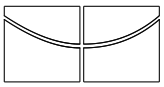

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Regressão Ordinal Bayesiana

por

Leonardo Oliveira Gois Cella

Orientador: Prof.º Eduardo Yoshio Nakano

Brasília, 2013

Leonardo Oliveira Gois Cella

Regressão Ordinal Bayesiana

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof.º Eduardo Yoshio Nakano

Universidade de Brasília

Brasília, 2013

Agradecimentos

À toda minha família, que torceu muito para eu conseguir vencer mais essa etapa. Em especial à minha mãe Elizabeth, que é a minha grande inspiração na vida; ao meu pai Antônio Gois, que sempre deu todo o suporte e motivação nos meus estudos e à minha irmã Déborah, cuja serenidade, amizade e (ás vezes) bom humor certamente tornaram essa caminhada menos árdua.

Aos meus amigos do peito, que além de me incentivarem e torcerem por mim, toleraram o meu "projeto monge". Em especial ao André, que tem o dom e a paciência de ensinar e me ajudou em muitos aspectos computacionais dessa dissertação; ao Tomaz, por ter "segurado as pontas" no nosso trabalho quando o tempo estava corrido para a finalização da dissertação; ao Bernardo, por ter o talento de me fazer rir de qualquer coisa e ao Alexis, que fala tanto que acabou me ajudando nesses 2 anos a esquecer toda a pressão do momento.

Ao Mestre Deo e toda família Deo de Jiu Jitsu. Começar o dia treinando certamente me motivava para os desafios diários.

Ao orientador Eduardo Yoshio Nakano, não só pela impecável orientação, mas também por ter se tornado um amigo ao qual posso contar e ser um exemplo de estatístico que me espelharei em toda carreira profissional.

Resumo

Este trabalho apresenta a inferência do modelo de regressão ordinal, considerando a ligação Logit e a abordagem da verossimilhança multinomial. Foi proposta uma reparametrização do modelo de regressão. As inferências foram realizadas dentro de um cenário bayesiano fazendo-se o uso das técnicas de MCMC (Markov Chain Monte Carlo). São apresentadas estimativas pontuais dos parâmetros e seus respectivos intervalos HPD, assim como um teste de significância genuinamente bayesiano $FBST$ (Full Bayesian Significance Test) para os parâmetros de regressão. A metodologia adotada foi aplicada em dados simulados e ilustrada por um problema genético que verificou a influência de um certo tipo de radiação na ocorrência de danos celulares. A abordagem da verossimilhança multinomial combinada à reparametrização do modelo é de fácil tratamento devido ao aumento da capacidade computacional e do avanço dos métodos MCMC. Além disso, o $FBST$ se mostrou um procedimento simples e útil para testar a significância dos coeficientes de regressão, motivando assim a utilização de uma abordagem bayesiana na modelagem de dados ordinais.

Palavras-chave: verossimilhança multinomial, logit, MCMC, intervalos HPD, FBST.

Abstract

This work presents inferences of ordinal regression models considering the Logit link functions and the multinomial likelihood approach. A new reparametrization was proposed for the regression model. The inferences were performed in a bayesian scenario, using the MCMC (Markov Chain Monte Carlo) technics. Point estimates of the parameters and their respective HPD credibility intervals are presented, as well a Full Bayesian Significance Test (FBST) for the regression parameters. This methodology was applied on simulated data and illustrated in a genetic problem which was to verify the influence of certain radiation on the occurrence of cellular damage. The multinomial likelihood approach combined with the model reparametrization is easy to treat due the increasing computing power and the advancement of MCMC methods. Moreover, the FBST proved being a simple and useful procedure for testing the significance of regression coefficients, thus motivating the use of a bayesian approach in ordinal data modeling.

Keywords: multinomial likelihood, logit, MCMC, HPD credibility intervals, FBST.

Sumário

Capítulo 1

Introdução

Os modelos de regressão logística são utilizados com o objetivo de relacionar uma variável resposta qualitativa com um conjunto de váriáveis explicativas. Em seu caso mais simples, em que a variável resposta é dicotômica, tem-se o modelo de regressão logística binária. É possivel, no entanto, ter um número maior de categorias, configurando uma variável com resposta politômica, resultando em um modelo de regressão logística multi-categórico.

No caso em que a variável resposta é nominal, tem-se um modelo de regressão logística multi-categórico nominal, ou simplesmente modelo logístico nominal. Ao se tratar de uma variável categórica ordinal, a presença de uma ordenação na resposta torna pouco razoável a aplicação da regressão nominal pelo fato da mesma não considerar essa ordenação. O caráter ordinal da variável resposta permitiria a adoção de modelos mais refinados, como os modelos de regressão logística multi-categórico ordinal (ou modelo logístico ordinal).

A presença de variáveis categóricas ordinais é muito comum em muitas áreas de estudo, principalmente quando não é possível obter medidas exatas das variáveis. As variáveis ordinais podem surgir através de mecanismos completamente distin-

tos. Anderson (1984) classifica as variáveis ordinais em: i) variáveis quantitativas agrupadas e; ii) variáveis categóricas naturalmente ordenadas. O primeiro tipo de variáveis representa uma versão categorizada de uma variável quantitativa. Como exemplo, tem-se que a nota de um aluno em uma escola pode ser medida ordinalmente através da categorização [0,3), [3,5), [5,7), [7,9) e [9,10], por exemplo (esse exemplo representa a categorização das notas em menções, adotada pela Universidade de Brasília). Já o segundo tipo de variável ordinal ocorre quando é avaliada uma informação não quantificável, associada a níveis de uma escala originalmente ordinal. Como exemplo, tem-se a concordância sobre um assunto, cuja variável pode apresentar vários níveis desde Concordar Plenamente a Discordar Plenamente.

Neste contexto, devido à importância das variáveis ordinais, o objetivo deste trabalho é realizar inferências em um modelo de regressão ordinal dentro de um contexto bayesiano. O método consistirá em determinar as estimativas pontuais e intervalares de cada um dos parâmetros do modelo e também propor um teste de significância genuinamente bayesiano (FBST - Full Bayesian Significance Test) para testar a significância estatística dos coeficientes das variáveis explicativas (Pereira and Stern, 1999). A inferência será realizada exclusivamente pela distribuição a posteriori dos parâmetros, que será aproximada através de simulações via MCMC - Markov Chain Monte Carlo (Gamerman, 1997). Todas as simulações, estimativas dos parâmetros e cálculo do e-valor (valor de evidência do FBST) serão realizadas através do software R (R Core Team, 2013).

O segundo e terceiro capítulos irão expor as abordagens clássicas da regressão

logística binária e da regressão logística multi-categórica (nominal e ordinal), respectivamente. O quarto capítulo revisa os principais conceitos da Inferência Bayesiana, enquanto a definição e propriedades do $FBST$ são discutidas no quinto capítulo. Em seguida, o modelo de regressão ordinal é formulado em uma abordagem bayesiana no capítulo 6. Simulações visando a comparação das abordagens clássica e bayesiana no modelo ordinal, bem como o desempenho do $FBST$, estão presentes no sétimo capítulo. Por fim, o oitavo capítulo contém uma aplicação a dados reais e o nono capítulo apresenta as conclusões obtidas com esse trabalho.

Capítulo 2

Regressão Logística Clássica

Os modelos de regressão vêm tendo aplicabilidade e desenvolvimento crescente nas últimas décadas, tanto em função do avanço teórico científico, quanto em razão do rápido aumento da capacidade de processamento computacional, que permite o uso de complexos algoritmos numéricos e viabilizam muitas das estimativas necessárias a esta classe de modelos (Barreto, 2011).

A sua utilização tem por finalidade descrever a relação entre uma variável resposta e uma ou mais variáveis explicativas. Essa metodologia é largamente utilizada nos mais diversos campos do conhecimento, como engenharias, administração, marketing, medicina, ciências agrárias, informática, dentre outras.

Em muitas aplicações, a variável resposta assume somente dois valores qualitativos comumente denotados por "sucesso" e "fracasso", podendo ser representada por uma variável indicadora binária de valores 0 e 1. Nesses casos a regressão logística é a abordagem mais popular, descrevendo a relação entre a ocorrência ou não de um evento de interesse e um conjunto de variáveis independentes.

Seja Y uma variável binária como descrito acima. Sua distribuição é a de Bernoulli, sendo especificada pelas probabilidades $P(Y = 1) = \pi$ de sucesso e

$P(Y = 0) = (1 - \pi)$ de fracasso. Sua média é $E(Y) = \pi$. Suponha $\boldsymbol{X}$ um vetor de variáveis explicativas $(X_1, X_2, ..., X_k)$, correspondendo às variáveis independentes do estudo. Agora, $\pi = P(Y = 1|\boldsymbol{x})$ denota a probabilidade de sucesso para os valores específicos das variáveis explicativas.

Nos modelos de regressão linear, $\mu = E(Y)$ é uma função linear de $\boldsymbol{X}$. Para uma resposta binária, um modelo análogo seria

$$E(Y_i|\boldsymbol{x}) = \pi_i = \alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} \quad , \tag{2.1}$$

ou simplesmente

$$E(Y_i|\boldsymbol{x}) = \pi_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} \tag{2.2}$$

com $\boldsymbol{\beta}^{'} = [\alpha_0 \quad \beta_1 \quad \cdots \quad \beta_k]$ sendo o vetor de coeficientes de regressão a serem estimados. Este é um modelo simples, em que as probabilidades de sucesso mudam linearmente em $\boldsymbol{x}$. No entanto, considerar o modelo de regressão linear quando a variável resposta é binária traz os seguintes problemas (Kutner et al., 2004):

1. **Não normalidade dos erros**: Cada erro $\varepsilon_i = Y_i - (\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})$ pode assumir somente dois valores:

$$\mathrm{Y}_i = 1: \quad \varepsilon_i = 1 - (\alpha_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$$

$$\mathrm{Y}_i = 0: \quad \varepsilon_i = -\alpha_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik} \quad .$$

   Claramente, o modelo de regressão linear, que assume ε_i normalmente distribuído, não é apropriado.

2. **Heterocedasticidade**: Outro problema com os erros ε_i é que eles não têm variâncias iguais quando a resposta é uma variável indicadora:

$$\sigma^2(Y_i) = \pi_i(1-\pi_i) = EY_i(1-EY_i) \quad .$$

A variância de ε_i é a mesma de Y_i, pois $\varepsilon_i = Y_i - \pi_i$ e π_i é uma constante:

$$\sigma^2(\varepsilon_i) = (EY_i)(1-EY_i)$$

$$\sigma^2(\varepsilon_i) = (\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})(1 - \alpha_0 - \beta_1 x_{i1} - ... - \beta_k x_{ik}) \quad . \tag{2.3}$$

Pela equação 2.3, nota-se que $\sigma^2(\varepsilon_i)$ depende de cada x_i, fazendo com que a variância dos erros difiram para cada nível de $\boldsymbol{x}$.

3. **Restrição no modelo**: Como o que está sendo modelado são probabilidades, teremos a seguinte restrição para a resposta média do modelo:

$$0 \leq EY_i = \pi_i \leq 1 \quad . \tag{2.4}$$

Funções de resposta lineares como a do modelo de regressão linear podem não atender satisfatoriamente a essa restrição, o que seria claramente uma impropriedade matemática.

Com as dificuldades apontadas acima, principalmente a restrição em 2.4, usar o modelo de regressão linear para modelar probabilidades não é uma boa escolha. Frequentemente isso implica em transformações que acabam por tornar a resposta esperada a ser estimada pelo modelo em uma função não linear. A transformação mais comumente usada utiliza a função exponencial para garantir valores compreendidos no intervalo $(0, 1)$, resultando na função de resposta logística (Agresti, 2012):

$$E(Y_i|\boldsymbol{x}) = \pi_i = \frac{exp(\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})}{1 + exp(\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})} \quad . \tag{2.5}$$

Em notação matricial:

$$E(Y_i|\boldsymbol{x}) = \pi_i = \frac{exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})} \quad . \tag{2.6}$$

Assim, o objetivo da transformação foi atingido, já que, enquanto $\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$ pode assumir qualquer valor real, π_i está restrito ao intervalo $(0, 1)$. Este modelo em 2.5 assume que os Y_i são variáveis aleatórias Bernoulli independentes, com parâmetro $E(Y_i) = \pi_i$.

A transformação de π que é utilizada para obtenção da forma aditiva é a transformação logito, e é definida da seguinte forma (Agresti, 2012):

$$logit[\pi] = log\left(\frac{\pi}{1-\pi}\right) = \alpha_0 + \beta_1 x_1 + ... + \beta_k x_k = \boldsymbol{x}^{'}\boldsymbol{\beta} \quad . \tag{2.7}$$

Dessa forma, 2.7 possui várias das propriedades desejadas de um modelo de regressão linear. A logito é contínua, linear nos seus parâmetros e pode assumir qualquer valor real.

Um dos fatores que tornam a regressão logística mais interessante e contribuiu para sua popularização é a interpretação simples e útil dos coeficientes do modelo (Kutner et al., 2004). Ela é baseada na razão de chances. O termo $\left(\frac{\pi}{1-\pi}\right)$ é conhecido como chance de sucesso. Assim, um incremento unitário em algum x_k ocasionará que a chance de sucesso estimado anteriormente a este incremento seja multiplicado por $exp(\beta_k)$, mantidas constantes todas as demais variáveis explicativas do modelo (Agresti, 2012).

A estimação dos parâmetros do modelo é feita pelo método de Máxima Verossimilhança. Assumindo que cada uma das n respostas da amostra é uma variável de Bernoulli independente, onde

$$P(Y_i = 1|\boldsymbol{x}) = \pi_i$$

$$P(Y_i = 0|\boldsymbol{x}) = 1 \text{ - } \pi_i \quad ,$$

podemos representar suas distribuições de probabilidade como se segue:

$$P_{Y_i}(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad y_i = 0, 1; \quad i = 1, \cdots, n \quad . \tag{2.8}$$

Como as n observações Y_i são independentes, a função de probabilidade conjunta é dada por

$$P(y_1, \cdots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad . \tag{2.9}$$

Substituindo π_i por 2.6 obtemos a expressão da função de verossimilhança (Kutner et al., 2004):

$$P(y_1, \cdots, y_n|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{y_i} \left(\frac{1}{1+exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{1-y_i} \quad . \tag{2.10}$$

As estimativas de máxima verossimilhança dos parâmetros são os valores que maximizam a função 2.10 acima. Não há formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos para tal fim, como por exemplo o de Newton-Raphson (Kutner et al., 2004).

Capítulo 3

Regressão Logística Multi-Categórica

Apesar da Regressão Logística ser mais frequentemente usada para modelar a relação entre uma variável resposta dicotômica e um conjunto de variáveis independentes, em muitas situações a variável resposta possui mais de dois níveis. Nesse contexto, uma das abordagens utiliza uma generalização da regressão logística vista no capítulo anterior, chamada de regressão logística multi-categórica, (Kutner et al., 2004).

Exemplos da regressão logística multi-categórica ocorrem em diferentes campos:

- Nos negócios, quando um consultor de marketing deseja relacionar a escolha dos consumidores por um tipo de produto (produto A, B ou C) segundo a idade, o sexo e outras possíveis variáveis explicativas desses consumidores.

- Na medicina, relacionando a severidade de uma determinada doença (baixa, média ou grave) com características dos pacientes (se é fumante, se faz atividade física e etc).

Com a ajuda dos exemplos acima percebe-se que há duas situações distintas envolvendo a variável resposta. No primeiro, ela é puramente qualitativa, sem ne-

nhuma ordenação. Teríamos um regressão logística multi-categórica nominal. Já no segundo exemplo, a severidade da doença possui uma ordem natural, o que leva a uma regressão logística multi-categórica ordinal (Kutner et al., 2004). Apesar de ser possível a modelagem de dados ordinais a partir da regressão logística multi-categórica nominal, a perda de informação referente à ordem das categorias e a maior complexidade desse modelo incentivam a adoção do modelo logístico multi-categórico ordinal. As duas abordagens serão descritas a seguir.

3.1 Regressão Nominal Clássica

Seja J o número de categorias de uma variável resposta Y, e $[\pi_1, \pi_2, \cdots, \pi_J]$ as respectivas probabilidades, satisfazendo $\sum_{j=1}^{J} \pi_j = 1$. Com n observações independentes, a distribuição de probabilidades do número de resultados dos J tipos é multinomial, especificando a probabilidade de todas as formas que as n observações podem se associar às J categorias (Agresti, 2007).

A i-ésima observação pode ser escrita a partir de J variáves resposta binárias, $Y_{i1}, \cdots, Y_{iJ}$, onde:

$$Y_{ij} = \begin{cases} 1, & \text{se a } i\text{-ésima resposta está na categoria } j \\ 0, & \text{caso contrário.} \end{cases}$$

Como somente uma categoria pode ser selecionada para a i-ésima variável resposta, teremos

$$\sum_{j=1}^{J} Y_{ij} = 1 \quad .$$

Seja agora π_{ij} a probabilidade da categoria j ser selecionada para a i-ésima resposta.

Então

$$\pi_{ij} = P(Y_{ij} = 1) \quad .$$

Para o caso da regressão logística binária (vista no capítulo anterior), tem-se que $J = 2$. Se denotarmos $Y_i = 1$ se a i-ésima resposta for da Categoria 1, e $Y_i = 0$ se a i-ésima resposta for da Categoria 2, então

$$\pi_i = \pi_{i1} \quad e \quad 1 - \pi_i = \pi_{i2} \quad .$$

Nesse contexto, o *Logit* de π_i é modelado utilizando um preditor linear. Como há somente 2 categorias na regressão logística binária, o *Logit* de fato compara a probabilidade da resposta ser da Categoria 1 com a probabilidade da resposta ser da Categoria 2

$$\pi_i^{'} = log\left[\frac{\pi_i}{1-\pi_i}\right] = log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \pi_{i12}^{'} = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{12} \quad .$$

Note que foi escrito $\pi_{i12}^{'}$ e $\boldsymbol{\beta}_{12}$ para enfatizar que o preditor linear está modelando o logaritmo da razão das probabilidades para as Categorias 1 e 2 (Kutner et al., 2004).

Generalizando para J categorias, tem-se $J(J-1)/2$ pares de categorias para comparação, e consequentemente $J(J-1)/2$ preditores lineares . Por exemplo, $J = 3$ resulta em 3 comparações, e, consequentemente, 3 modelos de regressão logística a serem estimados:

$$\pi_{i12}^{'} = log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{12},$$
$$\pi_{i13}^{'} = log\left[\frac{\pi_{i1}}{\pi_{i3}}\right] = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{13},$$
$$\pi_{i23}^{'} = log\left[\frac{\pi_{i2}}{\pi_{i3}}\right] = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{23}.$$

Ou seja, os modelos logísticos multi-categóricos nominais utilizam simultaneamente todos os pares de categorias, especificando a chance da resposta estar em uma categoria em comparação com outra (lembrando que não há ordenação nas categorias).

