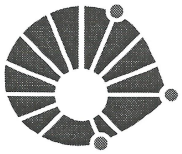
Pedro Luiz Baldoni

Modelos lineares generalizados mistos multivariados para caracterização genética de doenças

CAMPINAS

2014



Universidade Estadual de Campinas

Instituto de Matemática, Estatística
e Computação Científica

Pedro Luiz Baldoni

Modelos lineares generalizados mistos multivariados para caracterização genética de doenças

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em estatística.

Orientadora: Hildete Prisco Pinheiro

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Pedro Luiz Baldoni, e orientada pela Profa. Dra. Hildete Prisco Pinheiro.

Assinatura da Orientadora

Hildete Pinheiro

Campinas
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

B193m

Baldoni, Pedro Luiz, 1989-
Modelos lineares generalizados mistos multivariados para caracterização genética de doenças / Pedro Luiz Baldoni. – Campinas, SP : [s.n.], 2014.

Orientador: Hildete Prisco Pinheiro.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Modelos lineares generalizados. 2. Inferência estatística. 3. Genética quantitativa. I. Pinheiro, Hildete Prisco,1966-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Multivariate generalized linear mixed models for genetic characterization of diseases
Palavras-chave em inglês:
Generalized linear models
Statistical inference
Quantitative genetics
Área de concentração: Estatística
Titulação: Mestre em Estatística
Banca examinadora:
Hildete Prisco Pinheiro [Orientador]
Víctor Hugo Lachos Dávila
Clarice Garcia Borges Demétrio
Data de defesa: 26-02-2014
Programa de Pós-Graduação: Estatística

Dissertação de Mestrado defendida em 26 de fevereiro de 2014 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Hildete Pinheiro

Prof(a). Dr(a). HILDETE PRISCO PINHEIRO

Prof(a). Dr(a). VÍCTOR HUGO LACHOS DÁVILA

C B Demétrio

Prof(a). Dr(a). CLARICE GARCIA BORGES DEMÉTRIO

Abstract

Generalized Linear Mixed Models (GLMM) are a generalization of Linear Mixed Models (LMM) and of Generalized Linear Models (GLM). The class of models GLMM extends the normality assumption of the data and allows the use of several other probability distributions, for example, accommodating the over dispersion often observed and also the correlation among observations in longitudinal or repeated measures studies. However, the likelihood theory of the GLMM class is not straightforward since its likelihood function has not closed form and involves a high order dimensional integral.

In order to solve this problem, several methodologies were proposed in the literature, from classical techniques as numerical quadratures, for example, up to sophisticated methods involving EM algorithm, MCMC methods and penalized quasi-likelihood. These methods have advantages and disadvantages that must be evaluated in each problem. In this work, the penalized quasi-likelihood method (Breslow and Clayton 1993) was used to model infection data in a population of dairy cattle because demonstrated to be robust in the problems faced in the likelihood theory of this data. Moreover, the other methods do not show to be treatable faced to the complexity existing in quantitative genetics.

Additionally, simulation studies are presented in order to verify the robustness of this methodology. The stability of these estimators and the robust theory of this problem are not completely studied in the literature.

Keywords: Generalized linear models, Statistical inference, Quantitative genetics.

Resumo

Os Modelos Lineares Generalizados Mistos (MLGM) são uma generalização natural dos Modelos Lineares Mistos (MLM) e dos Modelos Lineares Generalizados (MLG). A classe dos MLGM estende a suposição de normalidade dos dados permitindo o uso de várias outras distribuições bem como acomoda a superdispersão frequentemente observada e também a correlação existente entre

observações em estudos longitudinais ou com medidas repetidas. Entretanto, a teoria de verossimilhança para MLGM não é imediata uma vez que a função de verossimilhança marginal não possui forma fechada e envolve integrais de alta dimensão.

Para solucionar este problema, diversas metodologias foram propostas na literatura, desde técnicas clássicas como quadraturas numéricas, por exemplo, até métodos sofisticados envolvendo algoritmo EM, métodos MCMC e quase-verossimilhança penalizada. Tais metodologias possuem vantagens e desvantagens que devem ser avaliadas em cada tipo de problema. Neste trabalho, o método de quase-verossimilhança penalizada (Breslow e Clayton 1993) foi utilizado para modelar dados de ocorrência de doença em uma população de vacas leiteiras pois demonstrou ser robusto aos problemas encontrados na teoria de verossimilhança deste conjunto de dados. Além disto, os demais métodos não se mostram calculáveis frente à complexidade dos problemas existentes em genética quantitativa.

