelo LQ*

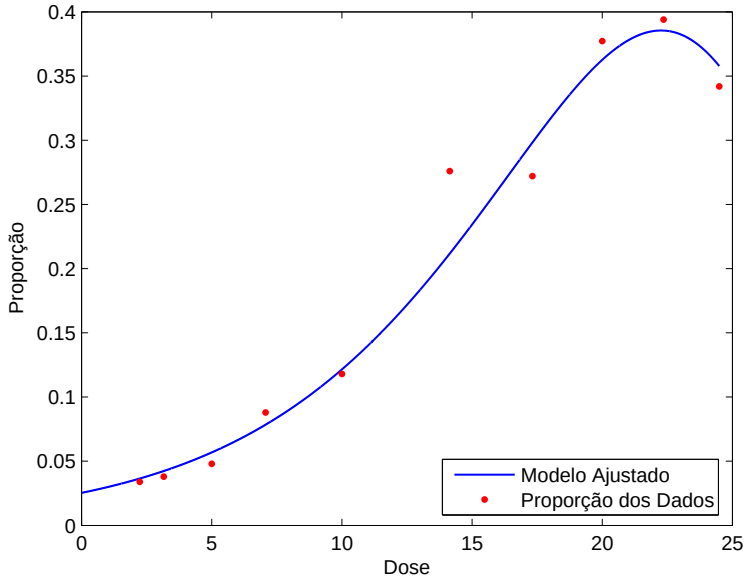

Figura 5.10 *Freqüência de células com um MN do modelo LQ*

obtidas nos dois modelos ajustados em todas as categorias, linear (L), linear-quadrático (LQ) e o proposto por Madruga *et al.* [1994] (MAD) .

Observa-se que o erro quadrático médio foi um pouco menor no modelo LQ e MAD para as categorias com *um MN* e com *dois ou mais MN*, e o modelo L e MAD teve erro menor na categoria com *zero MN*. Mas conclui-se que os dois ajustes propostos neste trabalho foram bons, levando a pequenos erros de estimação que podem ser apresentados

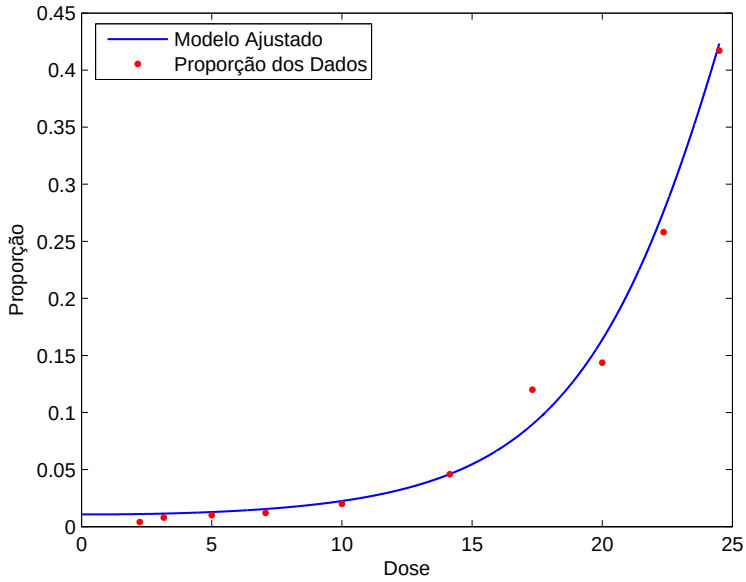

Figura 5.11 *Freqüência de células com dois MN do modelo LQ*

Tabela 5.10 *Erro quadrático médio empírico nos modelos ajustados L, LQ e MAD*

Modelo	0 MN	1 MN	2 ou mais MN
L	0,0004	0,0008	0,0007
LQ	0,0005	0,0007	0,0002
MAD	0,0004	0,0006	0,0003

também nos gráficos apresentados. Os modelos propostos também podem ser considerados melhor que o propostor por Madruga *et al.* [1994] por não fazer uso dos dados duas vezes.

Capítulo 6

Conclusões e Recomendações

Neste trabalho foi apresentada a metodologia de estimação Bayesiana dos parâmetros em Modelos de Regressão Logística proposto por Groenewald e Mokgatlhe[2005], esta proposta se mostrou eficiente no processo de simulação da distribuição a posteriori a partir da implementação computacional do Amostrador de Gibbs.