No entanto, não é necessário desenvolver todos os $J(J-1)/2$ modelos logísticos. Na prática, uma categoria é escolhida como base, e então todas as demais categorias são comparadas a ela. A escolha da categoria base, também chamada de categoria (nível) de referência, é arbitrária (Agresti, 2007). Por exemplo, utilizando a categoria J como base, é necessário considerar apenas as $J-1$ comparações a essa categoria. O *Logit* para a j-ésima comparação é (Kutner et al., 2004)

$$\pi_{ijJ}^{'} = log\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{jJ}, \quad j = 1, 2, \cdots, J-1 \quad . \tag{3.1}$$

Como todas as comparações são feitas com a categoria J, reescreve-se na equação acima: $\pi_{ijJ}^{'} = \pi_{ij}^{'}$ e $\boldsymbol{\beta}_{jJ} = \boldsymbol{\beta}_j$, ou seja

$$\pi_{ij}^{'} = log\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{j}, \quad j = 1, 2, \cdots, J-1 \quad . \tag{3.2}$$

A razão para se considerar apenas os $J-1$ *Logits* está no fato de que qualquer outro *Logit* pode ser obtido através deles (Kutner et al., 2004). Por exemplo, para $J = 4$, a comparação das Categorias 1 e 2 (tendo a quarta categoria como base) pode ser obtida por

$$\begin{aligned} log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] &= log\left[\frac{\pi_{i1}}{\pi_{i4}} \times \frac{\pi_{i4}}{\pi_{i2}}\right] \\ &= log\left[\frac{\pi_{i1}}{\pi_{i4}}\right] - log\left[\frac{\pi_{i2}}{\pi_{i4}}\right] \\ &= \boldsymbol{x}_i^{'}\boldsymbol{\beta}_1 - \boldsymbol{x}_i^{'}\boldsymbol{\beta}_2 \quad . \end{aligned}$$

Em geral, para comparar as categorias k e l:

$$log\left[\frac{\pi_{ik}}{\pi_{il}}\right] = \boldsymbol{x}_i^{'}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_l) \quad . \tag{3.3}$$

A interpretação para os coeficientes do modelo nominal (composta pelos $J(J-1)/2$ modelos logísticos) seguem a mesma lógica da regressão logística clássica, baseada em razões de chances. Só deve-se prestar atenção em quais categorias estão sendo comparadas para os parâmetros por vez analisados (Agresti, 2012).

Dadas as $J-1$ expressões logísticas em 3.2, obtem-se as $J-1$ expressões diretas para as probabilidades de cada categoria em termo dos $J-1$ preditores lineares $\boldsymbol{x}^{'}\boldsymbol{\beta}_j$. As expressões resultantes são (Kutner et al., 2004)

$$\pi_{ij} = \frac{exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta}_j)}{1 + \sum_{k=1}^{J-1} exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta}_k)} \quad . \tag{3.4}$$

A estimação dos $J-1$ vetores de parâmetros $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_{J-1}$ é feita simultaneamente pelo método de máxima verossimilhança. Para isso é necessário obter a função de verossimilhança dos dados.

A fim de fixar idéias, suponha que há $J = 4$ categorias e a terceira categoria é escolhida para a i-ésima resposta. Ou seja, para o i-ésimo caso teremos

$$Y_{i1} = 0, \quad Y_{i2} = 0, \quad Y_{i3} = 1 \quad e \quad Y_{i4} = 0.$$

A probabilidade dessa resposta é

$$\begin{aligned} P(Y_i = 3) = \pi_{i3} &= [\pi_{i1}]^0 \text{ x } [\pi_{i2}]^0 \text{ x } [\pi_{i3}]^1 \text{ x } [\pi_{i4}]^0 \\ &= \prod_{j=1}^{4} [\pi_{ij}]^{y_{ij}} \quad . \end{aligned}$$

Assim, para n observações independentes e J categorias, a função de probabilidade conjunta é dada por (Kutner et al., 2004)

$$P(y_1, \cdots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} [\pi_{ij}]^{y_{ij}} \right] \quad . \tag{3.5}$$

Sendo $\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}$ e $y_{iJ} = 1 - \sum_{j=1}^{J-1} y_{ij}$, a expressão fica

$$P(y_1, \cdots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \left[\left(\prod_{j=1}^{J-1} [\pi_{ij}]^{y_{ij}} \right) \left(\left[1 - \sum_{j=1}^{J-1} \pi_{ij} \right]^{1 - \sum_{j=1}^{J-1} y_{ij}} \right) \right] \quad .$$

Substituindo π_{ij} por 3.4, obtem-se a expressão da função de verossimilhança desejada, considerando a categoria J como base (Kutner et al., 2004):

$$P(y_1, \cdots, y_n | \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_{J-1}) =$$

$$= \prod_{i=1}^{n} \left\{ \left[\prod_{j=1}^{J-1} \left(\frac{exp(\boldsymbol{x}_i' \boldsymbol{\beta}_j)}{1 + \sum_{k=1}^{J-1} exp(\boldsymbol{x}_i' \boldsymbol{\beta}_k)} \right)^{y_{ij}} \right] \left[\left(\frac{1}{1 + \sum_{k=1}^{J-1} exp(\boldsymbol{x}_i' \boldsymbol{\beta}_k)} \right)^{1 - \sum_{j=1}^{J-1} y_{ij}} \right] \right\} \quad . \tag{3.6}$$

As estimativas dos parâmetros em $\boldsymbol{\beta_j}$ são os valores que maximizam essa função. Não existem formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos como por exemplo o de Newton-Raphson (Kutner et al., 2004).

3.2 Regressão Ordinal Clássica

A variável resposta de escala ordinal de um estudo pode ser fruto de duas abordagens. A primeira representa uma versão categorizada de uma variável contínua. Como exemplo, temos a variável número de anos de estudo, que pode ser medida ordinalmente por meio da categorização 0-8,9-12, 13-16, 17 ou mais anos. Já a segunda

ocorre quando é avaliada uma informação não quantificável, associada a níveis de uma escala categórica ordinal. Este tipo, que consiste de uma coleção de categorias naturalmente ordenadas, origina-se de casos onde uma medida precisa nem sempre é possível. A classe social, classificada como "baixa", "média" ou "alta" e a filosofia política ("liberal", "moderada" ou "conservadora") servem como ilustração.

Nos dois tipos de abordagens mencionados acima é sensato imaginar que a variável observada é uma categorização de uma variável contínua latente. No caso de variáveis contínuas agrupadas a variável latente pode ser considerada a variável subjacente não observada. Já no caso de variáveis categóricas naturalmente ordenadas, a variável latente é uma avaliação sobre uma escala contínua subjacente (Anderson, 1984). Por exemplo, em um estudo que pergunta a ideologia política dos entrevistados de acordo com cinco categorias (muito liberal, pouco liberal, moderado, pouco conservador, muito conservador), na prática pode haver diferenças na ideologia política de pessoas que se classificaram na mesma categoria. Assim, é possível imaginar uma medida contínua que messa precisamente qual a ideologia política de uma pessoa (variável latente).

A seguir será descrito o tipo de modelo mais utilizado para o desenvolvimento da regressão logística multi-categórica ordinal, ou simplesmente regressão logística ordinal. Apesar de respostas ordinais poderem também ser analisadas com as técnicas de regressão logística multi-categórica nominal, levar em conta a ordem das categorias resulta em um modelo mais parcimonioso e de mais fácil interpretação (Kutner et al., 2004).

3.2.1 Modelos Cumulativos

A relação de ordem entre as classes da variável dependente faz com que a tarefa de modelar a probabilidade de ocorrência de uma das suas classes seja feita em termos de probabilidades acumuladas (Agresti, 2007).

A probabilidade cumulativa de uma variável Y é a probabilidade de Y assumir valores iguais ou menores que um determinado ponto.

Sendo Y uma variável ordinal com J classes, a probabilidade de se observar uma classe inferior ou igual a j, para um determinado vetor de observações das variáveis independentes $\boldsymbol{X}$, é dado por

$$P(Y \leq j|\boldsymbol{x}) = \pi_1 + \cdots + \pi_j, \quad j = 1, \cdots, J \quad , \tag{3.7}$$

com $\pi_1 = P(Y = 1|\boldsymbol{x}), \pi_2 = P(Y = 2|\boldsymbol{x}), \cdots, \pi_J = P(Y = J|\boldsymbol{x})$. Naturalmente, as probabilidades acumuladas refletem a ordenação natural $P(Y \leq 1|\boldsymbol{x}) \leq P(Y \leq 2|\boldsymbol{x}) \leq \cdots \leq P(Y \leq J-1|\boldsymbol{x})$. Modelos para probabilidades cumulativas não utilizam a última categoria, $P(Y \leq J|\boldsymbol{x})$ visto que ela é necessariamente igual a 1 (informação referente à ultima classe é redundante).

A variável ordinal pode ser interpretada como a operacionalização de uma outra variável continua não medida (latente), como vimos anteriormente. Assim, a variável manifesta (Y) resulta do "corte" da variável latente (Y^*) em J classes ordinais e mutuamente exclusivas (Okura, 2008).

Suponha $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_J = \infty$ os pontos de corte da escala contínua de Y^*, dado que a variável resposta observada satisfaz

$$Y = j \quad se \quad \alpha_{j-1} < Y^* \leq \alpha_j, \quad j = 1, 2, \cdots, J \quad . \tag{3.8}$$

Em outras palavras, observa-se Y na categoria j quando a variável latente cair no j-ésimo intervalo de valores.

Suponha agora que a variável latente Y^* é determinada pelas variáveis explicativas de forma linear:

$$Y^* = \boldsymbol{x}^{'}\boldsymbol{\beta} + \varepsilon \quad , \tag{3.9}$$

onde $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)$ é o vetor de parâmetros e ε é uma variável aleatória com distribuição F.

Desses fatos, segue que a distribuição de probabilidade da variável observada Y é dada por

$$P(Y \leq j|\boldsymbol{x}) = F(\alpha_j - \boldsymbol{x}^{'}\boldsymbol{\beta}) \quad . \tag{3.10}$$

Para verificar 3.10, basta notar que:

$P(Y \leq j|\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) + P(Y = 2|\boldsymbol{x}) + \cdots + P(Y = j|\boldsymbol{x}) = P(\alpha_0 \leq Y^* \leq \alpha_1|\boldsymbol{x}) + P(\alpha_1 \leq Y^* \leq \alpha_2|\boldsymbol{x}) + \cdots + P(\alpha_{j-1} \leq Y^* \leq \alpha_j|\boldsymbol{x}) = F_{Y^*|\boldsymbol{x}}(\alpha_j) - F_{Y^*|\boldsymbol{x}}(\alpha_0)$

Como $\alpha_0 = -\infty$, temos $F_{Y^*|\boldsymbol{x}}(\alpha_0) = 0$. Assim $F_{Y^*|\boldsymbol{x}}(\alpha_j) = P(\boldsymbol{x}^{'}\boldsymbol{\beta} + \varepsilon \leq \alpha_j) = P(\varepsilon \leq \alpha_j - \boldsymbol{x}^{'}\boldsymbol{\beta}) = F(\alpha_j - \boldsymbol{x}^{'}\boldsymbol{\beta})$, onde F é a função de distribuição da variável aleatória ε.

O inverso da função F, isto é, F^{-1} é designada função de ligação ($Link$), por fazer a associação linear entre a parte aleatória do modelo, $P(Y \leq k)$, e a parte sistemática ($\boldsymbol{x}^{'}\boldsymbol{\beta}$). Ou seja

$$Link(P(Y \leq j)) = \alpha_j - \boldsymbol{x}^{'}\boldsymbol{\beta} \quad . \tag{3.11}$$

Várias são as opções para se usar como função de ligação, cuja utilização no modelo ordinal é recomendável de acordo com o tipo de distribuição de probabili-

dades que as classes da variável dependente apresentam. Esta escolha deve ser feita com cuidado, pois uma escolha inapropriada pode comprometer a significância do modelo e sua capacidade preditiva (Agresti, 2007). As cinco principais funções de ligação estão descritas na tabela abaixo (Agresti, 2012):

Nome	**Função Link**$(\boldsymbol{F^{-1}})$
Logit	$log\left[\frac{P(Y\leq j)}{P(Y>j)}\right]$
Complemento Log-log	$log(-log(1-P(Y\leq j))$
Log-log negativo	$-log(-log(P(Y\leq j))$
Cauchit	$Tan(\pi(P(Y\leq j)-0,5))$
Probit	$\phi^{-1}(P(Y\leq j))$, onde ϕ é a função de distribuição da N(0,1)

Na prática, a função de ligação *Logit* é a mais utilizada, devido a sua interpretação interessante dos coeficientes do modelo e da sua matemática simples. Essa será a abordagem explorada neste texto.

O modelo é proposto através de uma analogia com a regressão logística usual, de forma que o *Logit* das probabilidades cumulativas são (Kutner et al., 2004):

$$Logit[P(Y_i\leq j|\boldsymbol{x})]=log\left[\frac{P(Y_i\leq j|\boldsymbol{x})}{1-P(Y_i\leq j|\boldsymbol{x})}\right]=\alpha_j-\beta_1x_{i1}-...-\beta_kx_{ik}\quad j=1,\cdots,J-1\quad . \tag{3.12}$$

Consequentemente,

$$P(Y_i\leq j|\boldsymbol{x})=\frac{exp(\alpha_j-\boldsymbol{x}_i^{'}\boldsymbol{\beta})}{1+exp(\alpha_j-\boldsymbol{x}_i^{'}\boldsymbol{\beta})}\quad j=1,\cdots,J-1\quad . \tag{3.13}$$

O modelo ordinal definido anteriormente permite estimar o logaritmo da probabilidade da variável dependente tomar os valores de classes inferiores ou iguais a j, comparativamente com a probabilidade de tomar os valores das classes superiores a j.

Para J=3, por exemplo, o modelo usa $Logit[P(Y_i \leq 1)|\boldsymbol{x}] = log[\pi_1/(\pi_2 + \pi_3)]$ e $Logit[P(Y_i \leq 2)|\boldsymbol{x}] = log[(\pi_1 + \pi_2)/\pi_3]$. Cada logito cumulativo utiliza todas as categorias da variável resposta.

Note que os coeficientes de regressão ($\boldsymbol{\beta} = \beta_1, \cdots, \beta_K$) em 3.12 não apresentam índice j, obrigando o modelo a pressupor que os efeitos das variáveis independentes sobre o $P(Y_i \leq j)$ é igual para todas as classes (Kutner et al., 2004). Assim, a resposta observada em cada classe apenas se encontra deslocada para a direita ou para a esquerda, em função de α_j. Isso resulta em um modelo mais parcimonioso. Para um $\beta_k > 0$, um aumento em algum X_k resulta na diminuição da probabilidade de a variável dependente tomar valores de ordem inferiores ou iguais a j (mantendo as demais variáveis explicativas constantes), ou seja, quando X_k aumenta, Y aumenta. Já para um $\beta_k < 0$, quando X_k aumenta, Y diminui.

A interpretação do modelo pode usar as razões de chance para as probabilidades cumulativas e os seus complementos (Agresti, 2012). Para dois valores x_1 e x_2 de uma das variáveis explicativas X_k do estudo, a razão de chances comparando as probabilidades cumulativas, para todas as classes da variável dependente, é dada por (mantendo as demais variáveis explicativas constantes):

$$\frac{P(Y \leq j|X_k = x_2)/P(Y > j|X_k = x_2)}{P(Y \leq j|X_k = x_1)/P(Y > j|X_k = x_1)} \quad . \tag{3.14}$$

O log dessa razão de chances é a diferença entre os logitos cumulativos para esses dois valores de X_k. Isso é igual a $-\beta_k(x_2 - x_1)$. Se $x_2 - x_1 = 1$, a chance da variável resposta assumir valores menores para qualquer categoria é multiplicado por $e^{-\beta_k}$ para cada unidade acrescida em X_k.

Os parâmetros do modelo $\alpha_1, \cdots, \alpha_{J-1}$ e $\boldsymbol{\beta}$ são estimados simultaneamente pelo método de Máxima Verossimilhança. Para isso, é necessário a obtenção da função de verossimilhança para os dados, lembrando que o modelo pressupõe que as curvas de probabilidade das $J-1$ classes da variável dependente são iguais para todas as classes e são calculadas de forma cumulativa.

A partir de 3.5, para n observações independentes e J categorias, a função de verossimilhança é dada por (Agresti, 2012)

$$P(y_1, \cdots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} [\pi_{ij}]^{y_{ij}} \right] = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} [P(Y_i \leq j|\boldsymbol{x}) - P(Y_i \leq j-1|\boldsymbol{x})]^{y_{ij}} \right] \quad . \tag{3.15}$$

Substituindo $P(Y_i \leq J|\boldsymbol{x}) = 1$, $P(Y_i \leq 0|\boldsymbol{x}) = 0$ e $P(Y_i \leq j|\boldsymbol{x})$, $j = 1, \cdots, J-1$, por 3.13 encontra-se a expressão desejada da função de verossimilhança, em termos de $\alpha_1, \cdots, \alpha_{J-1}$ e $\boldsymbol{\beta}$:

$$P(y_1, \cdots, y_n|\alpha_1, \cdots, \alpha_{J-1}, \boldsymbol{\beta}) =$$

$$= \prod_{i=1}^{n} \left[\left(\frac{exp(\alpha_1 - \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + exp(\alpha_1 - \boldsymbol{x}_i'\boldsymbol{\beta})} \right)^{y_{i1}} \left(\prod_{j=2}^{J-1} \left(\frac{exp(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + exp(\alpha_j - \boldsymbol{x}_i'\boldsymbol{\beta})} - \frac{exp(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})}{1 + exp(\alpha_{j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})} \right)^{y_{ij}} \right) \left(\frac{1}{1 + exp(\alpha_{J-1} - \boldsymbol{x}_i'\boldsymbol{\beta})} \right)^{y_{iJ}} \right]. \tag{3.16}$$

As estimativas de máxima verossimilhança são os valores dos parâmetros que maximizam 3.16. Não existem formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos como por exemplo o de Newton-Raphson (Kutner et al., 2004).

Capítulo 4

Inferência Bayesiana

Toda a teoria tratada nos capítulos anteriores são oriundas do paradigma estatístico clássico (ou frequentista). A essência desse paradigma está em pensar no processo aleatório que produziu os dados observados. Por exemplo, o lançamento de uma única moeda resulta em um de dois resultados possíveis: "cara" ou "coroa". O frequentista imagina o processo aleatório sendo repetido um número infinito de vezes, baseando-se na experiência da estabilidade da frequência relativa de ocorrência de eventos, quando são realizadas muitas repetições do experimento. Logo, para o exemplo da moeda, os dados consistem no número de caras, y, em n lançamentos independentes. A proporção de caras é y/n. Os frequentistas formulam o conceito de probabilidade de cara θ como o valor de y/n, com n tendendo ao infinito. Consequentemente o parâmetro θ pode ser estimado a partir dos dados observados, por exemplo, através da maximização da função de verossimilhança $P(y|\theta, n)$ ($\hat{\theta} = y/n$).

Essa idéia de repetição do processo aleatório gerador dos dados ou, equivalentemente, da amostragem repetida, é fundamental para a abordagem clássica. Segundo ela, os métodos estatísticos devem derivar-se através do respectivo comportamento num número indefinido de repetições - hipotéticas - efetuadas nas mesmas condições.

Sustenta-se assim a definição de duas estatísticas muito utilizadas: p-valores e intervalos de confiança (Dobson and Barnett, 2008).