Adicionalmente, estudos de simulação são apresentados para verificar a robustez de tal metodologia. A estabilidade dos estimadores e a teoria de robustez para este problema não estão completamente desenvolvidos na literatura.

Palavras-chave: Modelos lineares generalizados, Inferência estatística, Genética quantitativa.

Sumário

Aos meus pais ...

Agradecimentos

Aos meus pais que sempre incentivaram e apoiaram meus estudos e que, muitas vezes, abriram mãos de seus sonhos para que eu pudesse realizar os meus. Agradeço também a minha família por todo o apoio e compreensão nos momentos frequentes de ausência e de ansiedade nestes últimos anos.

Meus sinceros agradecimentos a minha namorada Fernanda, pelo companheirismo, compreensão e, principalmente, por me surpreender cada dia mais com o seu amor.

À Professora Hildete Prisco Pinheiro, por sua orientação acadêmica e pessoal durante o desenvolvimento desta dissertação.

Ao Professor Rodrigo Labouriau, com quem tive o prazer de trabalhar durante seis meses na Universidade de Aarhus, Dinamarca, que me recebeu de braços abertos e é responsável por grande parte da realização deste trabalho.

Aos Professores Víctor Hugo Lachos e Clarice Borges Demétrio pelas valiosas sugestões e críticas construtivas, e por aceitarem participar da banca da defesa desta dissertação.

Aos professores do Departamento de Estatística, que contribuíram de maneira direta e/ou indireta em minha formação durante os sete anos em que estive presente nesta instituição. Em especial, agradeço ao professor Maurício Zevallos, excelente docente, com quem trabalhei durante os últimos anos de graduação e nos primeiros momentos de pós-graduação.

Aos colegas de pós-graduação de Campinas e de Foulum, Dinamarca, que me acompanharam durante estes dois anos de mestrado. Meu especial agradecimento aos colegas Rafael Maia e Beatriz

Cuyabano por todo o apoio e conselhos acadêmicos.

Agradeço à CAPES e ao Banco Santander pelo apoio financeiro.

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A classe dos Modelos Lineares Generalizados Mistos (MLGM) é uma extensão natural dos Modelos Lineares Mistos (MLM) e dos Modelos Lineares Generalizados (MLG) (McCullagh e Nelder 1989). Como tais, os MLGM são de grande importância e possuem diversas aplicações dada a sua capacidade de modelar a super dispersão dos dados (Williams 1982) e a dependência entre observações em estudos longitudinais (Stiratelli, Laird e Ware 1984) ou em dados com medidas repetidas (Breslow 1984), quando incorporamos efeitos aleatórios. Dentro do contexto de genética quantitativa, a inclusão de tais efeitos é uma etapa fundamental na definição do modelo estatístico, pois é por meio do uso de componentes aleatórias que será modelada e inferida a dependência genética existente.

Além disto, a classe dos MLGM permite acomodar outras distribuições da família exponencial como as distribuições gama, inversa gaussiana, binomial e poisson, por exemplo, além de permitir o uso de funções de respostas não lineares.

Neste trabalho, o foco principal é a apresentação da classe dos MLGM via discussão teórica e prática por meio da análise de um conjunto de dados de vacas leiteiras dinamarquesas. Aspectos inferenciais desta classe de modelos podem ser estudados e questionamentos genéticos acerca da amostra em estudo podem ser respondidos por meio de tal análise. Tais questionamentos verificam,

por exemplo, se os programas de seleção genética aos quais a população de animais em produção está sujeita contribui para a melhoria ou piora da resistência a doenças com a ocorrência registrada no sistema de produção.

No decorrer deste trabalho, extensões multivariadas dos MLGM serão apresentadas, i.e., extensões que permitem a análise simultânea do valor genético de várias características simultaneamente. Tais extensões são fundamentais para avaliar o impacto de programas de seleção genética na resistência a doenças. Adicionalmente, estudos de validação cruzada são apresentados para estudar aspectos de robustez dos modelos ajustados. Estudos similares podem, eventualmente, ser desenvolvidos com dados brasileiros no futuro próximo.