A seleção do modelo foi feita com o uso do *Fator de Bayes* (Kass e Raftery[1995]), do BIC (*Bayesian Information Criterion*) e com o procedimento proposto por Pereira e Stern [1999], o FBST (*Full Bayesian Statistical Test*). Devido o uso de distribuições logísticas, com uso dos parâmetros estimados, percebe-se no Fator de Bayes uma certa influência na seleção do modelo utilizando esta metodologia, já que o Fator de Bayes é influenciado pelas prioris adotadas (Kass e Raftery [1995]). O BIC por depender da dimensão do espaço paramétrico e do tamanho da amostra não foi percebido influência na seleção do modelo nesta aplicação devido a dimensão do espaço paramétrico ser relativamente pequeno (≤ 5) e o tamanho da amostra também ser pequena (n=40) já o FBST foi coerente com a seleção feita dependendo somente da distribuição a posteriori e selecionando dentre os modelos, o melhor, mas sendo feita esta seleção em conjunto com outras técnicas de seleção como em Pereira e Stern [2001].

A implementação do *FBST* a partir da metodologia proposta por Groenewald e Mokgatlhe[2005], foi facilitada devido ser de "fácil" implementação a maximizaçào da distribuição a posteriori do modelo e a aproximação de Monte Carlo, na parte de integração visto na Secção 3.2.

Nas aplicações feitas neste trabalho a proposta de Groenewald e Mokgatlhe[2005] adequou-se satisfatoriamente aos modelos propostos de Regressão Dicotômica e Policotômica. Na regressão Policotômica a metodologia de estimação possibilitou utilizar formas estruturais diferentes no preditor do modelo para tentar contornar o problema que há na natureza dos dados de Madruga *et al.*[1994].

A metodologia prosposta de Groenewald e Mokgatlhe [2005] necessita de um número relativamente grande de iterações no Amostrador de Gibbs devido a autocorrelação natural que há nas amostras geradas por este amostrador e por fazer uso de variáveis latentes com distribuição uniforme, mas se mostrou bastante eficiente no processo de estimação em Modelos de Regressão Logística. Devido a simplicidade de sua implementação e a possibilidade de contornar problemas clássicos que existem no uso de metodologia Bayesiana, como não conseguir determinar de forma analitica a distribuição a posteriori, a adoção desta metodologia se mostra adequada a aplicação de vários outros conjuntos de dados com variáveis resposta qualitativa.

Com isso, os objetivos deste trabalho foram alcançados com êxito. Como recomendações para trabalhos futuros, podem-se destacar:

- o uso da proposta de Groenewald e Mokgatlhe [2005] para modelos com formas estruturais diferentes em Modelos de Regressão Logística Policotômica;
- o uso da proposta de Groenewald e Mokgatlhe [2005], utilizando o procedimento FBST na seleção de Modelos de Regressão Logística Policotômica, usando distribuições *a priori* Não-informativas.
- o uso da proposta de Groenewald e Mokgatlhe [2005], para Modelos de Regressão Logística Ordinal, para os dados de Madruga *et al.* [1994].
- implementar o FBST para seleção de modelos com o uso do nível de significância empírico;
- aplicar a técnica para a transformação complemento $\log(-\log(1-\pi))$.

Bibliografia

AGRESTI, A. **Categorical Data Analysis, 2 ed**. New Jersey, John Wiley & Sons, 2002.

ALBERT, J. H; CHIB, S. Bayesian analysis of binary and polychotomous response data. **Journal of the American Statistical Association**, 1993, 88, 669-679.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. **Introdução à Teoria da resposta ao Item : Conceitos e Aplicações**. Caxambu: 14º SINAPE, 2000.

BEDRICK, E. J.; CHRISTENSEN, R,; JOHNSON, W. Bayesian binomial regression: predicting survival at a trauma center. **The American Statistician**, 1997, 51, 3, 211-218.

BERGER, J. O.; PERICCHI, L. R. **Objective Bayesian Methods for Models Selection: introduction and comparison**. Cagliari: workshop Bayesian Model Selection, 1997.

BERNARDO, J. M.; SMITH, A. F. M. **Bayesian Theory**. New York: John Wiley & Sons, 1994.

BOLSTAD, W. M. **Introdution to Bayesian Statistics**. New Jersey: John Wiley & Sons, 2004.

BOX, G. E. P.; TIAO, G. C. **Bayesian Inference in Statistical Analysis**. London: Addison Wesley Pub., 1973.

CASELLA, G.; BERGER, R. L. **Statistical Inference, 2nd ed**. Pacific Grove: Duxbury, 2002.