As definições frequentistas para o p-valor e para um intervalo de confiança de $(1-\alpha)\times 100\%$ são, respectivamente:

1. O p-valor é a probabilidade de observar dados mais extremos de uma estatística daqueles que foram observados na amostra (se o processo aleatório for repetido) dado que a hipótese nula é verdadeira.

2. Se n repetições do processo aleatório fossem repetidas um número grande de vezes, e para cada uma se construisse um intervalo de confiança de $(1-\alpha)\times 100\%$, espera-se que $(1-\alpha)\times 100\%$ dos intervalos gerados conteriam o real valor do parâmetro desejado. Por definição, o número esperado de vezes que um intervalo de confiança de $(1-\alpha)\times 100\%$ não contém o valor real do parâmetro é $\alpha\times 100\%$ das vezes.

Como visto, as definições acima dependem do conceito de múltiplas repetições do processo aleatório gerador dos dados. Alternativamente, seriam mais naturais as definições que contam somente com os dados observados em mãos. Consequentemente, as definições ideais seriam (Dobson and Barnett, 2008):

1. O p-valor é a probabilidade estimada da hipótese nula ser verdadeira, dado os dados observados.

2. Um intervalo de confiança de $(1-\alpha)\times 100\%$ é um intervalo que contém o valor real do parâmetro desejado com uma probabilidade de $(1-\alpha)$.

Essas definições ideais são possíveis a partir de um paradigma estatístico alternativo: a análise **bayesiana**.

4.1 O Paradigma Bayesiano

Os métodos bayesianos passam, em certo sentido, por uma extensão do modelo clássico, extensão que tem raiz na seguinte divergência fundamental (Paulino et al., 2003):

No modelo clássico o parâmetro $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, é um escalar ou vetor desconhecido, mas fixo, isto é, igual ao valor particular que indexa a distribuição da família $\boldsymbol{F}$ que descreve "apropriadamente" o processo que gera as observações. No modelo bayesiano $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, é tomado como um escalar ou vetor aleatório (não observável). A filosofia bayesiana é, nesse ponto, a seguinte: o que é desconhecido é incerto, e toda a incerteza deve ser quantificada em termos de probabilidade. A idéia passa a ser a de tentar reduzir esta incerteza.

A intensidade da incerteza a respeito de $\boldsymbol{\theta}$ pode assumir diferentes graus. No paradigma bayesiano, esses diferentes graus de incerteza são representados através de modelos probabilísticos para $\boldsymbol{\theta}$, $h(\boldsymbol{\theta})$ (Ehlers, 2007), designados de distribuição *a priori* de $\boldsymbol{\theta}$. Essa distribuição de probabilidade, geralmente subjetiva, representa a informação inicial, anterior ou externa em relação à experiência (Paulino et al., 2003). Assim, diferentes pesquisadores e especialistas em determinado assunto podem ter diferentes graus de incerteza sobre $\boldsymbol{\theta}$ (especificando distribuições distintas).

Essa informação que dispoe-se sobre $\boldsymbol{\theta}$, resumida probabilisticamente através de $h(\boldsymbol{\theta})$, pode ser aumentada observando-se uma quantidade aleatória $\boldsymbol{Y}$ relacionada

com $\boldsymbol{\theta}$ (Paulino et al., 2003). A distribuição amostral $P(\boldsymbol{y}|\boldsymbol{\theta})$ (verossimilhança) define esta relação. A idéia de que após observar $\mathbf{Y}$ a quantidade de informação sobre $\boldsymbol{\theta}$ aumenta é bastante intuitiva, e o Teorema de Bayes é a regra de atualização utilizada para quantificar este aumento de informação (Dobson and Barnett, 2008):

$$h(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{h(\boldsymbol{y})} \quad . \tag{4.1}$$

Essa equação fornece a probabilidade *a posteriori* de $\boldsymbol{\theta}$ (dado os dados) , como função da *verossimilhança*, $P(\boldsymbol{y}|\boldsymbol{\theta})$, e da distribuição *a priori* de $\boldsymbol{\theta}$, $h(\boldsymbol{\theta})$.

Como o denominador em 4.1 não depende de $\boldsymbol{\theta}$, ele é simplesmente uma constante normalizadora de $h(\boldsymbol{\theta}|\boldsymbol{y})$, podendo-se reescrever a fórmula de Bayes em termos proporcionais como (Dobson and Barnett, 2008)

$$h(\boldsymbol{\theta}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{\theta})h(\boldsymbol{\theta}), \tag{4.2}$$

ou seja,

distribuição a posteriori $\propto$ verossimilhança$\times$ distribuição a priori

Como se verifica, a função de verossimilhança tem papel importante na fórmula de Bayes pois representa o meio através do qual os dados $\boldsymbol{y}$ transformam o conhecimento a priori sobre $\boldsymbol{\theta}$ (Paulino et al., 2003). Ela é também a chave para a metodologia frequentista, mas os parâmetros são tratados como fixos nessa abordagem.

Em resumo, para os bayesianos, a distribuição *a posteriori* incorpora, através do Teorema de Bayes, toda a informação disponível sobre o parâmetro (informação

inicial + informação da experiência ou da amostra). Daqui decorre que todos os procedimentos de inferência bayesiana são baseados exclusivamente em $h(\boldsymbol{\theta}|\boldsymbol{y})$.

A seguir serão descritos os insumos e definições para a formulação posterior do modelo de regressão ordinal em um contexto bayesiano.

4.2 Prioris

A mais clara diferença entre as abordagens bayesiana e clássica começa no fato da primeira se basear num mecanismo formal que permite a introdução de informação anterior ou externa à experiência em questão (informação *a priori*), e a sua combinação, após apropriada formalização, com a informação contida nos dados resultantes dessa experiência (Paulino et al., 2003).

A informação *a priori* que se pretende incorporar na análise é a informação de um especialista do problema que contém, até dominantemente, elementos subjetivos. Para a realização da análise bayesiana, deve-se tentar traduzir tal informação numa distribuição *a priori* subjetiva, havendo diversos métodos na literatura para tal fim.

Há, no entanto, casos em que não existe informação *a priori* palpável ou em que o conhecimento *a priori* é pouco significativo relativamente à informação amostral, conduzindo-se a distribuições *a priori* minimamente informativas e que, em geral, são apelidadas de **distribuições não informativas**.

A primeira idéia de não informação *a priori* que se pode ter é pensar em todos os possíveis valores de $\boldsymbol{\theta}$ como igualmente prováveis, isto é, com uma distribuição *a priori* uniforme (Dobson and Barnett, 2008). Nesse caso $h(\boldsymbol{\theta}) \propto c$, então $h(\boldsymbol{\theta}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{\theta})$. Logo, utilizar uma priori vaga ou não informativa fornece uma posteriori

que é completamente dependente dos dados. Nesse caso, os resultados das análises bayesiana e frequentista são muito semelhantes, embora os resultados bayesianos ainda possuam a interpretação "ideal" dos p-valores e dos intervalos de confiança, vistos no começo do capítulo. Essa escolha de prioris, no entanto, pode trazer algumas dificuldades técnicas, tais como se o intervalo de variação de $\boldsymbol{\theta}$ for ilimitado então sua distribuição é imprópria, ou seja, ela não integra para 1 (Ehlers, 2007).

O problema mais grave ocorre quando a distribuição *a posteriori* do parâmetro de interesse é imprópria, inviabilizando a realização de inferências (Paulino et al., 2003). Infelizmente não se conhecem condições que assegurem distribuições *a posteriori* próprias em situações gerais com distribuições *a priori* impróprias, e a verificação de que a *posteriori* é própria nem sempre é trivial. Devido a estes problemas costuma-se substituir as distribuições *a priori* impróprias por distribuições próprias difusas que as aproximem, como é o caso de distribuições Normais com uma enorme variância (Paulino et al., 2003).

4.3 Estimativas Pontuais

A distribuição *a posteriori* dos parâmetros $\boldsymbol{\theta}$ de um modelo contém toda a informação a respeito destes parâmetros. No entanto, ás vezes é necessário resumir a informação contida na posteriori através de um valor numérico, chamado de estimativa pontual. Assim, toda posteriori é resumida em um valor para cada parâmetro (Ehlers, 2007).

Na estatística clássica o problema de estimação é resolvido, por exemplo, encontrando-se o estimador de máxima verossimilhança da quantidade desconhe-

cida (porém fixa) $\boldsymbol{\theta}$. A abordagem bayesiana adota o conceito de Função Perda como auxilio na escolha do estimador de $\boldsymbol{\theta}$.

A Função Perda, denotada por $\mathrm{L}(\boldsymbol{a}, \boldsymbol{\theta})$, determina a perda sofrida ao se tomar a decisão $\boldsymbol{a}$ (que é uma função do resultado observado da experiência aleatória) dado o real estado $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Esta perda é expressa como um número real (cuja interpretação física pode ser, por exemplo, de uma perda monetária).

Para fins de estimação, a ação $\boldsymbol{a} = \boldsymbol{a}^*$ indica o valor da estimativa usada para o parâmetro $\boldsymbol{\theta}$. Ou seja, com base no resultado $\boldsymbol{y}$ da experiência aleatória, escolhe-se uma ação $\boldsymbol{a}^*$, que resulta numa perda $\mathrm{L}(\boldsymbol{a}^*, \boldsymbol{\theta})$.

A perda esperada a posteriori é dada por (Ehlers, 2007)

$$E[L(\boldsymbol{a}^*,\boldsymbol{\theta}|\boldsymbol{y})] = \int L(\boldsymbol{a}^*,\boldsymbol{\theta}|\boldsymbol{y})h(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

e o objetivo é o de escolher $\boldsymbol{a}^*$ de tal forma que a perda esperada a posteriori seja minimizada. $\boldsymbol{a}^*$ é chamado de **estimador de Bayes**, e depende da função perda que é usada.

Uma escolha usual para a função perda é a *Função Perda Quadrática*, definida como $\mathrm{L}(\boldsymbol{a}, \boldsymbol{\theta}) = (\boldsymbol{a} - \boldsymbol{\theta})^2$. O estimador de Bayes dessa função perda é a média a posteriori (Paulino et al., 2003).

4.4 Estimação por Intervalos

Com a estimação pontual, toda a informação contida na posteriori é resumida em um único número (ou em k números referentes ao número de parâmetros). É importante também associar alguma informação sobre o quão precisa é a especificação deste

número, e esse é o papel dos intervalos de credibilidade, ou "intervalos de confiaça" bayesianos.

Um intervalo de credibilidade fornece a probabilidade de um dado parâmetro pertencer a este intervalo. Desse modo, podemos especificar, digamos, um intervalo de probabilidade de $1 - \alpha$ e, então, de posse da distribuição da quantidade de interesse, determinar os limites que estabelecem essa probabilidade (Ehlers, 2007). Formalmente:

- **Intervalo de Credibilidade**: C é um intervalo de credibilidade $1 - \alpha$ para θ se $P(\theta \in C) \geq 1 - \alpha$.

A exigência de que a probabilidade acima possa ser maior que o nível de credibilidade é essencialmente técnica, visto que o desejo é o de menor intervalo possível. No entanto, a desigualdade será útil para θ com distribuição discreta, quando nem sempre é possível satisfazer a igualdade.

Uma infinidade de intervalos fica possível através da definição acima, mas o interesse está no intervalo com menor comprimento possível. Pode-se mostrar que os intervalos de comprimento mínimo são obtidos tomando-se os valores de θ com maior densidade a posteriori. Esses são os intervalos HPD (Highest Posterior Density) (Ehlers, 2007):

- **Intervalo HPD**: Um intervalo de credibilidade C de $(1 - \alpha)\%$ para θ é HPD se $C = \{\theta \in \Theta : h(\theta|\boldsymbol{y}) \geq k(\alpha)\}$ onde $k(\alpha)$ é a maior constante tal que $P(\theta \in C) \geq 1 - \alpha$.

Com essa definição, todos os pontos dentro do intervalo HPD terão densidade a

posteriori maior do que qualquer ponto fora do intervalo.

4.5 Testes de Hipóteses

O teste de hipóteses é um dos pilares da inferência estatística. A sua utilização é praticamente certa nas pesquisas dos mais variados campos do conhecimento.

Uma hipóteses estatística é uma afirmação ou conjetura sobre o(s) parâmetro(s) $\boldsymbol{\theta}$ da distribuição de probabilidades de uma característica da população.

Sendo $\boldsymbol{\theta}$ o parâmetro (ou vetor de parâmetros) de interesse, e seja $\boldsymbol{\Theta}$ o seu espaço paramétrico (correspondente ao conjunto de valores que $\boldsymbol{\theta}$ pode assumir), o problema de testar $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta_0}$ contra $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta_1} = \boldsymbol{\Theta} - \boldsymbol{\Theta_0}$ num cenário bayesiano é conceitualmente mais simples do que num contexto clássico. Dada a interpretação probabilística direta das hipóteses em confronto, é suficiente calcular as respectivas probabilidades à posteriori (Paulino et al., 2003):

- $P(H_0|\boldsymbol{y}) = P(\boldsymbol{\theta} \in \boldsymbol{\Theta_0}|\boldsymbol{y})$;
- $P(H_1|\boldsymbol{y}) = P(\boldsymbol{\theta} \in \boldsymbol{\Theta_1}|\boldsymbol{y}) = 1 - P(H_0|\boldsymbol{y})$.

Caso se pretenda optar por uma das hipóteses em função de uma grandeza relativa, pode-se calcular o Odds (Chance) à posteriori a favor de H_0 (ou de H_0 sobre H_1) (Paulino et al., 2003):

$$O(H_0, H_1|\boldsymbol{y}) = \frac{P(H_0|\boldsymbol{y})}{P(H_1|\boldsymbol{y})} = \frac{P(H_0|\boldsymbol{y})}{1 - P(H_0|\boldsymbol{y})} \quad . \tag{4.3}$$

A evidência a favor de H_0 antes da observação dos dados pode ser calculada de forma análoga, denominada Odds à priori:

$$O(H_0, H_1) = \frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{1 - P(H_0)} \quad . \tag{4.4}$$

A fim de se avaliar a influência dos dados observados $\boldsymbol{y}$ na alteração da credibilidade relativa (a priori) de H_0 e H_1, opta-se por contrapor a razão a posteriori $O(H_0, H_1|\boldsymbol{y})$ com a razão das probabilidades a priori $O(H_0, H_1)$ através da divisão dessas razões (Odds Ratio ou Razão de chances):

$$B(\boldsymbol{y}) = \frac{O(H_0, H_1|\boldsymbol{y})}{O(H_0, H_1)} = \frac{\frac{P(H_0|\boldsymbol{y})}{1-P(H_0|\boldsymbol{y})}}{\frac{P(H_0)}{1-P(H_0)}} \quad . \tag{4.5}$$

$B(\boldsymbol{y})$ é denominado **Fator de Bayes** a favor de H_0 (ou contra H_1). Aplicando o logaritmo no Fator de Bayes, obtem-se uma relação aditiva entre as quantidades transformadas:

$$lnB(\boldsymbol{y}) = lnO(H_0, H_1|\boldsymbol{y}) - lnO(H_0, H_1) \quad . \tag{4.6}$$

Essas quantidades são chamadas de pesos de evidência. Assim, o Fator de bayes $B(\boldsymbol{y})$ ou o seu logaritmo representam um peso relativo da evidência contida nos dados observados a favor de uma ou outra das hipóteses em confronto. $lnB(\boldsymbol{y})$ é visto como o peso da evidência à posteriori descontado do correspondente peso da evidência à priori (Paulino et al., 2003).

Valores do Fator de Bayes muito grandes ou muito pequenos em relação a 1 representam uma tendência forte nos dados a favor de uma hipótese contra a outra, ou seja, uma hipótese é muito mais ou muito menos provável do que era a priori.

4.5.1 Hipóteses Precisas

O problema de se testar hipóteses precisas do tipo $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ contra $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta_0}$, onde $\boldsymbol{\theta_0} \in \boldsymbol{\Theta}$ (muito comum em regressão, quando é testada a significância de um dado coeficiente) fica inviabilizada com a metodologia atrás definida quando $\boldsymbol{\Theta}$ é um

espaço paramétrico contínuo (Paulino et al., 2003). Isso se deve à impossibilidade de calcular-se o Fator de Bayes, visto que $P(\boldsymbol{\theta} = \boldsymbol{\theta_0}) = 0 \quad \forall \quad \boldsymbol{\theta_0} \in \boldsymbol{\Theta}$.

Há argumentos de que esse tipo de problema é irrealista e que a hipótese nula não é senão uma representação abreviada de uma bola centrada em $\boldsymbol{\theta_0}$, $V_\epsilon(\boldsymbol{\theta_0})$, com raio ϵ escolhido de modo que os pontos sejam praticamente indistinguíveis a posteriori de $\boldsymbol{\theta_0}$. Uma vez transformando o problema para $H_0 : \boldsymbol{\theta} \in V_\epsilon(\boldsymbol{\theta_0})$ contra $H_1 : \boldsymbol{\theta} \notin V_\epsilon(\boldsymbol{\theta_0})$, fica-se em condições de se aplicar o procedimento baseado nas chances *a posteriori* (Paulino et al., 2003).

No entanto, consolidou-se na literatura a abordagem de Jeffreys, Savage e Good, que defendem o fato de $\boldsymbol{\theta_0}$ possuir, à priori, uma ordem de importância diferente da que é atribuída aos demais valores de $\boldsymbol{\theta}$, devido a própria formulação de $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$.

Dessa forma, esse juízo deve ser formalizado e integrado na distribuição a priori, que passará a ter uma natureza mista com uma massa pontual concentrada em $\boldsymbol{\theta_0}$, $p_0 = P\{H_0\}$, e uma distribuição contínua da massa restante, $1 - p_0 = P\{H_1\}$, em $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$. Ou seja, sendo $h(\boldsymbol{\theta})$ a priori para $\boldsymbol{\theta}$ tem-se

$$h(\boldsymbol{\theta}) = \begin{cases} p_0, & \text{se} \quad \boldsymbol{\theta} = \boldsymbol{\theta_0} \\ (1 - p_0)h_1(\boldsymbol{\theta}), & \text{se} \quad \boldsymbol{\theta} \neq \boldsymbol{\theta_0}, \end{cases}$$

onde $h_1(\boldsymbol{\theta})$ é a distribuição à priori dos valores de $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$.

Consequentemente, a distribuição a posteriori mista de $\boldsymbol{\theta}$ pode ser definida como

$$h(\boldsymbol{\theta}|\boldsymbol{y}) = \begin{cases} \frac{p_0 P(\boldsymbol{y}|\boldsymbol{\theta_0})}{h(\boldsymbol{y})}, & \text{se} \quad \boldsymbol{\theta} = \boldsymbol{\theta_0} \\ \frac{(1-p_0)h_1(\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{\theta})}{h(\boldsymbol{y})}, & \text{se} \quad \boldsymbol{\theta} \neq \boldsymbol{\theta_0}, \end{cases}$$

onde $h(\boldsymbol{y}) = p_0 P(\boldsymbol{y}|\boldsymbol{\theta_0}) + (1 - p_0) \int_{\boldsymbol{\theta} \neq \boldsymbol{\theta_0}} h_1(\boldsymbol{\theta}) P(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$.

Assim, tem-se o Fator de Bayes:

$$B(\boldsymbol{y}) = \frac{O(H_0|\boldsymbol{y})}{O(H_0)} = \frac{\frac{P(H_0|\boldsymbol{y})}{1-P(H_0|\boldsymbol{y})}}{\frac{P(H_0)}{1-P(H_0)}} = \frac{\frac{P(\theta=\theta_0|\boldsymbol{y})}{1-P(\theta=\theta_0|\boldsymbol{y})}}{\frac{P(\theta=\theta_0)}{1-P(\theta=\theta_0)}} = \frac{\frac{\frac{p_0 P(\boldsymbol{y}|\boldsymbol{\theta_0})}{h(\boldsymbol{y})}}{1-\frac{p_0 P(\boldsymbol{y}|\boldsymbol{\theta_0})}{h(\boldsymbol{y})}}}{\frac{p_0}{1-p_0}} \quad . \tag{4.7}$$

Capítulo 5

FBST

A utilização da medida de evidência clássica denominada p-valor já é consagrada nas pesquisas e publicações dos mais variados campos do conhecimento. No entanto, muito é discutido na literatura acerca dos conflitos entre essa medida e as medidas de evidência bayesianas, alertando para o fato de que em algumas situações o p-valor pode não ser uma boa medida de evidência para uma hipótese estatística precisa.

Tal incompatibilidade entre p-valores e medidas de evidência bayesianas em testes de hipóteses precisas começaram a ser discutidas em Jeffreys (1961). Nele, Jeffreys aponta que uma hipótese nula precisa pode ser fortemente rejeitada utilizando-se testes de significância clássicos enquanto a sua probabilidade posterior em uma abordagem bayesiana é alta (dando evidências a seu favor), baseando-se em uma priori pouco provável para a hipótese nula e uma distribuição difusa da probabilidade remanescente para a hipótese alternativa (Shafer, 1976). Esse conflito entre as abordagens clássica e bayesiana ficou conhecido como *Paradoxo de Lindley* após Dennis Lindley intitular o conflito como paradoxo em Lindley (1957). Esse paradoxo, no entanto, estava baseada em grandes tamanhos de amostra e (até certo ponto) no pressuposto de que é plausível para $\boldsymbol{\theta}$ se igualar a $\boldsymbol{\theta_0}$ exatamente (Berger and Selke,

1987).

Ampliações que não necessitavam necessariamente da premissa de grandes amostras e nem da exatidão na hipótese nula começaram a ser desenvolvidas em Edwards et al. (1963) e Dickey (1977) e, posteriormente, expandidas em Berger and Selke (1987) e Berger and Delampady (1987). É ilustrado que a evidência contra uma hipótese nula (medida através da probabilidade posterior, fator de Bayes ou da verossimilhança comparativa) pode diferir drasticamente do p-valor, sendo essa conflito exarcebado com grandes amostras (sendo o extremo ilustrado pelo *Paradoxo de Lindley*). A conclusão geral tirada em Berger and Selke (1987) e Berger and Delampady (1987) é que os p-valores podem ser medidas altamente enganosas acerca da evidência trazida pelos dados contra a hipótese nula precisa, em determinadas situações.

O desentendimento entre as abordagens é causado por uma série de de fatores. Primeiramente, a abordagem clássica do p-valor não considera no seu cálculo a hipótese alternativa (Pereira and Stern, 1999), enquanto na abordagem bayesiana são envolvidas as duas hipóteses. A atribuição de uma dada probabilidade *a priori* para a hipótese nula atenua o conflito, visto que a probabilidade remanescente distribuida de uma forma difusa para a hipótese alternativa pode resultar em uma probabilidade posterior pequena para essa hipótese, favorecendo assim a hipótese nula. Lindley (1997) discute que a utilização do Fator de Bayes para testes de significância em hipóteses precisas é controverso.

Os conflitos citados somados com o desconforto de se trabalhar com um modelo

misto fez com que fossem desenvolvidas novas medidas de evidência. Proposta por Pereira and Stern (1999), a medida de evidência genuinamente bayesiana é obtida através do procedimento denominado *Full Bayesian Significance Test* ($FBST$). O termo "genuinamente bayesiano" decorre do fato da medida de evidência proposta ser baseada somente na distribuição posterior.

O $FBST$ testa as hipóteses baseando-se no cálculo da probabilidade *a posteriori* da região HPD ($HighestPosteriorDensity$) que é "tangente" ao conjunto que define a hipótese nula. Segundo Pereira and Stern (1999) o teste é intuitivo, com fácil caracterização geométrica e pode ser implementado a partir de técnicas de otimização e integração numérica. Define-se então a medida de evidência em favor de uma hipótese precisa como:

- **Definição**: Considere uma variável aleatória $\boldsymbol{Y}$ e um modelo estatístico paramétrico a ela associado, isto é, uma quíntupla ($\mathcal{Y}$,A,F,$\boldsymbol{\Theta}$,h), onde $\mathcal{Y}$ é o espaço amostral, conjunto dos possíveis valores de $\boldsymbol{y}$, A é uma sigma-álgebra conveniente de subconjuntos de $\mathcal{Y}$, F é uma classe de distribuições de probabilidade em A, indexadas no espaço paramétrico $\boldsymbol{\Theta}$ e h é uma densidade a priori em $\boldsymbol{\Theta}$. Considere também que, após a observação de $\boldsymbol{y}$, obtem-se a densidade *a posteriori* de $\boldsymbol{\theta}$, $h(\boldsymbol{\theta}|\boldsymbol{y})$, restringindo essa a uma função densidade de probabilidade.

  Seja T_ϕ o subconjunto do espaço paramétrico onde a densidade posterior é maior do que ϕ:

$$T_\phi = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} | h(\boldsymbol{\theta}|\boldsymbol{y}) > \phi\} \quad .$$

A credibilidade de T_ϕ é a sua probabilidade a posteriori: $\kappa = \int_{T_\phi} h(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$. Agora, seja h^* o máximo da densidade posterior sob a hipótese nula (H_0), obtido através de $\boldsymbol{\theta^*}$:

$$\boldsymbol{\theta^*} \in sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta_0}} h(\boldsymbol{\theta}|\boldsymbol{y}), \quad h^* = h(\boldsymbol{\theta^*}|\boldsymbol{y}) \quad .$$

Assim, $T^* = T_{h^*}$ é o conjunto "tangente" à hipótese nula com credibilidade κ^*.

A medida de evidência de Pereira-Stern em favor de H_0, sendo H_0 um subconjunto $\boldsymbol{\Theta_0}$ de $\boldsymbol{\Theta}$ (com $dim(\boldsymbol{\Theta_0}) < dim(\boldsymbol{\Theta})$), é a probabilidade complementar do conjunto T^*

$$ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) = 1 - \kappa^* = 1 - P(\boldsymbol{\theta} \in T^*|\boldsymbol{y}) \quad , \tag{5.1}$$

e o procedimento (ou teste) $FBST$ consiste em aceitar H_0 sempre que $ev(\boldsymbol{\Theta_0}; \boldsymbol{y})$ é grande.

Percebe-se então que a medida de evidência proposta em favor de uma hipótese nula (H_0) considera todos os pontos do espaço paramétrico cujos valores da densidade posterior são, no máximo, tão grandes quanto seu supremo em $\boldsymbol{\Theta_0}$. Um valor grande de $ev(\boldsymbol{\Theta_0}; \boldsymbol{y})$ significa que o subconjunto $\boldsymbol{\Theta_0}$ cai em uma região do espaço paramétrico de alta probabilidade posterior, ou seja, T^* tem probabilidade posterior "pequena", portanto, os dados suportam a hipótese nula H_0. Por outro lado, um valor pequeno da evidência levaria à rejeição da hipótese nula (Pereira and Stern, 1999).

Interessante notar que, enquanto na abordagem clássica dos p-valores o olhar está no conjunto C dos pontos amostrais no mínimo tão inconsistentes a H_0 quanto $\boldsymbol{y}$, através do $FBST$ olha-se o conjunto tangente T^* de pontos no espaço paramétrico que são mais consistentes com $\boldsymbol{y}$ do que H_0 (Pereira et al., 2008).

Embora a definição da medida de evidência aqui estabelecida tenha sido feita para hipóteses em que a dimensão é menor do que a do espaço paramétrico, nada impede a sua utilização para o caso $dim(\boldsymbol{\Theta_0}) = dim(\boldsymbol{\Theta})$.

5.1 Propriedades do FBST

Madruga et al. (2001) mostrou que o $FBST$ respeita alguns princípios estatísticos conhecidos, e demonstrou algumas propriedades desse procedimento. Em Pereira et al. (2008) também listou-se propriedades desejáveis de um teste estatístico que são satisfeitas pelo e-valor e no seu resultante $FBST$. De uma forma geral, essas propriedades estão listadas abaixo:

1- *O $FBST$ não viola o Princípio da Verossimilhança*

O Princípio da Verossimilhança estabelece que, se tem-se dois modelos estatísticos, o primeiro dando observações $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ de uma variável aleatória $\boldsymbol{X}$, com função densidade de probabilidade (ou função de probabilidade, no caso discreto) $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$, e o segundo dando observações $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, de uma variável aleatória $\boldsymbol{Y}$, com função densidade de probabilidade (ou função de probabilidade, no caso discreto) $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})$, sendo o espaço paramétrico $\boldsymbol{\Theta}$ comum aos dois modelos. Se

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = K(x, y) f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

com $\boldsymbol{K}$ independente de $\boldsymbol{\theta}$, então as inferências feitas para $\boldsymbol{\theta}$ devem ser idênticas tanto por $\boldsymbol{x}$ quanto por $\boldsymbol{y}$.

Alguns autores (Paulino et al. (2003), por exemplo) mostram exemplos de violação deste princípio na abordagem clássica, sendo essa contraditória a esse princípio. Já os procedimentos "bayesianos" o seguem automaticamente.

Como mencionado anteriormente, a utilização inferencial do e-valor é um procedimento genuinamente bayesiano. Ele é, de fato, uma probabilidade *a posteriori* bem definida de um subconjunto do espaço paramétrico (Pereira et al., 2008).

No entanto, para haver uma credibilidade definitiva em tal definição, o $FBST$ deve ser caracterizado dentro da abordagem da Teoria da Decisão. Devem ser chamados "bayesianos" apenas os procedimentos que minimizam funções perdas esperadas, a solução coerente para o problema de decisão. Assim, seja a determinação de um valor de K (ponto crítico cujo valor depende da função perda escolhida) de modo que:

- **Rejeitar** H_0 se $ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) \leq K$;
- **Aceitar** H_0 se $ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) > K$.

Por exemplo, Madruga et al. (2001) consideram $\mathbf{D}$ o espaço de decisões tal que $\mathbf{D} = \{AceitarH_0(d_0), RejeitarH_0(d_1)\}$, e L a função perda em $\boldsymbol{D} \times \boldsymbol{\Theta} \to \Re^+$ definida por:

$$L(RejeitarH_0, \theta) = a[1 - \mathbf{1}(\theta \in T^*)] \quad e$$

$$L(AceitarH_0, \theta) = b + c\mathbf{1}(\theta \in T^*), \quad com \quad a, b, c > 0. \tag{5.2}$$

Nota-se que a função de perda L pune severamente a decisão de aceitar H_0 quando $\boldsymbol{\theta}$ é, de fato, mais "provável" que $\boldsymbol{\theta_0}$, ou seja, quando $\boldsymbol{\theta} \in T^*$. Madruga et al. (2001) prova que a minimização da esperança posterior da função de perda definida acima, bem como de outras funções de perda, sob H_0, levam ao procedimento $FBST$, o que o torna completamente bayesiano e confirma a não violação do Princípio da Verossimilhança.

2- *O $FBST$ não viola o Princípio da Surpresa Mínima*

O $FBST$ respeita, também, o Princípio da Surpresa Mínima, como sugerido por Good (1988), pois defende que os pontos mais importantes são aqueles mais apoiados pelos dados, portanto, são os pontos do espaço paramétrico que têm maiores valores da densidade posterior.

3- *O $FBST$ está de acordo com o princípio jurídico Onus Probandi*

O princípio jurídico *Onus Probandi*, também conhecido como *Benefício da Dúvida* ou *In Dubio Pro Reo* parte do princípio que toda afirmação precisa de sustentação, de provas para ser levada em consideração, e quando não são oferecidos, essa afirmação não tem valor argumentativo e deve ser desconsiderada em um raciocínio lógico. Assim, uma hipótese em questão não é rejeitada se não há evidências suficientes contra ela.

4- *A necessidade de aproximações no cálculo do e-valor se restringe a maximizações e integrações numéricas*

Métodos assintóticos não são necessários para o cálculo do e-valor, cálculo esse

que sempre envolve dois passos: otimização restrita e integração da distribuição posterior.

5- *O $FBST$ não requer a atribuição de probabilidades a priori positivas para as hipóteses precisas e nem a eliminação de parâmetros de confusão*

O $FBST$ não exige a adoção de uma distribuição *a priori* que associa uma probabilidade positiva para o subconjunto que define a hipótese nula precisa. Esta é uma característica de coerência mais relevante do $FBST$ em relação ao teste de Jeffreys para hipóteses precisas.

A distribuição posterior é suficiente para o cálculo do e-valor, sem complicações quanto a dimensão tanto dos parâmetros quanto do espaço amostral. Essa característica evita a necessidade da eliminação de parâmetros de perturbação, um problema que afeta algumas aplicações, como pode ser visto em Basu (1977).

6- *A hipótese alternativa também é considerada no cálculo do e-valor*

Como o espaço paramétrico completo é usado no cálculo do e-valor, a hipótese alternativa é sempre inerentemente considerada. Como citado em Pereira and Wechsler (1993), os testes de significância clássicos ás vezes desconsideram a hipótese alternativa, afetando as inferências.

7- *Sob as condições, i) o valor verdadeiro do parâmetro deve estar contido no suporte da distribuição a priori e ii) existência de um estimador consistente de $\boldsymbol{\theta}$, a medida de evidência é consistente, ou seja, é tal que*

$$lim_{n\to\infty} ev(\boldsymbol{\Theta_0}, \boldsymbol{y}) = \begin{cases} 1, & \boldsymbol{\theta} = \boldsymbol{\theta_0} \\ 0, & \boldsymbol{\theta} \neq \boldsymbol{\theta_0}. \end{cases}$$

Isso indica que, quando a dimensão da amostra é grande, o procedimento de Pereira-Stern conduz o experimentador à decisão correta. Essa propriedade é uma consequência direta da propriedade de *consistência da distribuição posterior*, discutida e demostrada por vários autores, por exemplo, em Bernardo and Smith (1994).

8- *A medida de evidência corrigida é invariante sob transformações um-a-um do parâmetro de interesse*

A medida de evidência é invariante apenas sob transformações lineares. Porém, pode ser feita uma correção a fim de ter-se invariância sob qualquer transformação um-a-um do parâmetro de interesse (Madruga et al., 2001). Essa correção se baseia na substituição da região T^* pela região S^*, dada por

$$S^* = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} : \frac{h(\boldsymbol{\theta}|\boldsymbol{y})}{h(\boldsymbol{\theta})} > sup_{\boldsymbol{\Theta_0}} \frac{h(\boldsymbol{\theta}|\boldsymbol{y})}{h(\boldsymbol{\theta})} \right\}$$

e do cálculo da evidência (corrigida) em favor de H_0 como $ev_c(\boldsymbol{\Theta_0}, \boldsymbol{y}) = 1 - P(\boldsymbol{\theta} \in S^*|\boldsymbol{y})$ (Madruga et al., 2001). A medida de evidência corrigida coincide com a medida sem correção quando se utiliza uma priori uniforme. Ela considera desfavorável aqueles pontos que têm aumento relativo na probabilidade à priori, com relação à posteriori, maior que o aumento relativo máximo sob H_0.

5.2 Regra de Decisão no $FBST$

Com a regra de decisão descrita anteriormente (correspondente à rejeição de H_0 se $ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) \leq K$) e a função perda definida em 5.2, Madruga et al. (2001) mostram que o valor de corte é $K = \frac{(b+c)}{(a+c)}$.

Na prática, a escolha dos valores de a, b e c não é simples e envolve a opinião do pesquisador sobre o erro mais (ou menos) danoso na sua decisão.

Outra alternativa, baseando-se nas idéias de Santis (2004), é a de estabelecer uma escala de evidência em termos do $ev(\boldsymbol{\Theta_0}; \boldsymbol{y})$, levando as seguintes regras de decisão do $FBST$:

- se $ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) < w_1$, rejeita-se H_0 (evidência devisiva);
- se $w_1 \leq ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) \leq w_0$, não decisão (evidência fraca);
- se $ev(\boldsymbol{\Theta_0}; \boldsymbol{y}) > w_0$, aceita-se-se H_0 (evidência decisiva).

com $w_i \in [0; 1](i = 0; 1)$ representando os pontos de corte do $ev(\boldsymbol{\Theta_0}; \boldsymbol{y})$.

De uma forma geral, nos casos em que a evidência obtida é muito próxima de zero (ou de 1), a decisão natural é rejeitar (ou aceitar) H_0. Nas demais situações, pode-se estabelecer o nível de significância do teste, como é feito nos testes clássicos.

5.3 Implementação do $ev(\boldsymbol{\Theta_0}; \boldsymbol{y})$

A determinação de $ev(\boldsymbol{\Theta_0}; \boldsymbol{y})$ envolve as duas etapas descritas a seguir:

- **1ª Etapa** - *Etapa de Otimização:* Consiste em maximizar a densidade posterior $h(\boldsymbol{\theta}|\boldsymbol{y})$ sob H_0. Em outras palavras, consiste em obter o valor de $\boldsymbol{\theta^*}$ que maximiza a densidade posterior, ou seja,

$$h(\boldsymbol{\theta^*}|\boldsymbol{y}) = sup_{\boldsymbol{\Theta_0}} h(\boldsymbol{\theta}|\boldsymbol{y}) \quad .$$

- **2ª Etapa** - *Etapa de Integração:* Consiste em integrar a densidade posterior $h(\boldsymbol{\theta}|\boldsymbol{y})$ sob o conjunto complementar a T^*, ou seja,

$$I = \int_{T^{*C}} h(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \quad ,$$

com $T^{*C} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : h(\boldsymbol{\theta}|\boldsymbol{y}) \leq h(\boldsymbol{\theta^*}|\boldsymbol{y})\}$.

Na maioria dos casos, não é possível a realização dessas etapas de forma analítica, sendo necessário a utilização de métodos numéricos, tais como o *Método de Monte Carlo* (nos casos de baixa dimensão do espaço paramétrico) e os *Métodos MCMC* (nos casos de alta dimensão do espaço paramétrico).

Capítulo 6

Inferência Bayesiana no Modelo de Regressão Ordinal

A abordagem clássica ajusta o modelo de regressão ordinal através do método de máxima verossimilhança. Um problema dessa abordagem é que as estimativas de certos coeficientes podem não existir. Em experimentos multifatoriais, por exemplo, a proximidade do número de efeitos e do número de réplicas pode levar à falha na convergência das estimativas (Chipman and Hamada, 1993). Adicionalmente, as inferências a respeito do modelo são baseadas na teoria assintótica associada, sendo sua precisão questionável em tamanhos de amostra pequenos.