CHEN, M.; DEY, D. K.; SHAO, Q. A new skewed link model for dichotomous quantal response data. **Journal of the American Statistical Association**, 1999, 94, 1172-1186.

CHEN, M.; SHAO, Q; IBRAHIM, J. G. **Monte Carlo Methods in Bayesian Computation**. New Yorq: Springer, 2000.

CHIB, S. Marginal likelihood from the Gibbs output. **Journal of the American Statistical Association**, 1995, 90, 1313-1321.

CORDEIRO, G. M. **Modelos Lineares Generalizados**. Campinas: VII Simpósio Nacional de Probabilidade e Estatística, 1986.

DEY, D. K; GHOSH, S. K; MALLICK, B. K. **Generalized Linear Models: a Bayesian Perspective**. New York-Basel: Marcel Dekker, 2000.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis, 3 ed**. New York: John Wiley & Sons, 1998.

FAHRMEIR, L; TUTZ, G. **Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd ed**. New York: Springer, 2001.

GAMERMAN, D. **Simulação Estocástica via Cadeias de Markov**. XII Simpósio Nacional de Probabilidade e Estatística. Associação Brasileira de Estatística, 1996.

GELFAND, A. E.; SMITH, A. F. M. Sampling-Based Approaches to Calculating Marginal Densities. **Journal of the American Statistical Association**, 85, 398-409, 1990.

GEMAN, S.; GEMAN, D. Stochastic Relaxation, Gibbs Distribuitions and the Bayesian Restoration of Images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 1984, 6, 721-741.

GILKS, W. R; RICHARDSON, S; SPIEGELHALTER, D. J. **Markov Chain Monte Carlo in Practice**. London: Chapman and Hall, 1996.

GROENEWALD, P. C. N; MOKGATLHE, L. Bayesian computation for logistic regression. **Computational Statistics** & **Data Analysis**, 48, 857-868, 2005.

HAIR, J. F; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Analise Multivariada de Dados, 5 ed**. Porto Alegre: Bookman, 2005.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression, 2 ed**. New York: John Wiley & Sons, 2001.

JAMES, B. R. **Probabilidade: um curso em nível intermediário**. Rio de Janeiro: IMPA, 1981.

JOHNSON, W. The detection of the influencial observations for allocation, separation and the determination of probabilities in bayesian framework. **Journal of the Bussines and Economic Statistics**, 5, 3, 369-381, 1987.

KASS, R. E.; RAFTERY, A. E. Bayes Factor. **Journal of the American Statistical Association**, 1995, 90, 773-795.

LEHMANN, E. L. **Testing Statistical Hypotheses**. New York: John Wiley & Sons, 1959.

MADRUGA, M. R.; PEREIRA, C. A. de B.; GAY-RABELO, M. N. Bayesian dosimetry: radiation dose versus frequencies of cells with aberrations. **Envirometrics**, 1994, 5, 47-56.

MADRUGA, M. R.; ESTEVES, L. G.; WECHSLER, S. On the Bayesian of Pereira-Stern test. **Test**, 2001, 10, 291-299.

McCULLOCH, C. E; SEARLE, S. R. **Generalized, Linear, and Mixed Models**. New York: John Wiley & Sons, 2001.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introdution to the Theory of Statistical**. Singapore: McGraw-Hill, 1974.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J; WASSERMAN, W. **Applied Linear Statistical Models, 4ed**. : McGraw-Hill, 1996.

O'HAGAN, A. **Kendall's Advanced Theory of Statistics 2B**: Bayesian Inference. London: Edward Arnold, 1994.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003.

PEREIRA, C. A. de B.; STERN, J. M. Evidence and Credibility: a full Bayesian test of precise hypothesis. **Entropy**, 1, 99-110, 1999.

PEREIRA, C. A. de B.; STERN, J. M. Model Selection: Full Bayesian Approach. **Environmetrics**, 12, 559-568, 2001.

ROSS, S. M. **Estochastic Process.** 2 ed. New York:New York: John Wiley & Sons, 1995.

ROYALL, M. R. On the Probability of Observing Misleading Statistical Evidence. **Journal of the American Statistical Association**, 2000, 451, 760-780.

TANNER, T.A.; WONG, W.H. The Calculation of Posterior Distribution by Data Augmentation. **Journal of the American Statistical Association**, 1987, 82, 528-549.

THE MATHWORKS, Inc. **MATLAB**: The Language of Technical Computing. Version 7.7.0.19920(R14). 2004.