A utilização de técnicas bayesiana pode ser vista como uma solução para o ajuste do modelo. Essa abordagem mantém toda estrutura do modelo cumulativo, além de promover uma descrição mais precisa da distribuição da variável resposta Y. Ao mesmo tempo, a questão da não convergência é resolvida, e as aproximações assintóticas não são necessárias para inferência (Chipman and Hamada, 1993). A modelagem nesse contexto requer a especificação de distribuições *a priori* para os parâmetros que, combinadas com a *verossimilhança*, resultam na distribuição *a posteriori*. Em posse da distribuição posterior, toda inferência bayesiana pode ser

desenvolvida (estimativas pontuais, intervalos de credibilidade e testes de hipóteses para os parâmetros do modelo).

Os parâmetros de interesse no Modelo Ordinal são:

- $\boldsymbol{\beta}^{'} = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_k] \rightarrow$ coeficientes do Modelo Ordinal referente às variáves explicativas ($\beta_z \in \Re, z = 1, \cdots, k$);

- $\boldsymbol{\alpha}^{'} = [\alpha_1, \alpha_2, \cdots, \alpha_{J-1}] \rightarrow$ pontos de corte da variável contínua subjacente Y^* ($\alpha_j \in \Re, j = 1, \cdots, J-1$, com $\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}$).

Para cada um dos $\beta_z^{'}s$, prioris com o espaço amostral pertencente aos números reais podem ser consideradas. A restrição que envolve os $\alpha_j^{'}s$ ($\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}$) deve ser levada em consideração na tarefa de se atribuir distribuições a priori para esses parâmetros. Como visto no Capítulo 4, prioris informativas e não informativas podem ser atribuídas, dependendo da quantidade de informação disponível. No caso de prioris informativas, Johnson and Albert (1998) discutem uma abordagem alternativa: um método de especificar estimativas a priori para as probabilidades cumulativas de sucesso ao invés de atribuir prioris diretamente para os parâmetros.

O cálculo da *verossimilhança* pode ser derivado de duas abordagens (que geram funções distintas) que, consequentemente, resultam em duas formas da distribuição posterior em um contexto bayesiano. Como visto no Capítulo 3, as variáveis resposta Y_i são independentes e multinomialmente distribuídas, com as probabilidades de sucesso para cada categoria $(\pi_{i1}, \pi_{i2}, \cdots, \pi_{iJ})$ definidas pela forma do modelo cumulativo. A partir disso desenvolve-se a função de verossimilhança em 3.16. Essa é a "real" verossimilhança dos dados, utilizada no contexto clássico para encontrar

os valores dos parâmetros que a maximizam (Estimadores de Máxima Verossimilhança). A primeira abordagem bayesiana (e também a mais intuitiva) seria a de combinar essa verossimilhança com distribuições a priori para os parâmetros do modelo, resultando na seguinte distribuição posterior:

$$h(\boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{x}) \propto h(\boldsymbol{\alpha}, \boldsymbol{\beta}) * \\ * \prod_{i=1}^{n} \left[\left(\frac{exp(\alpha_1 - \boldsymbol{X}_i'\boldsymbol{\beta})}{1 + exp(\alpha_1 - \boldsymbol{X}_i'\boldsymbol{\beta})} \right)^{Y_{i1}} \left(\prod_{j=2}^{J-1} \left(\frac{exp(\alpha_j - \boldsymbol{X}_i'\boldsymbol{\beta})}{1 + exp(\alpha_j - \boldsymbol{X}_i'\boldsymbol{\beta})} - \frac{exp(\alpha_{j-1} - \boldsymbol{X}_i'\boldsymbol{\beta})}{1 + exp(\alpha_{j-1} - \boldsymbol{X}_i'\boldsymbol{\beta})} \right)^{Y_{ij}} \right) \left(\frac{1}{1 + exp(\alpha_{J-1} - \boldsymbol{X}_i'\boldsymbol{\beta})} \right)^{Y_{iJ}} \right]. \tag{6.1}$$

No entanto, essa não foi a abordagem adotada nos primeiros desenvolvimentos do modelo cumulativo ordinal através de métodos bayesianos exatos, no começo dos anos 90 (como em Albert and Chib (1993) e etc), pela função em 6.1 ser, na época, complicada no sentido de se normalizar e de se amostrar diretamente dela. Utilizou-se então a ideia de *Dados Aumentados* (*Data Augmentation*).

Dados Aumentados é uma técnica que se utiliza de variáveis latentes como estratégia para converter a verossimilhança em uma forma (como a da regressão normal clássica) que o amostrador simples de Gibbs (Anexo B) possa ser utilizado (Congdon, 2005). No contexto da regressão binária e multi-categórica (nominal ou ordinal), a forma das distribuições condicionais completas são simplificadas quando se assume que uma variável contínua latente está associada à variável categórica observada (Chipman and Hamada, 1993). Para introduzir essa abordagem, será considerada primeiramente a Regressão Logística definida no Capítulo 2. Em seguida será mostrado como o *Dados Aumentados*/amostrador de Gibbs pode ser generalizado para lidar com dados multinomiais em que as categorias são ordenadas.

6.1 Aumento de Dados e Amostrador de Gibbs para dados binários

6.1.1 Aumento de Dados

Suponha que n variáveis binárias independentes $Y_1, \cdots, Y_n$ foram observadas, onde Y_i segue uma distribuição de Bernoulli com probabilidade de sucesso π_i. O modelo de regressão binário é definido como $\pi_i = F(\boldsymbol{x_i'}\boldsymbol{\beta})$, onde F é uma função de distribuição conhecida que faz a ligação entre as probabilidades π_i com o preditor linear $\boldsymbol{x_i'}\boldsymbol{\beta}$. Como discutido no Capítulo 2, o modelo logístico é obtido se F é a função de distribuição logística.

Seja $F = \Phi$ a função de distribuição de uma normal padrão, gerando o modelo *Probit* (McCullagh and Nelder, 1989). Sendo $h(\boldsymbol{\beta})$ a densidade a priori de $\boldsymbol{\beta}$, a densidade a posteriori $h(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{x})$ será dada por

$$h(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{x}) \propto h(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(\boldsymbol{x_i'}\boldsymbol{\beta})^{y_i}(1 - \Phi(\boldsymbol{x_i'}\boldsymbol{\beta}))^{1-y_i}, \tag{6.2}$$

que é altamente intratável para o uso do amostrador de Gibbs (Albert and Chib, 1993).

O método do aumento de dados introduz n variáveis latentes independentes $Y_1^*, \cdots, Y_n^*$, onde os $Y_i^{*\prime}s \sim N(\boldsymbol{x_i'}\boldsymbol{\beta}, 1)$ e define-se $Y_i = 1$ se $Y_1^* > 0$ e $Y_i = 0$ caso contrário. Observe que se os $Y_i^{*\prime}s$ são conhecidos e uma priori normal multivariada é escolhida para $\boldsymbol{\beta}$, então a distribuição posterior de $\boldsymbol{\beta}$ pode ser derivada utilizando os resultados da regressão linear normal bayesiana (Albert and Chib, 1993). No entanto, os $Y_i^{*\prime}s$ são obviamente desconhecidos mas, dado os dados Y_i, a distribuição de Y_i^* segue uma normal truncada. Essas informações, combinadas com o amostra-

dor de Gibbs, permitem a simulação a partir da distribuição posterior exata de $\boldsymbol{\beta}$ (Albert and Chib, 1993), como é descrito abaixo.

6.1.2 Amostrador de Gibbs

A densidade posterior conjunta de $\boldsymbol{\beta}$ e $\boldsymbol{Y^*} = (Y_1^*, \cdots, Y_n^*)$ é dada por (Albert and Chib, 1993)

$$h(\boldsymbol{\beta}, \boldsymbol{Y^*}|\boldsymbol{y}, \boldsymbol{x}) \propto h(\boldsymbol{\beta}) \prod_{i=1}^{n} \{I(Y_i^* > 0)I(y_i = 1) + I(Y_i^* \leq 0)I(y_i = 0)\}\Phi(Y_i^* - \boldsymbol{x_i'}\boldsymbol{\beta}), \qquad (6.3)$$

sendo $I(.)$ uma função indicadora. Essa distribuição é difícil de se normalizar e amostrar valores diretamente dela. No entanto, o cálculo da distribuição posterior marginal de $\boldsymbol{\beta}$ utilizando o amostrador de Gibbs, requer somente a distribuição posterior de $\boldsymbol{\beta}$ condicionada a $\boldsymbol{Y^*}$ e a distribuição posterior de $\boldsymbol{Y^*}$ condicionada a $\boldsymbol{\beta}$. Essas distribuições condicionais são de formas clássicas conhecidas: $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{y^*}$ segue uma normal multivariada e $Y_i^*|\boldsymbol{y}, \boldsymbol{\beta}$ segue uma normal truncada à esquerda de zero se $y_i = 1$ e uma normal truncada à direita de zero se $y_i = 0$ (Albert and Chib, 1993). Assim, a tarefa de simulação a partir dessas duas distribuições é computacionalmente fácil, o que torna possível o uso do amostrador simples de Gibbs.

6.2 Aumento de Dados e Amostrador de Gibbs para dados ordinais

6.2.1 Aumento de Dados

Como citado no Capítulo 3, para cada resposta ordinal Y_i com J categorias há uma variável latente Y_i^* em escala contínua. A correspondência se dá através dos pontos de corte $\alpha_0, \alpha_1, \cdots, \alpha_J$ onde $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_J = \infty$. Se $\alpha_{j-1} < Y^* \leq \alpha_j$, então $Y = j$ é observado. A distribuição da variável Y^* é determinada pela forma

da função de ligação F^{-1}, já que $Link(P(Y \leq j)) = \alpha_j - \boldsymbol{x}'\boldsymbol{\beta}$, ou equivalentemente $Link(P(Y^* \leq \alpha_j)) = \alpha_j - \boldsymbol{x}'\boldsymbol{\beta}$.

Adotando-se a abordagem de *Dados Aumentados*, a *verossimilhança*, como função do conjunto de parâmetros e dados desconhecidos, passa a ser reescrita em termos das variáveis latentes Y_i^* como (Johnson and Albert, 1998)

$$P(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{Y^*}) = \prod_{i=1}^{N} F(Y_i^* - \boldsymbol{x_i'}\boldsymbol{\beta}))I(\alpha_{y_i-1} \leq Y_i^* \leq \alpha_{y_i}), \tag{6.4}$$

onde I(.) é uma função indicadora.

6.2.2 Amostrador de Gibbs

A densidade posterior conjunta das variáveis latentes e dos parâmetros do modelo é obtida através da verossimilhança em 6.4 e da informação a priori acerca dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ (Albert and Chib, 1993):

$$h(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{Y^*}|\boldsymbol{y}, \boldsymbol{x}) \propto h(\boldsymbol{\beta})h(\boldsymbol{\alpha}) \prod_{i=1}^{N} F(Y_i^* - \boldsymbol{x_i'}\boldsymbol{\beta}))I(\alpha_{y_i-1} \leq Y_i^* \leq \alpha_{y_i}), \tag{6.5}$$

adotando-se prioris independentes para $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$. As distribuições condicionais completas, requeridas para o algoritmo amostrador de Gibbs, são (Chipman and Hamada, 1993):

$$h(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{y^*}, \boldsymbol{x}, \boldsymbol{y}) \propto h(\boldsymbol{\beta}) \prod_{i=1}^{N} f(Y_i^* - \boldsymbol{x_i'}\boldsymbol{\beta}))$$

$$h(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{y^*}, \boldsymbol{x}, \boldsymbol{y}) \propto h(\boldsymbol{\alpha})I(\alpha_{y_i-1} \leq Y_i^* \leq \alpha_{y_i})$$

$$h(\boldsymbol{Y^*}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{x}, \boldsymbol{y}) \propto \prod_{i=1}^{N} f(Y_i^* - \boldsymbol{x_i'}\boldsymbol{\beta}))I(\alpha_{y_i-1} \leq Y_i^* \leq \alpha_{y_i}),$$

sendo necessária a especificação das distribuições a priori independentes de $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ e da função de ligação F^{-1}. Johnson and Albert (1998) estimula a utilização do

modelo probito (o que leva Y^* a ser normalmente distribuído), já que com esse link (e certas prioris de $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$) todas distribuições condicionais completas acima tem formas analíticas tratáveis, sendo possível a utilização do amostrador simples de Gibbs.

Albert and Chib (1993) adotam prioris uniformes para $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ (respeitando a ordenação dos $\alpha' s$), gerando uma densidade posterior proporcional à verossimilhança em 6.4. A distribuição dos $Y_i^{*\prime} s$ dado $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ são normais truncadas independentes, onde os pontos de truncamento são definidos pelos valores correntes dos pontos de corte da categoria. A distribuição posterior de $\boldsymbol{\beta}$ condicionada a $\boldsymbol{\alpha}$ e Y^* segue uma normal multivariada (a mesma que a densidade posterior dos parâmetros da regressão para o modelo normal clássico quando a variância é conhecida e igual a um). Finalmente, a distribuição condicional dos componentes de $\boldsymbol{\alpha}$, por exemplo α_j, dados os valores correntes de $\boldsymbol{\beta}$ e Y^* é uniformemente distribuída no intervalo $(max_{y_i=j-1} \mathrm{Y}_i^*, min_{y_i=j} \mathrm{Y}_i^*)$.

Chipman and Hamada (1993) atribuem uma priori normal multivariada para $\boldsymbol{\beta}$ e uma priori normal multivariada truncada para $\boldsymbol{\alpha}$, devido a restrição $\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}$. Essas escolhas resultam nas seguintes distribuições condicionais completas: $h(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{y^*}, \boldsymbol{x}, \boldsymbol{y})$ segue uma normal multivariada, $h(\alpha_j|\boldsymbol{\beta}, \boldsymbol{y^*}, \boldsymbol{x}, \boldsymbol{y})$ segue uma normal truncada pelos $Y_i^{*\prime} s$ e $h(\boldsymbol{Y_i^*}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{x}, \boldsymbol{y})$ segue uma normal truncada pelos $\alpha_j' s$.

6.3 Dados Aumentados vs Verossimilhança Multinomial

Apesar de ter tido papel fundamental no desenvolvimento da abordagem bayesiana para os modelos de regressão ordinal nos anos 90, o método de *Dados Aumentados* apresenta algumas desvantagens e é criticado em artigos mais recentes, como em Ishwaran (2000) e Parmigiani et al. (2003).

O grande diferencial para a época (e o que contribuiu para a popularidade e difusão) do procedimento de Albert and Chib (1993) era contar com um amostrador de Gibbs que utiliza atualizações exatas para o parâmetro de locação no modelo probito. No entanto, a mudança da função de ligação resulta em mudanças expressivas no amostrador de Gibbs, tais como a perda dessa propriedade. Isso se deve ao fato das distribuições condicionais completas dos parâmetros da regressão $\boldsymbol{\beta}$ não serem de formas conhecidas (Johnson and Albert, 1998). Com isso, os exemplos de aplicação da década de 90 utilizavam em sua maioria a função *Probit*, como em Albert and Chib (1993), Chipman and Hamada (1993) e Cowles (1996).

Outro problema da abordagem de *Dados Aumentados* proposta em Albert and Chib (1993) é a ineficiência da amostragem dos pontos de corte $\alpha_j's$. Lembrando que a variável latente Y^* é uniformemente distribuída no intervalo $(max_{y_i=j-1}Y_i^*, min_{y_i=j}Y_i^*)$. No entanto, quando há um número alto de observações em categorias adjacentes, esse intervalo tende a ser pequeno e a movimentação do componente $\boldsymbol{\alpha}$ é mínimo. Isso resulta em um lento avanço no amostrador de Gibbs (Congdon, 2005), com autocorrelações altas e que com frequência não desaparecem

(Ishwaran, 2000). Surgiram então métodos que buscavam acelerar a convergência dos pontos de corte. Por exemplo, Cowles (1996) discutiu a utilização de um passo Metropolis-Hastings multivariado que atualizasse os pontos de corte e as variáveis latentes simultaneamente. Esse procedimento também é discutido em Johnson and Albert (1998). Albert and Chib (1998) utilizaram o Metropolis-Hastings com *Dados Aumentados* através de uma reparametrização dos pontos de corte, que os tornam sem ordenação:

$$\gamma_2 = log(\alpha_2) \quad ; \quad \gamma_j = log(\alpha_j - \alpha_{j-1}), \quad 2 \leq j \leq J-1,$$

ou seja,

$$\alpha_2 = exp(\gamma_2) \quad ; \quad \alpha_j = \alpha_{j-1} + \sum_{v=3}^{j} exp(\gamma_v),$$

assumindo $\alpha_1 = 0$, pois é adotado que $\boldsymbol{x_i'\beta}$ inclui um intercepto e, para resolver o problema de identificabilidade, alguma restrição nos $\alpha's$ deve ser incorporada (Johnson and Albert, 1998). Mesmo com essa evolução dos métodos visando algoritmos mais velozes, de uma forma geral a adição e simulação de tantas variáveis latentes quanto o número de observações diminui a velocidade de convergência da cadeia de Markov, especialmente para o caso de amostras grandes (Ishwaran, 2000).

Uma vantagem de se utilizar o procedimento de *Dados Aumentados* na modelagem ordinal é a análise de resíduos. Enquanto na abordagem da verossimilhança multinomial existem $J-1$ resíduos para cada observação, há somente um resíduo na abordagem de variáveis latentes: $y_i^* - \boldsymbol{\beta x_i}$ (Johnson and Albert, 1998). O crescimento de dimensão, de 1 para $J-1$, complica a análise de resíduos. Não somente há mais resíduos a serem examinados, mas os $J-1$ resíduos para cada observação

são correlacionados. Consequentemente, não há uma forma clara como os resíduos clássicos (Pearson, deviance e etc) podem ser analisados. Já os resíduos da abordagem de aumento de dados são independentes e distribuídos conforme a função F. Desvios da estrutura do modelo são refletidos por valores atípicos dessas quantidades em relação a amostras tiradas de F (Johnson and Albert, 1998). Contudo, há alternativas para a abordagem multinomial como a discutida em Congdon (2005), onde novos y_i são amostrados e a concordância entre as categorias preditas e as atuais categorias é avaliada (esse procedimento pode também ser utilizado na abordagem de *Dados Aumentados*).

Como visto, o grande motivo para a utilização de *Dados Aumentados* eram as restrições da época. A combinação desse método com certas condições (função de ligação *Probit*, por exemplo) permitiram o ajuste dos modelos binários, nominais e ordinais a partir de técnicas bayesianas. Atualmente, o seu uso não é mais uma necessidade visto que a evolução dos algoritmos $MCMC$ e da capacidade computacional permitem ao pesquisador a utilização de qualquer uma das duas abordagens da verossimilhança, com qualquer função de ligação desejada e sem prejuízos significativos em termos de tempo e/ou qualidade das estimativas.

6.4 Abordagens adotadas no trabalho

A abordagem da verossimilhança multinomial será adotada nesse trabalho. A exemplo de Albert and Chib (1998), uma reparametrização nos pontos de corte é sugerida a fim de os tornar sem ordenação e facilitar a atribuição de distribuições a priori:

$$\begin{cases} \alpha_1 = \alpha_1 \\ \alpha_2 = \alpha_1 + \lambda_2 \\ \alpha_3 = \alpha_2 + \lambda_3 = \alpha_1 + \lambda_2 + \lambda_3 \\ \vdots \\ \alpha_{J-1} = \alpha_{J-2} + \lambda_{J-1} = \alpha_1 + \sum_{v=2}^{J-1} \lambda_v \end{cases}$$

Ao contrário de Albert and Chib (1998), essa reparametrização não necessita de fixar $\alpha_1 = 0$, visto que adota-se que $\boldsymbol{\beta}$ não contém intercepto (como proposto em McCullagh and Nelder (1989), Agresti (2012) e etc). Os parâmetros de interesse passam a ser: $\boldsymbol{\beta}' = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_k]$, α_1 e $\lambda_2, \cdots, \lambda_{J-1}$. A verossimilhança, anteriormente definida em 3.16, passa a ter a seguinte forma, com $\lambda_1 = 0$,

$$\prod_{i=1}^{n} \left[\left(\frac{exp(\alpha_1 - \boldsymbol{X}_i'\boldsymbol{\beta})}{1+exp(\alpha_1 - \boldsymbol{X}_i'\boldsymbol{\beta})} \right)^{Y_{i1}} \left(\prod_{j=2}^{J-1} \left(\frac{exp[(\alpha_1 + \sum_{v=1}^{j} \lambda_v) - \boldsymbol{X}_i'\boldsymbol{\beta}]}{1+exp[(\alpha_1 + \sum_{v=1}^{j} \lambda_v) - \boldsymbol{X}_i'\boldsymbol{\beta}]} - \frac{exp[(\alpha_1 + \sum_{v=1}^{j-1} \lambda_v) - \boldsymbol{X}_i'\boldsymbol{\beta}]}{1+exp[(\alpha_1 + \sum_{v=1}^{j-1} \lambda_v) - \boldsymbol{X}_i'\boldsymbol{\beta}]} \right)^{Y_{ij}} \right) *$$

$$* \left(\frac{1}{1+exp[(\alpha_1 + \sum_{v=2}^{J-1} \lambda_v) - \boldsymbol{X}_i'\boldsymbol{\beta}]} \right)^{Y_{iJ}} . \tag{6.6}$$

As prioris consideradas para os parâmetros do modelo ordinal em 6.6 serão independentes uma das outras e terão as seguintes distribuições:

- $\alpha_1 \sim N(\mu_0, \sigma_0^2)$;
- $\lambda_v \sim Gama(\delta_v, \phi_v), \quad v = 2, \cdots, J-1$;
- $\beta_z \sim N(\mu_z, \sigma_z^2), \quad z = 1, \cdots, k$.

Com as prioris definidas para os parâmetros do Modelo Ordinal e a *verossimilhança* em 6.6, a *posteriori* pode ser calculada através da Fórmula de Bayes (4.1):

$$h(\alpha_1, \lambda_2, \cdots, \lambda_{J-1}, \beta_1, \cdots, \beta_k | \boldsymbol{y}, \boldsymbol{x}) \propto P(\boldsymbol{y}, \boldsymbol{x} | \alpha_1, \lambda_2, \cdots, \lambda_{J-1}, \beta_1, \cdots, \beta_k) h(\alpha_1, \lambda_2, \cdots, \lambda_{J-1}, \beta_1, \cdots, \beta_k)$$

$h(\alpha_1, \lambda_2, \cdots, \lambda_{J-1}, \beta_1, \cdots, \beta_k | \boldsymbol{y}, \boldsymbol{x}) \propto$

$$\propto \prod_{i=1}^{n} \left\{ \left[\left(\frac{1}{\sigma_0\sqrt{2\pi}} exp\left(\frac{-(\alpha_1-\mu_0)^2}{2\sigma_0^2} \right) \right) \left(\frac{exp(\alpha_1 - \boldsymbol{X}'_i\boldsymbol{\beta})}{1+exp(\alpha_1-\boldsymbol{X}'_i\boldsymbol{\beta})} \right)^{Y_{i1}} \right] \left[\prod_{j=2}^{J-1} \left(\frac{\phi_j^{\delta_j}}{\Gamma(\delta_j)} exp(-\phi_j\lambda_j)\lambda_j^{\delta_j - 1} \right) * \right.$$

$$* \left(\frac{exp[(\alpha_1+\sum_{v=1}^{j}\lambda_v) - \boldsymbol{X}'_i\boldsymbol{\beta}]}{1+exp[(\alpha_1+\sum_{v=1}^{j}\lambda_v) - \boldsymbol{X}'_i\boldsymbol{\beta}]} - \frac{exp[(\alpha_1+\sum_{v=1}^{j-1}\lambda_v) - \boldsymbol{X}'_i\boldsymbol{\beta}]}{1+exp[(\alpha_1+\sum_{v=1}^{j-1}\lambda_v) - \boldsymbol{X}'_i\boldsymbol{\beta}]} \right)^{Y_{ij}} \left[\left(\frac{1}{1+exp[(\alpha_1+\sum_{v=2}^{J-1}\lambda_v) - \boldsymbol{X}'_i\boldsymbol{\beta}]} \right)^{Y_{iJ}} \right] *$$

$$\left. * \prod_{z=1}^{k} \left[\frac{1}{\sigma_z\sqrt{2\pi}} exp\left(\frac{-(\beta_z-\mu_z)^2}{2\sigma_z^2} \right) \right] \right. . \tag{6.7}$$

Note que a *posteriori* (6.7) não apresenta forma analítica fechada, no entanto, ela pode ser estimada empiricamente através dos métodos $MCMC$ como o amostrador de Gibbs (Anexo B) e o algoritmo Metropolis-Hastings (Anexo C).

Capítulo 7

Simulações

Este capítulo tem por objetivo apresentar os resultas de simulações computacionais de dados ordinais e sua consequente modelagem em diversos cenários, referentes ao número de categorias da variável resposta Y e do número e tipo das variáveis explicativas X_k. Serão obtidos:

- Estimativas pontuais para cada um dos $J-1+k$ parâmetros do Modelo Ordinal, contidos na posteriori em 6.7. A média a posteriori será o estimador de Bayes utilizado.

- Intervalos de credibilidade do tipo HPD de 95% para os $J-1+k$ parâmetros do Modelo Ordinal.

- Realização do $FBST$ para testar se cada um dos $\beta_k's$ é igual a zero, bem como as $\binom{k}{2}+\binom{k}{3}+\cdots+\binom{k}{k}$ combinações possíveis de $\beta's$ serem iguais a zero.

  Testes de hipóteses para os $J-1$ $\alpha_j's$ do Modelo Ordinal não serão realizados por não apresentarem sentido prático. Usualmente eles não são de interesse, exceto para a estimação das probabilidades da variável resposta (Agresti, 2012).

As distribuições a priori consideradas para os parâmetros serão não informativas, obtidas adotando-se uma dada média e variância grande (prioris difusas):

- $\alpha_1 \sim N(0, 1000)$;
- $\lambda_v \sim Gama(1/1000, 1/1000), \quad v = 2, \cdots, J-1$;
- $\beta_z \sim N(0, 1000), \quad z = 1, \cdots, k$.

A utilização de prioris não informativas permitirão a comparação com os resultados da abordagem clássica, e, consequentemente, a comparação do desempenho do $FBST$ em relação ao p-valor. Todas as simulações foram realizadas através dos algoritmos descritos no Anexo A e utilizando o Software R (R Core Team, 2013).

7.1 Resultados

7.1.1 Simulação com $J = 3$ categorias e 1 variável explicativa binária

A Tabela 7.1 apresenta os resultados obtidos em simulações de um modelo logístico ordinal considerando $J = 3$ categorias e uma única variável explicativa binária. As simulações foram realizadas para diferentes tamanhos de amostras (n=30, 50, 100 e 200) e considerando os seguintes valores dos parâmetros: $\alpha_1 = -0,5$, $\alpha_2 = 0,5$ e $\beta = 0,15$.

Tabela 7.1: Inferência Clássica e Bayesiana dos Parâmetros do Modelo Logístico Ordinal ($\alpha_1 = -0,5$, $\alpha_2 = 0,5$ e $\beta = 0,15$).

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
$\boldsymbol{n = 30}$						
$\alpha_1 = -0,5$	-0,339	(-1,309 : 0,630)	-	-0,332	(-1,283 : 0,707)	-
$\alpha_2 = 0,5$	0,914	(-0,105 : 1,934)	-	0,930	(-0,143 : 1,950)	-
$\beta = 0,15$	0,136	(-1,184: 1,457)	0,840	0,141	(-1,205 : 1,484)	0,998
$\boldsymbol{n = 50}$						
$\alpha_1 = -0,5$	-0,693	(-1,448 : 0,061)	-	-0,699	(-1,437 : 0,086)	-
$\alpha_2 = 0,5$	0,375	(-0,360 : 1,111)	-	0,370	(-0,382 : 1,121)	-
$\beta = 0,15$	-0,252	(-1,280 : 0,775)	0,630	-0,257	(-1,296 : 0,771)	0,975
$\boldsymbol{n = 100}$						
$\alpha_1 = -0,5$	-0,635	(-1,192 : -0,079)	-	-0,636	(-1,202 : -0,079)	-
$\alpha_2 = 0,5$	0,290	(-0,253 : 0,834)	-	0,290	(-0,263 : 0,838)	-
$\beta = 0,15$	0,331	(-0,404 : 1,068)	0,390	0,336	(-0,406 : 1,074)	0,859
$\boldsymbol{n = 200}$						
$\alpha_1 = -0,5$	-0,862	(-1,270 : -0,453)	-	-0,864	(-1,286 : -0,459)	-
$\alpha_2 = 0,5$	0,173	(-0,216 : 0,564)	-	0,172	(-0,227 : 0,568)	-
$\beta = 0,15$	-0,283	(-0,799 : 0,232)	0,281	-0,285	(-0,787 : 0,245)	0,760

A Tabela 7.2 apresenta os resultados obtidos em simulações de um modelo logístico ordinal considerando $J = 3$ categorias e uma única variável explicativa binária. As simulações foram realizadas para diferentes tamanhos de amostras (n=30, 50, 100 e 200) e considerando os seguintes valores dos parâmetros: $\alpha_1 = -0,5$, $\alpha_2 = 0,5$ e $\beta = 1,0$.

Tabela 7.2: Inferência Clássica e Bayesiana dos Parâmetros do Modelo Logístico Ordinal ($\alpha_1 = -0,5$, $\alpha_2 = 0,5$ e $\beta = 1,0$).

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
$\boldsymbol{n = 30}$						
$\alpha_1 = -0,5$	0,136	(-0,699 : 0,971)	-	0,157	(-0,681 : 1,011)	-
$\alpha_2 = 0,5$	1,338	(0,363 : 2,313)	-	1,368	(0,394 : 2,397)	-
$\beta = 1,0$	0,480	(-1,002 : 1,963)	0,526	0,488	(-1,058 : 2,034)	0,947
$\boldsymbol{n = 50}$						
$\alpha_1 = -0,5$	-0,574	(-1,354 : 0,206)	-	-0,577	(-1,389 : 0,204)	-
$\alpha_2 = 0,5$	0,240	(-0,524 : 1,005)	-	0,235	(-0,583 : 0,992)	-
$\beta = 1,0$	0,814	(-0,272 : 1,901)	0,142	0,847	(-0,291 : 1,927)	0,558
$\boldsymbol{n = 100}$						
$\alpha_1 = -0,5$	-0,462	(-1,005 : 0,081)	-	-0,462	(-1,001 : 0,094)	-
$\alpha_2 = 0,5$	0,379	(-0,161 : 0,920)	-	0,379	(-0,161 : 0,930)	-
$\beta = 1,0$	0,852	(0,090 : 1,613)	0,028	0,866	(0,107 : 1,634)	0,189
$\boldsymbol{n = 200}$						
$\alpha_1 = -0,5$	-0,537	(-0,933 : -0,141)	-	-0,538	(-0,940 : -0,145)	-
$\alpha_2 = 0,5$	0,449	(0,056 : 0,842)	-	0,449	(0,055 : 0,845)	-
$\beta = 1,0$	1,207	(0,651 : 1,762)	0,00002	1,215	(0,660 : 1,780)	0,00015

A Tabela 7.3 apresenta os resultados obtidos em simulações de um modelo logístico ordinal considerando $J = 3$ categorias e uma única variável explicativa binária. As simulações foram realizadas para diferentes tamanhos de amostras (n=30, 50, 100 e 200) e considerando os seguintes valores dos parâmetros: $\alpha_1 = -0,5$, $\alpha_2 = 0,5$ e $\beta = 2,0$.

Tabela 7.3: Inferência Clássica e Bayesiana dos Parâmetros do Modelo Logístico Ordinal ($\alpha_1 = -0,5$, $\alpha_2 = 0,5$ e $\beta = 2,0$).

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
$\boldsymbol{n = 30}$						
$\alpha_1 = -0,5$	-1,352	(-2,568 : -0,136)	-	-1,401	(-2,689 : -0,176)	-
$\alpha_2 = 0,5$	-0,268	(-1,317 : 0,781)	-	-0,308	(-1,420 : 0,765)	-
$\beta = 2,0$	1,210	(0,410, : 2,831)	0,143	1,322	(-0,345 : 3,107)	0,565
$\boldsymbol{n = 50}$						
$\alpha_1 = -0,5$	-0,776	(-1,549 : -0,003)	-	-0,785	(-1,588 : 0,024)	-
$\alpha_2 = 0,5$	0,163	(-0,569 : 0,896)	-	0,153	(-0,585 : 0,911)	-
$\beta = 2,0$	1,402	(0,183 : 2,621)	0,024	1,480	(0,225 : 2,751)	0,16
$\boldsymbol{n = 100}$						
$\alpha_1 = -0,5$	-1,026	(-1,623 : -0,429)	-	-1,036	(-1,632 : -0,432)	-
$\alpha_2 = 0,5$	-0,218	(-0,761 : 0,325)	-	-0,229	(-0,783 : 0,324)	-
$\beta = 2,0$	1,558	(0,596 : 2,520)	0,001	1,607	(0,656 : 2,625)	0,011
$\boldsymbol{n = 200}$						
$\alpha_1 = -0,5$	-0,660	(-1,043 : -0,276)	-	-0,661	(-1,048 : -0,276)	-
$\alpha_2 = 0,5$	0,309	(-0,061 : 0,680)	-	0,310	(-0,065 : 0,685)	-
$\beta = 2,0$	1,573	(0,962 : 2,184)	$\cong 0$	1,593	(0,983 : 2,222)	0

7.1.2 Simulação com $J = 3$ categorias e 1 variável explicativa quantitativa

A Tabela 7.4 apresenta os resultados obtidos em simulações de um modelo logístico ordinal considerando $J = 3$ categorias e uma única variável explicativa quantitativa. As simulações foram realizadas para diferentes tamanhos de amostras (n=30, 50, 100 e 200) e considerando os seguintes valores dos parâmetros: $\alpha_1 = -1,5$, $\alpha_2 = 0,0$ e $\beta = 0,1$.

Tabela 7.4: Inferência Clássica e Bayesiana dos Parâmetros do Modelo Logístico Ordinal ($\alpha_1 = -1,5$, $\alpha_2 = 0,0$ e $\beta = 0,10$).

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
$\boldsymbol{n = 30}$						
$\alpha_1 = -1,5$	-2,766	(-4,322 : -1,209)	-	-2,890	(-4,545 : -1,283)	-
$\alpha_2 = 0,0$	0,287	(-0,623 : 1,199)	-	0,285	(-0,643 : 1,285)	-
$\beta = 0,10$	-0,028	(-0,161 : 0,104)	0,675	-0,030	(-0,169 : 0,112)	0,983
$\boldsymbol{n = 50}$						
$\alpha_1 = -1,5$	-1,183	(-1,864 : -0,501)	-	-1,199	(-1,888 : -0,501)	-
$\alpha_2 = 0,0$	-0,120	(-0,703 : 0,463)	-	-0,136	(-0,745 : 0,455)	-
$\beta = 0,10$	0,102	(-0,029 : 0,233)	0,127	0,110	(-0,022 : 0,250)	0,495
$\boldsymbol{n = 100}$						
$\alpha_1 = -1,5$	-1,06	(-1,666 : -0,454)	-	-1,065	(-1,694 : -0,465)	-
$\alpha_2 = 0,0$	0,161	(-0,397 : 0,720)	-	0,161	(-0,397 : 0,729)	-
$\beta = 0,10$	0,115	(0,026 : 0,203)	0,011	0,117	(0,029 : 0,208)	0,085
$\boldsymbol{n = 200}$						
$\alpha_1 = -1,5$	-1,193	(-1,615 : -0,772)	-	-1,196	(-1,617 : -0,777)	-
$\alpha_2 = 0,0$	0,359	(-0,017 : 0,735)	-	0,361	(-0,007 : 0,748)	-
$\beta = 0,10$	0,146	(0,083 : 0,208)	0,000004	0,148	(0,086 : 0,212)	0,00013

7.1.3 Simulação com $J = 4$ categorias e 2 variáveis explicativas (binária e contínua)

A Tabela 7.5 apresenta os resultados obtidos em simulações de um modelo logístico ordinal considerando $J = 4$ categorias e duas variáveis explicativas, X_1 e X_2, onde X_1 é uma variável binária e X_2 é uma variável contínua. As simulações foram realizadas para diferentes tamanhos de amostras (n=50 e 200) e considerando os seguintes valores dos parâmetros: $\alpha_1 = -0,3$, $\alpha_2 = 1,2$, $\alpha_3 = 2,2$, $\beta_1 = -0,1$ e $\beta_2 = 1,1$.

Tabela 7.5: Inferência Clássica e Bayesiana dos Parâmetros do Modelo Logístico Ordinal ($\alpha_1 = -0,3$, $\alpha_2 = 1,2$, $\alpha_3 = 2,2$, $\beta_1 = -0,1$ e $\beta_2 = 1,1$).

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
$\boldsymbol{n = 50}$						
α_1 = -0,3	0,995	(-0,813 : 2,803)	-	1,256	(-0,687 : 3,239)	-
$\alpha_2 = 1,2$	2,199	(0,216 : 4,182)	-	2,471	(0,461 : 4,665)	-
$\alpha_3 = 2,2$	3,012	(0,945 : 5,078)	-	3,279	(1,166 : 5,565)	-
$\beta_1 = -0,1$	0,587	(-1,249 : 2,425)	0,531	0,649	(-1,324 : 2,612)	0,996
$\beta_2 = 1,10$	0,956	(0,499 : 1,413)	0,00004	1,067	(0,608 : 1,577)	0
$\beta_1 = \beta_2 = 0$	-	-	$\cong 0$	-	-	0
$\boldsymbol{n = 200}$						
α_1 = -0,3	-1,301	(-2,136 : -0,465)	-	-1,304	(-2,162 : -0,463)	-
$\alpha_2 = 1,2$	0,316	(-0,511 : 1,144)	-	0,328	(-0,529 : 1,155)	-
$\alpha_3 = 2,2$	1,483	(0,559 : 2,407)	-	1,494	(0,592 : 2,460)	-
$\beta_1 = -0,1$	-1,121	(-2,214 : -0,029)	0,044	-1,163	(-2,290 : -0,048)	0,562
$\beta_2 = 1,10$	1,016	(0,741 : 1,290)	$\cong 0$	1,053	(0,770 : 1,332)	0
$\beta_1 = \beta_2 = 0$	-	-	$\cong 0$	-	-	0

A Tabela 7.6 apresenta os resultados obtidos em simulações de um modelo logístico ordinal considerando $J = 4$ categorias e duas variáveis explicativas, X_1 e X_2, onde X_1 é uma variável binária e X_2 é uma variável contínua. As simulações foram realizadas para diferentes tamanhos de amostras (n=50 e 200) e considerando os seguintes valores dos parâmetros: $\alpha_1 = -0,3$, $\alpha_2 = 1,2$, $\alpha_3 = 2,2$, $\beta_1 = -1,5$ e $\beta_2 = 0,02$.

Tabela 7.6: Inferência Clássica e Bayesiana dos Parâmetros do Modelo Logístico Ordinal ($\alpha_1 = -0,3$, $\alpha_2 = 1,2$, $\alpha_3 = 2,2$, $\beta_1 = -1,5$ e $\beta_2 = 0,02$).

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
$\boldsymbol{n = 50}$						
$\alpha_1 =$ - 0,3	-0,652	(-1,554 : 0,249)	-	-0,658	(-1,576 : 0,285)	-
$\alpha_2 = 1,2$	1,240	(0,233 : 2,246)	-	1,270	(0,285 : 2,346)	-
$\alpha_3 = 2,2$	2,059	(0,793 : 3,325)	-	2,095	(0,848 : 3,419)	-
$\beta_1 = -1,5$	-1,759	(-2,971 : -0,547)	0,004	-1,836	(-3,098 : -0,603)	0,150
$\beta_2 = 0,02$	-0,028	(-0,113 : 0,057)	0,520	-0,032	(-0,124 : 0,053)	0,995
$\beta_1 = \beta_2 = 0$	-	-	0,010	-	-	0,137
$\boldsymbol{n = 200}$						
$\alpha_1 =$ - 0,3	-0,200	(-0,650 : 0,249)	-	-0,192	(-0,640 : 0,266)	-
$\alpha_2 = 1,2$	1,596	(1,076 : 2,116)	-	1,614	(1,099 : 2,133)	-
$\alpha_3 = 2,2$	2,576	(1,903 : 3,248)	-	2,596	(1,941 : 3,293)	-
$\beta_1 = -1.5$	-1,370	(-1,949 : -0,791)	0,000003	-1,385	(-1,954 : -0,803)	0,0002
$\beta_2 = 0,02$	0,047	(0,011 : 0,083)	0,010	0,048	(0,012 : 0,084)	0,240
$\beta_1 = \beta_2 = 0$	-	-	$\cong 0$	-	-	0

7.2 Comentários

As estimativas pontuais bayesianas, calculadas como a média da distribuição posterior, foram bem próximas das estimativas clássicas de verossimilhança. Essa proximidade se tornou maior a medida que o tamanho da amostra crescia. Logicamente tal fato também ocorreu em relação aos intervalos de confiança e os intervalos de credibilidade do tipo HPD. Os intervalos de confiança clássicos calculados são do tipo $Wald$. Eles são simétricos em torno da estimativa pontual e se baseiam na normalidade assintótica dos estimadores de máxima verossimilhança. Sua precisão é, portanto, questionável para amostras pequenas. Os intervalos HPD não são necessariamente simétricos, e nas simulações realizadas foram sempre mais conservadores que os intervalos de confiança.

A relação entre os p-valores e os intervalos de confiança no sentido de que se a hipótese nula de algum $\beta_k = 0$ for rejeitada a um nível de significância de $\alpha \times 100\%$

(p-valor $< \alpha$) o zero não estará contido no intervalo de confiança de $(1-\alpha) \times 100\%$, não se aplica entre os intervalos HPD e os e-valores. Por exemplo, na Tabela 7.2 com n = 100, o e-valor de 0,189 não rejeita a hipótese que $\beta = 0$, mas o intervalo HPD de 95% não contém o zero. Isso decorre do fato de que, enquanto o intervalo HPD é calculado no espaço paramétrico unidimensional (correspondente à dimensão do parâmetro em questão), o e-valor é calculado em todo espaço paramétrico Θ (multidimensional).

Os e-valores foram mais conservadores que os p-valores clássicos, havendo casos em que a hipótese nula seria rejeitada a um nível de 5% na abordagem clássica mas não seria rejeitada na abordagem bayesiana (Tabela 7.3 com n= 50, por exemplo). Esse cenário já esteve presente nos exemplos de Pereira and Stern (1999). Os e-valores são relacionados com a dimensão do modelo, de forma que eles crescem a medida que a dimensão cresce (Izbickil et al., 2012).

Para os casos em que se testava simultaneamente mais de um parâmetro β, observou-se uma incoerência dos p-valores em hipóteses aninhadas a exemplo do destacado em Izbickil et al. (2012). Na Tabela 7.6, o p-valor para o teste $\beta_1 = \beta_2 = 0$ (0,01) foi maior do que o p-valor para o teste $\beta_1 = 0$ (0,004). Isso significa que, se fosse adotado um nível de significância de 0,5% , rejeitaria-se a segunda hipótese mas não rejeitaria-se a primeira. Schervish (1996) questiona a utilização dos p-valores como medida de evidência nessas situações. Por outro lado, com as prioris não informativas adotadas, os e-valores dos testes $\beta_1 = \beta_2 = 0$ e $\beta_1 = 0$ foram respectivamente 0,137 e 0,150. No $FBST$, a hipótese de dimensão maior sempre

terá um e-valor menor do que o da hipótese aninhada de dimensão menor, fazendo com que a contradição dos p-valores não se aplique. Isso se deve ao fato do cálculo do e-valor ser sempre realizado no pleno espaço de dimensão igual ao número de parâmetros do modelo. Logo, as sub-hipóteses sempre seguem coerentemente a orientação da hipótese principal (Izbickil et al., 2012).

Capítulo 8

Aplicação a dados reais

Este capítulo apresenta uma aplicação da metodologia do Capítulo 6, ilustrando um problema genético como em Grazeffe et al. (2008) onde foi verificada a influência de certo tipo de radiação na ocorrência de danos celulares. A avaliação foi realizada através do teste cometa ou eletroforese em gel de células individuais para a detecção de quebras no DNA. Os indivíduos do estudo foram caramujos da espécie *Biomphalaria glabrata* selvagens, mantidos em laboratório durante várias gerações. Os caramujos foram expostos a taxas de radiação a 2,82 Kgy/h e separados em cinco grupos: grupos expostos a 2,5; 5; 10 e 20 Gy e grupo controle (submetidos às mesmas condições exceto quanto à radiação). Ao todo foram 48 animais e em cada um deles, foram coletadas células até completar 100 células não apoptóticas (células vivas). A análise visual do teste do cometa foi baseada no trabalho de Jaloszynski et al. (1997) e consistiu na classificação dos cometas em categorias (0 a 3) de acordo com o dano no DNA (extensão da migração do DNA). A classificação do dano é apresentada abaixo e ilustrada pela Figura 8.1. A Tabela 8.1 apresenta os dados dos danos celulares induzidos pelos diferentes níveis de radiação.

- **Classe zero células sem dano**: cometas com cabeça grande e sem cauda,

sem migração do DNA;

- **Classe um células pouco danificadas**: cometas com cauda bem curta, com pouca migração do DNA;
- **Classe dois células danificadas**: cometas com caudas longas, com migração intermediária de DNA;
- **Classe três células muito danificadas**: cometas com cabeças bem pequenas e uma cauda muito longa, com muita migração de DNA.

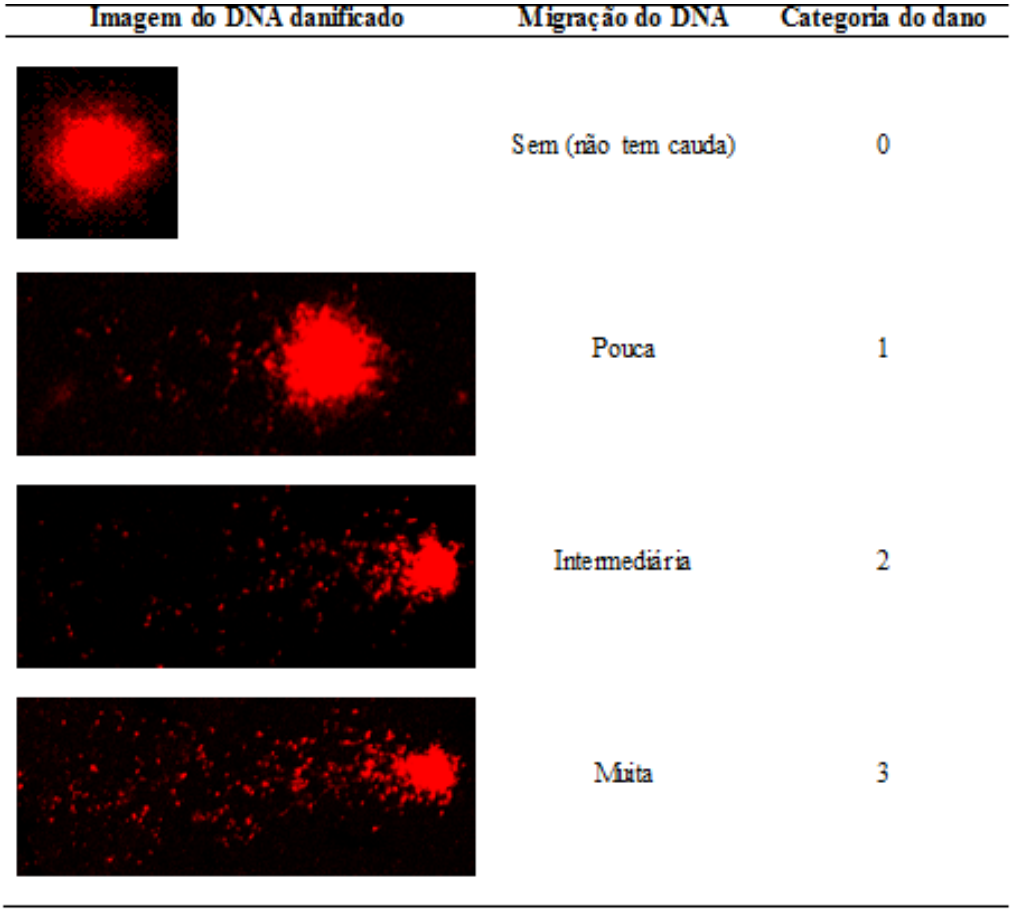

Figura 8.1: Método visual do teste do cometa para classificação da categoria do dano.

Tabela 8.1: Freqüência observada da categoria do dano, segundo a dose de radiação.

Dose (Gy)	**Categoria de dano no DNA**				**Total de células**
	0	1	2	3	
0	654	125	72	249	1100
2,5	442	178	105	175	900
5	197	253	173	277	900
10	159	296	264	281	1000
20	58	49	133	660	900

8.1 Análise Estatística

O objetivo da análise estatística é o de verificar se o nível de radiação influencia no dano celular. A variável resposta ordinal Y é dada pela categoria do dano celular (0,1,2 ou 3), enquanto a variável explicativa X corresponde ao nível de radiação (0, 2,5, 5, 10 ou 20 Gy). A escolha do modelo de regressão ordinal para a análise é uma consequência do próprio delineamento experimental. Para cada dose de radiação, o número de cometas foi fixado em 100 por animal. Logo, o vetor de frequências das categorias de cada cometa é um vetor multinomial com 4 categorias. Como as doses são variáveis ordinais, o modelo logístico multi-categórico ordinal é adequado. Em Grazeffe et al. (2008) esse problema foi solucionado utilizando-se uma partição do modelo logístico nominal, ou seja, ignorou-se a ordenação da variável resposta.

A Tabela 8.2 mostra as estimativas pontuais, intervalares e os testes de significância dos parâmetros do modelo, tanto para a abordagem clássica quanto para a abordagem Bayesiana. Na abordagem Bayesiana adotaram-se as mesmas prioris não informativas para os parâmetros que as utilizadas para as simulações do Capítulo 6. A exemplo de Grazeffe et al. (2008), a variável explicativa "*nível da radiação*" foi ajustada por um polinômio de segundo grau.

Tabela 8.2: Inferências clássicas e Bayesianas para o modelo.

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
α_1	0,105	(0,002 : 0,207)	-	0,105	(0,005 : 0,209)	-
α_2	1,051	(0,942 : 1,161)	-	1,052	(0,947 : 1,164)	-
α_3	1,827	(1,711 : 1,944)	-	1,828	(1,716 : 1,948)	-
β_1	0,147	(0,119 : 0,175)	$\cong 0$	0,147	(0,120 : 0,175)	0
β_2	-0,00036	(-0,002 : 0,001)	0,600	-0,00034	(-0,002 : 0,001)	0,997
$\beta_1 = \beta_2 = 0$	-	-	$\cong 0$	-	-	0

O e-valor = 0 para o teste $\beta_1 = \beta_2 = 0$ indica a presença de regressão. Já o coeficiente β_2 não foi significativo (e-valor = 0,997), podendo o termo quadrático do "nível de radiação" ser retirado da modelo. A Tabela 8.3 mostra as inferências para o modelo reduzido. As proporções observadas e estimadas de cada categoria do dano segundo os diferentes níveis de radiação, e para os dois modelos considerados (completo e reduzido), estão contidos na Tabela 8.4. A grande proximidade das proporções estimadas para os dois modelos confirmam que o termo quadrático do "nível de radiação" pode ser retirado (coerente com o e-valor próximo de 1), o que simplifica o modelo sem grandes perdas nas estimativas.

Tabela 8.3: Inferências clássicas e Bayesianas para o modelo reduzido.

Parametros	**Classico**			**Bayesiano**		
	EMV	IC (95%)	p-valor	Média Post	HPD (95%)	e-valor
α_1	0,088	(0,007 : 0,170)	-	0,088	(0,009 : 1,171)	-
α_2	1,034	(0,947 : 1,120)	-	1,034	(0,949 : 1,121)	-
α_3	1,810	(1,713 : 1,907)	-	1,811	(1,712 : 1,906)	-
β_1	0,140	(0,131 : 0,150)	$\cong 0$	0,140	(0,131 : 0,150)	0

O efeito estimado do nível de radiação foi $\hat{\beta} = 0,147$, enquanto a razão de chances estimada foi $\hat{exp}(\beta) = 0,869$. Assim, para qualquer categoria $j = 0$, 1, 2 ou 3 do dano celular fixada, o aumento de uma unidade no nível de radiação recebido faz com que a chance estimada de se estar na direção mais danosa ao invés da menos danosa (isto é, $Y > j$ ao invés de $Y \leq j$) aumente em 15%. Esse aumento é de no

Tabela 8.4: Probabilidades observadas e estimadas através do modelo completo e reduzido, segundo a dose de radiação.

Dose (Gy)	**Probabilidade**	**Categoria de dano no DNA**			
		0	1	2	3
0	Observada	0,595	0,114	0,065	0,226
	Modelo Completo	0,526	0,215	0,120	0,139
	Modelo Reduzido	0,522	0,216	0,122	0,141
2,5	Observada	0,491	0,198	0,117	0,194
	Modelo Completo	0,435	0,230	0,147	0,188
	Modelo Reduzido	0,435	0,230	0,147	0,189
5	Observada	0,214	0,281	0,192	0,308
	Modelo Completo	0,350	0,231	0,170	0,250
	Modelo Reduzido	0,352	0,231	0,170	0,248
10	Observada	0,159	0,296	0,264	0,281
	Modelo Completo	0,209	0,196	0,191	0,403
	Modelo Reduzido	0,212	0,197	0,192	0,399
20	Observada	0,064	0,054	0,148	0,733
	Modelo Completo	0,063	0,085	0,126	0,726
	Modelo Reduzido	0,062	0,084	0,125	0,729

mínimo 14% com uma probabilidade de 95% (intervalo HPD).

A Figura 8.2 ilustra as probabilidades estimadas para cada categoria do dano.

O ajuste do modelo multi-categórico nominal particionado em Grazeffe et al. (2008) produziu estimativas mais próximas das proporções observadas. No entanto, esse modelo possui 8 parâmetros no total, contra apenas 4 do modelo logístico ordinal. A qualidade do ajuste e o princípio da parsimônia fazem com que o modelo ordinal seja uma forte alternativa para a análise estatística do problema em questão.

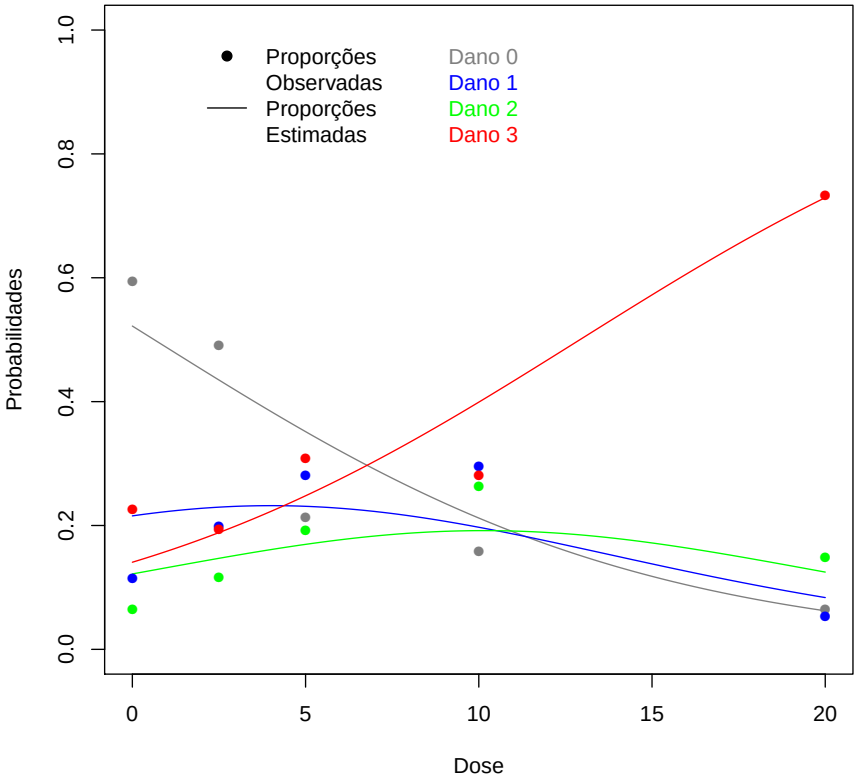

Figura 8.2: Probabilidades de danos empíricas e estimadas pelo modelo ordinal.

Capítulo 9

Conclusão

Neste trabalho foi apresentado o modelo logístico multi-categórico ordinal em uma abordagem bayesiana. Apesar de outras funções de ligação serem possíveis para a modelagem de dados ordinais, a *Logit* foi escolhida devido a sua interpretação interessante dos coeficientes do modelo (baseadas em razões de chances) e, até em consequência disso, da sua maior aplicabilidade na literatura.

A distribuição *a posteriori* dos parâmetros do modelo foi obtida com a utilização da verossimilhança derivada da distribuição multinomial da variável resposta Y, ao contrátio dos primeiros desenvolvimentos de uma abordagem bayesiana em modelos de regressão ordinal (como em Albert and Chib (1993)). Neles, a utilização de *Dados Aumentados* somados com a adoção da função de ligação *Probit* permitiam o ajuste do modelo, já que as restrições computacionais da época exigiam alternativas para o uso do $MCMC$. Essa combinação resulta em distribuições condicionais completas de formas conhecidas, viabilizando o uso do amostrador de Gibbs (Anexo B). Nos dias de hoje, o uso de *Dados Aumentados* não é mais uma necessidade devido o aumento da capacidade computacional e do avanço dos métodos $MCMC$.

Os testes de significância dos parâmetros do modelo foram feitos tomando como

medida de evidência estatística a medida de evidência bayesiana proposta por Pereira and Stern (1999). Não há muita discussão na literatura acerca de testes de significância para os parâmetros do modelo ordinal. Albert and Chib (1998) acessavam a significância de cada variável explicativa do modelo utilizando o Fator de Bayes, que compara o modelo completo com o modelo reduzido (excluindo-se a variável em questão) através das verossimilhanças marginais. O conceito de verossimilhança marginal pode ser visto em Kass and Raftery (1995). Chipman and Hamada (1993) sugeriram como evidência o "zero p value", definido como a proporção da distribuição posterior marginal de dado parâmetro de uma lado do zero:

$$p_k = 2min(\hat{F}_k(0), 1 - \hat{F}_k(0)),$$

onde $\hat{F}_k$ é a função de distribuição estimada do k-ésimo parâmetro de interesse. Valores pequenos de p_k indicam alta probabilidade de que o correspondente parâmetro é significante. Essa medida se iguala ao e-valor de Pereira and Stern (1999) quando $k = 1$ e a distribuição posterior é simétrica. A utilização do e-valor apresenta vantagens visto que é derivado de um procedimento genuinamente bayesiano e satisfaz diversas propriedades desejáveis de um teste estatístico.

Prioris não informativas foram adotadas para os parâmetros para a eventual comparação com os resultados clássicos em estudos de simulação. Notou-se que:

- As estimativas pontuais (estimativa de máxima verossimilhança e média *a posteriori*) e intervalares (intervalos de confiança e intervalos de credibilidade do tipo HPD) foram bem próximas, sobretudo para amostras grandes;

- Os e-valores foram mais conservadores que os p-valores clássicos, como em exemplos de simulações em Pereira and Stern (1999);

- Problemas referentes à hipóteses aninhadas, nos quais os p-valores apresentam contradições, não se verificam ao se utilizar o e-valor. Enquanto o p-valor é calculado no espaço amostral, o e-valor é calculado no espaço paramétrico completo, de forma que a hipótese de dimensão maior sempre terá um e-valor menor do que o da hipótese aninhada de dimensão menor.

Os resultados averiguados encorajam a utilização de uma abordagem bayesiana na modelagem de dados ordinais. A pouca ou não existente informação a priori permite se chegar a resultados bastante semelhante aos resultados clássicos, embora com interpretações de caráter mais sugestivos e interessantes nas inferências. Por outro lado, a possibilidade de se inserir informação a priori, seja ela diretamente sobre os parâmetros do modelo ou sobre as probabilidades cumulativas de sucesso (Johnson and Albert, 1998), configura mais uma vantagem para a aplicação de tal abordagem. Por fim, poder contar com uma medida de evidência para os testes de significância dos parâmetros que mantém as mais desejaveis propriedades do uso prático dos p-valores (sem, no entanto, ser sua mera versão bayesiana) e que é conceitualmente simples, teoricamente coerente e facilmente implementável pode encurtar ainda mais a distância da aplicação de métodos bayesianos nas pesquisas das mais diversas áreas.

Referências Bibliográficas

Agresti, A. (2007). *An Introduction Categorical Data Analysis*. John Wiley & Sons, 2nd edition.

Agresti, A. (2012). *Categorical Data Analysis*. John Wiley & Sons, 3rd edition.

Albert, J. and Chib, S. (1998). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. Technical report.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous respose data. *Journal of the American Statistical Association*, (88):669–679.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. John Wiley & Sons.

Barreto, A. S. (2011). *Modelos de Regressão: Teoria e Aplicações com o Programa Estatístico R*. Brasília: Ed. do Autor.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, (72):355–366.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, (2):317–352.

Berger, J. O. and Selke, T. (1987). Testing a point null hypothesis: The irreconciability of p-values and evidence. *Journal of the American Statistical Association*, (82):112–130.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley and Sons, New York.

Chipman, H. and Hamada, M. (1993). Bayesian analysis of ordered categorical data from industrial experiments. *IIQP Research Report*, (RR-93-06).

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley Series in Probability and Statistics.

Cowles, M. K. (1996). Accelerating monte carlo markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, (6):101–111.

Dickey, J. M. (1977). Is the tail area useful as an approximate bayes factor? *Journal of the American Statistical Association*, (72):138–142.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. CRC Press.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, (70):193–242.

Ehlers, R. S. (2007). *Inferência Bayesiana*. URL `www2.icmc.usp.br/~ehlers/bayes/`.

Gamerman, D. (1996). Simulação estocástica via cadeias de markov. *12 Sinape, Caxambu/MG*, pages 101–139.

Gamerman, D. (1997). *Morkov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman Hall. New York, ISBN: 0412818205.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling - based approaches to calculating marginal densities. *Journal of the American Statistical Association*, (85):398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741.

Good, I. J. (1988). *Surprise Index*. Encyclopedia of Statistical Sciences: John Wiley and Sons.

Grazeffe, V. S., Tallarico, L. F., Pinheiro, A. S., Kawano, T., Suzuki, M. F., Okazaki, K., Pereira, C. A. B., and Nakano, E. (2008). Establishment of the comet assay in the freshwater snail biomphalaria glabrata (say, 1818). *Mutation Research*, (654):58–63.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, (57):97–109.

Ishwaran, H. (2000). Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *Canadian Journal of Statistics*, (28):715–730.

Izbickil, R., Fossaluza, V., Hounie, A. G., Nakano, E. Y., and C. A. B. Pereira, C. A. B. (2012). Testing allele homogeneity: the problem of nested hypotheses. *BMC Genetics*, (13:):103.

Jaloszynski, P., Kujawski, M., Czub-Swierczek, M., Markowska, J., and Szyfter, K. (1997). Bleomycin-induced dna damage and its removal in lymphocytes of breast cancer patients studied by comet assay. *Mutation Research*, (385):223–233.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, U.K.: Oxford University Press, 3rd edition.

Johnson, V. E. and Albert, J. H. (1998). *Ordinal Data Modeling*. Springes-Verlag, New York.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, (90):773–795.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5th edition.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, (44):187–192.

Lindley, D. V. (1997). Some comments on bayes factors. *Journal of Statistical Planning and Inference*, (61):181–189.

Madruga, M. R., Esteves, L. G., and Wechsler, S. (2001). On the bayesianity of pereira-stern tests. *Test*, (10):291–299.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman Hall, 2nd edition.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, (21):1087–1092.

Okura, R. I. S. (2008). *Modelos de regressão para variáveis categóricas ordinais com aplicações ao problema de classificação*. Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo.

Parmigiani, G., Ashih, H., Samsa, G., Duncan, P., Lai, S., and Matchar, D. (2003). Cross-calibration of stroke disability measures: Bayesian analysis of longitudinal ordinal categorical data using negative dependence. *Journal of the American Statistical Association*, (98):273–281.

Paulino, C. D., Turkman, M. A. A., and Murteira, B. (2003). *Estatística Bayesiana*. Lisboa: Funda ¸cão Calouste Gulbenkian.

Pereira, C. and Wechsler, S. (1993). On the concept of p-value. *Brazilian Journal of Probability and Statistics*, (7):159–177.

Pereira, C. A. B. and Stern, J. M. (1999). Evidence and credibility: full bayesian significance test of precise hypothesis. *Entropy*, 1:99–110.

Pereira, C. A. B., Stern, J. M., and Wechsler, S. (2008). Can a significance test be genuinely bayesian? *Bayesian Analysis*, (1):79–100.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Santis, F. D. (2004). Statistical evidence and sample size determination for bayesian hypothesis testing. *Journal of the Statistical Planning and Inference*, (124):121–144.

Schervish, M. (1996). P values: what they are and what they are not. *Am Statistician*, (50):203–206.

Shafer, G. (1976). Lindley's paradox by glenn shafer. Technical Report 125, Stanford University, California.

Anexo A - Geração de dados

O Capítulo 7 se baseou na simulação de dados para a aplicação e comparação das abordagens clássica e bayesiana no contexto da regressão ordinal. Em cada um dos cenários discutidos e para cada tamanho de amostra n pré estabelecido houve a geração da variável resposta Y e das variáveis explicativas X_k como é descrito nos algoritmos abaixo:

- $J = 3$ **categorias ;** X **binário**

  1. gerar $x_i \sim Bernoulli(0,5)$, $i = 1, \cdots, n$
  2. especificar os valores dos parâmetros α_1, α_2 e β
  3. Para cada x_i, $i = 1, \cdots, n$, definir as probabilidades de cada categoria:

$$\pi_{i1} = \frac{exp(\alpha_1 - \beta x_i)}{1+exp(\alpha_1 - \beta x_i)}$$

$$\pi_{i2} = \frac{exp(\alpha_2 - \beta x_i)}{1+exp(\alpha_2 - \beta x_i)} - \frac{exp(\alpha_1 - \beta x_i)}{1+exp(\alpha_1 - \beta x_i)}$$

$$\pi_{i3} = \frac{1}{1+exp(\alpha_2 - \beta x_i)}$$

  4. gerar $Y_i \sim$Multinomial$(\pi_{i1}, \pi_{i2}, \pi_{i3})$, $i = 1, \cdots, n$

- $J = 3$ **categorias ;** X **contínuo**

  1. gerar $x_i \sim N(\mu, \sigma^2)$, $i = 1, \cdots, n$
  2. especificar os valores dos parâmetros α_1, α_2 e β

3. Para cada x_i, $i = 1, \cdots, n$, definir as probabilidades de cada categoria:

$$\pi_{i1} = \frac{exp(\alpha_1 - \beta x_i)}{1+exp(\alpha_1 - \beta x_i)}$$

$$\pi_{i2} = \frac{exp(\alpha_2 - \beta x_i)}{1+exp(\alpha_2 - \beta x_i)} - \frac{exp(\alpha_1 - \beta x_i)}{1+exp(\alpha_1 - \beta x_i)}$$

$$\pi_{i3} = \frac{1}{1+exp(\alpha_2 - \beta x_i)}$$

4. gerar $Y_i \sim$Multinomial$(\pi_{i1}, \pi_{i2}, \pi_{i3})$, $i = 1, \cdots, n$

- $J = 4$ **categorias ;** X_1 **binário e** X_2 **quantitativo**

1. gerar $x_{i1} \sim Bernoulli(0, 5)$, $i = 1, \cdots, n$

2. gerar $x_{i2} \sim N(\mu, \sigma^2)$, $i = 1, \cdots, n$

3. especificar os valores dos parâmetros α_1, α_2, α_3, β_1 e β_2

4. Para cada x_{i1} e x_{i2}, $i = 1, \cdots, n$, definir as probabilidades de cada categoria:

$$\pi_{i1} = \frac{exp(\alpha_1 - \beta_1 x_{i1} - \beta_2 x_{i2})}{1+exp(\alpha_1 - \beta_1 x_{i1} - \beta_2 x_{i2})}$$

$$\pi_{i2} = \frac{exp(\alpha_2 - \beta_1 x_{i1} - \beta_2 x_{i2})}{1+exp(\alpha_2 - \beta_1 x_{i1} - \beta_2 x_{i2})} - \frac{exp(\alpha_1 - \beta_1 x_{i1} - \beta_2 x_{i2})}{1+exp(\alpha_1 - \beta_1 x_{i1} - \beta_2 x_{i2})}$$

$$\pi_{i3} = \frac{exp(\alpha_3 - \beta_1 x_{i1} - \beta_2 x_{i2})}{1+exp(\alpha_3 - \beta_1 x_{i1} - \beta_2 x_{i2})} - \frac{exp(\alpha_2 - \beta_1 x_{i1} - \beta_2 x_{i2})}{1+exp(\alpha_2 - \beta_1 x_{i1} - \beta_2 x_{i2})}$$

$$\pi_{i4} = \frac{1}{1+exp(\alpha_3 - \beta_1 x_{i1} - \beta_2 x_{i2})}$$

5. gerar $Y_i \sim$Multinomial$(\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})$, $i = 1, \cdots, n$

Anexo B- Amostrador de Gibbs

O Amostrador de Gibbs (Geman and Geman (1984);Gelfand and Smith (1990)) é um esquema de simulação estocástica utilizando cadeias de Markov, cuja função geradora é formada pelas densidades condicionais completas. Torna-se possível gerar amostras de uma distribuição marginal sem a necessidade de se calcular analiticamente a sua densidade.

Seja $\boldsymbol{\theta} = (\theta_1, \quad \theta_2, \cdots, \quad \theta_k)$ um vetor paramétrico k-dimensional, onde θ_i $i = 1, 2, \cdots, k$, são variáveis aleatórias cuja densidade *a posteriori* é dada por:

$$h(\boldsymbol{\theta}|\boldsymbol{y}) = h(\theta_1, \quad \theta_2, \cdots, \quad \theta_k),$$

onde $\boldsymbol{y}$ representa o conjunto dos dados observados.

Suponha que o interesse esteja na geração de uma amostra de $h(\boldsymbol{\theta}|\boldsymbol{y})$ e que a geração direta da posteriori conjunta é extremamente complicada/custosa, mas que as gerações das condicionais $h(\theta_i|\boldsymbol{\theta_{(i)}}, \boldsymbol{y})$ (onde $\boldsymbol{\theta_{(i)}}$ é o vetor $\boldsymbol{\theta}$ sem o i-ésimo componente) são possíveis de ser realizadas. O algoritmo de Gibbs então fornece uma forma alternativa de geração baseada em sucessivas gerações das distribuições condicionais $h(\theta_i|\boldsymbol{\theta_{(i)}}, \boldsymbol{y})$, $i = 1, 2, \cdots, k$. O algoritmo é descrito da seguinte forma (Gamerman, 1996):

(1) inicializa-se o contador de iterações da cadeia $j = 1$, e escolhe-se arbitrariamente os valores iniciais $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, \quad \theta_2^{(0)}, \cdots, \quad \theta_k^{(0)})$;

(2) obtem-se um novo valor $\boldsymbol{\theta^{(j)}} = (\theta_1^{(j)}, \quad \theta_2^{(j)}, \cdots, \quad \theta_k^{(j)})$ a partir de $\boldsymbol{\theta^{(j-1)}}$ através de sucessivas gerações de valores

$$\theta_1^{(j)} \sim h(\theta_1 | \theta_2^{(j-1)}, \cdots, \theta_k^{(j-1)}, \boldsymbol{y})$$

$$\theta_2^{(j)} \sim h(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \cdots, \theta_k^{(j-1)}, \boldsymbol{y})$$

$$\vdots$$

$$\theta_k^{(j)} \sim h(\theta_k | \theta_1^{(j)}, \cdots, \theta_{k-1}^{(j)}, \boldsymbol{y});$$

(3) o contador é atualizado de j para $j+1$ e retorna-se a (2) até a convergência.

Assume-se que a convergência é atingida em uma iteração cuja distribuição esteja arbitrariamente próxima da distribuição de equilíbrio $h(\boldsymbol{\theta}|\boldsymbol{y})$ e não no sentido formal e inatingível do número de iterações tendendo ao infinito.

A forma de obter-se uma amostra de tamanho n é replicar a cadeia m vezes até a convergência (período de burn-in). Após a convergência, todas gerações de uma mesma cadeia são gerações da distribuição de equilíbrio e sucessivos valores dessa cadeia também formam uma amostra de $h(\boldsymbol{\theta}|\boldsymbol{y})$. Com a convergência da cadeia, a amostra pode ser retirada tomando-se saltos entre os valores gerados, de forma a evitar a dependência entre o valor gerado e o valor anterior.

Taxas e diagnósticos de convergência para as cadeias podem ser utilizados tais como Geman and Geman (1984), Gelfand and Smith (1990) entre outros.

Anexo C - Metropolis-Hastings

O algoritmo de Metropolis-Hastings (Metropolis et al. (1953) e Hastings (1970)) é utilizado para gerar amostras de uma distribuição conjunta nos casos onde as densidades condicionais apresentam formas (distribuições) desconhecidas.

O interesse está em gerar valores de uma densidade $h(\theta_i|\boldsymbol{\theta_{(i)}}, \boldsymbol{y})$. Para simplificar a notação, seja $\theta_i = \theta$ e $h(\theta)$ a distribuição marginal desejada.

Suponha que a cadeia esteja no estado $\theta^{(j-1)}$ e um valor θ' é gerado de uma distribuição proposta $q(.|\theta^{(j-1)})$ (núcleo de transição que define a função geradora do novo estado da cadeia). O novo valor θ' é aceito com probabilidade

$$\alpha(\theta^{(j-1)}, \theta') = min\left(1, \frac{h(\theta')q(\theta^{(j-1)}|\theta')}{h(\theta^{(j-1)})q(\theta'|\theta^{(j-1)})}\right),$$

onde $h(.)$ é o núcleo da distribuição posterior desejada. O algoritmo de Metropolis-Hastings é dado pelos seguintes passos:

(1) inicia-se arbitrariamente com um ponto qualquer $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, \quad \theta_2^{(0)}, \cdots, \quad \theta_k^{(0)})$ e também o contador j=1;

(2) Gera-se um novo valor θ' da distribuição $q(\theta'|\theta^{(j-1)})$

(3) Calcula-se a probabilidade de aceitação $\alpha(\theta^{(j-1)}, \theta')$ e simula-se g da Distribuição Uniforme Contínua no intervalo $[0, 1]$, ou seja, $g \sim U(0, 1)$

(4) Se $g \leq \alpha(\theta^{(j-1)}, \theta')$ aceita-se o novo valor $\theta' = \theta^{(j)}$ e faz-se $j = j + 1$. Caso contrário, a cadeia permance em $\theta^{(j-1)}$ e reinicia-se o processo a partir do passo 2 até a convergência.

O núcleo de transição $q(.)$ define apenas uma proposta de movimento que pode ou não ser confirmado por α. Por esse motivo, $q(.)$ é normalmente chamado de proposta e quando olhado como uma densidade (ou distribuição) condicional, é chamado de densidade (distribuição) proposta.

Escolha de $q(.|\theta^{(j-1)})$

(a) "Cadeias passeio aleatório": $q(\theta|\theta') = q_1(|\theta' - \theta|)$, onde $q_1(.)$ é uma densidade multivariada. Neste caso, $\theta' = \theta + \epsilon$, onde ϵ é a variável incremento com distribuição $q_1(.)$. No caso em que $q_1(\epsilon) = q_1(-\epsilon)$ tem-se

$$\alpha = min\left(1, \frac{h(\theta')}{h(\theta^{(j-1)})}\right);$$

(b) "Cadeias independentes": $q(\theta|\theta') = q_1(\theta')$ (Hastings, 1970), forma uma cadeia independente da iteração anterior com

$$\alpha = min\left(1, \frac{h(\theta')q(\theta^{(j-1)})}{q(\theta')h(\theta^{(j-1)})}\right);$$

(c) "Cadeias simétricas": se $h(\theta) \propto c(\theta)\psi(\theta)$, onde $c(\theta)$ é uma densidade que pode ser amostrada e ψ uniformemente limitada, toma-se

$$q(\theta|\theta') = c(\theta')$$

Neste caso particular (o mais eficiente na prática), tem-se que

$$\alpha = min\left(1, \frac{\psi(\theta')}{\psi(\theta^{(j-1)})}\right);$$