O interesse de pesquisadores em compreender estruturas genéticas, dentro do contexto de programas de melhoramento genético e de seleção animal, justifica a importância de estudos estatísticos sobre este tema. A análise da transmissão genética dentro de uma população requer a utilização de modelos estatísticos detalhados em virtude da grande complexidade inerente ao processo de herdabilidade.

Muito embora a teoria de máxima verossimilhança seja largamente utilizada e estudada nos MLM e MLG, durante muitos anos esta se restringiu aos modelos mais simples nas classes dos MLMG devido à necessidade de se calcular, numericamente, integrais de alta dimensão. Entretanto, avanços computacionais e teóricos permitiram novas abordagens no contexto da teoria de verossimilhança e, consequentemente, modelos mais complexos puderam ser ajustados a dados reais.

Métodos para maximização da função de verossimilhança por meio do algoritmo EM e do algoritmo Newton-Raphson foram apresentados por McCulloch (1997). Neste trabalho, o autor utilizou métodos de simulação estocástica, via MCMC, nas etapas destes algoritmos para solucionar o problema do cálculo de integrais de alta dimensão. Besag, York e Mollié (1991) apresentaram uma abordagem bayesiana para a solução deste tipo de problema por meio da utilização de amostras da distribuição *a posteriori* via amostrador de Gibbs.

Aproximações da função de verossimilhança marginal, utilizando o método de Laplace e funções de quase-verossimilhança, foram estudadas por Breslow e Clayton (1993). A metodologia desenvolvida pelos autores utiliza a quase-verossimilhança penalizada e pseudo-verossimilhança para estimação dos parâmetros de locação e escala do MLGM, apresentando resultados inferenciais bastante razoáveis em modelos hierárquicos, por exemplo.

Ainda na introdução deste trabalho é apresentada a classe dos MLGM. Adicionalmente, uma breve discussão é apresentada sobre a teoria genética envolvida na construção dos modelos estatísticos utilizados em melhoramento genético animal.

O restante deste trabalho está organizado da seguinte maneira. Aspectos inferenciais dos MLGM utilizados no problema em questão são apresentados no Capítulo 2. Em seguida, o Capítulo 3 apresenta a análise dos dados de vacas leiteiras dinamarquesas e um breve estudo de simulação sobre a teoria de robustez dos modelos ajustados. Por fim, uma discussão geral acerca do problema estudado é apresentada no Capítulo 4.

1.1 Modelos Lineares Generalizados Mistos

Nesta seção iremos definir a classe dos MLGM, explorar as consequências da introdução dos efeitos aleatórios e discutir alguns aspectos inferenciais destes modelos.

1.1.1 O modelo

Os dados consistem em n observações em que, para $i = 1, \ldots, n$, a i-ésima observação é dada pela tripla $(y_i, \mathbf{x}_i^{'}, \mathbf{z}_i^{'})$. Aqui, y_i é a realização de uma variável aleatória Y_i, representando a variável resposta, e $\mathbf{x}_i^{'}$ e $\mathbf{z}_i^{'}$ são vetores de variáveis explanatórias associadas aos efeitos fixos e aleatórios, respectivamente. No contexto de melhoramento animal, o *pedigree*[1] será introduzido e especificado em termos do vetor de variáveis explanatórias associado aos efeitos aleatórios (Seção 1.2.2). Tais

[1]Diagrama apresentando a genealogia de um indivíduo e de seus ancestrais diretos com o objetivo de analisar ou seguir a herança de determinada característica.

observações podem ser agrupadas se considerarmos observações repetidas de uma mesma unidade amostral, por exemplo.

Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ o vetor de variáveis resposta das n unidades amostrais. Para especificar o modelo, iniciaremos com a distribuição condicional de $\mathbf{Y}$ dado $\mathbf{U}$, em que $\mathbf{U}$ é um vetor q-dimensional de efeitos aleatórios. Analogamente aos MLG, assumiremos que $\mathbf{Y}$ consistirá de observações condicionalmente independentes com função densidade de acordo com o modelo exponencial com parâmetro de dispersão (Jorgensen, Labouriau e Lundbye-Christensen 1996):

$$\begin{aligned} Y_i|\mathbf{U} = \mathbf{u} &\sim \text{indep.} \quad f_{Y_i|\mathbf{U}=\mathbf{u}}(y_i) \\ f_{Y_i|\mathbf{U}=\mathbf{u}}(y_i) &= \exp\{[y_i\gamma_i - b(\gamma_i)]/a(\phi) - c(y_i, \phi)\} \end{aligned} \tag{1.1.1}$$

para funções dadas $a(\cdot), b(\cdot)$ e $c(\cdot)$. A função $a(\phi)$ é comumente encontrada na forma $a(\phi) = \phi \cdot a_i$, em que ϕ, chamado de parâmetro de dispersão, é constante dentre todas as observações e a_i são constantes conhecidas. A função $b(\cdot)$ é a função geradora de cumulantes.

A Tabela 1.1 apresenta as funções $a(\cdot), b(\cdot)$ e $c(\cdot)$ para as distribuições normal, gama, binomial e poisson.

Tabela 1.1: Funções $a(\cdot), b(\cdot)$ e $c(\cdot)$ para as distribuições Normal, Gama, Binomial e Poisson

	Normal (μ, σ^2)	Gama (α, β)	Binomial (n, p)	Poisson (λ)
γ	μ	$-\frac{\beta}{\alpha}$	$\log\left[\frac{p}{1-p}\right]$	$\log(\lambda)$
ϕ	σ^2	$\frac{1}{\alpha}$	1	1
$a(\phi)$	ϕ	ϕ	1	1
$b(\gamma)$	$\frac{\gamma^2}{2}$	$-\log(-\gamma)$	$n\log[1+e^{\gamma}]$	e^{γ}
$c(y, \phi)$	$\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$	$\log(\phi)/\phi - \log\Gamma(1/\phi)$ $+(1/\phi - 1)\log(y)$	$\log\left[\binom{n}{y}\right]$	$\log(y!\,)$

Analogamente aos MLG, iremos modelar uma transformação da média, que será possivelmente

uma função de γ_i, como um modelo linear nos efeitos fixos e aleatórios:

$$\begin{aligned} \mathrm{E}[Y_i|\mathbf{U}=\mathbf{u}] &= \mu_i \\ g(\mu_i) &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}, \end{aligned} \tag{1.1.2}$$

em que a função $g(.)$ é dada, chamada de função de ligação, e $\boldsymbol{\beta}$ é o vetor de parâmetros de efeitos fixos. Para completar a especificação do modelo, devemos atribuir uma distribuição de probabilidade aos efeitos aleatórios:

$$\mathbf{U} \sim f_{\mathbf{U}}(\mathbf{u}). \tag{1.1.3}$$

Alguns resultados imediatos desta classe de modelos podem ser obtidos por meio das seguintes relações, que valem sob as condições de regularidade da família exponencial (Casella e Berger 1990). Com o intuito de facilitar a leitura, toda a notação condicional será expressa em termos de $\mathbf{u}$ e não em termos de $\mathbf{U}=\mathbf{u}$.

$$\mathrm{E}\left[\frac{\partial \log f_{Y_i|\mathbf{U}}(y_i|\mathbf{u})}{\partial \gamma_i}\Big|\mathbf{u}\right] = 0 \tag{1.1.4}$$

$$\mathrm{Var}\left[\frac{\partial \log f_{Y_i|\mathbf{U}}(y_i|\mathbf{u})}{\partial \gamma_i}\Big|\mathbf{u}\right] = -\mathrm{E}\left[\frac{\partial^2 \log f_{Y_i|\mathbf{U}}(y_i|\mathbf{u})}{\partial \gamma_i^2}\Big|\mathbf{u}\right]. \tag{1.1.5}$$

Usando (1.1.1) e (1.1.4), temos:

$$\begin{aligned} \mathrm{E}\left[\left\{Y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\} \Big/ a(\phi)\Big|\mathbf{u}\right] &= 0 \\ \mathrm{E}[Y_i|\mathbf{u}] &= \frac{\partial b(\gamma_i)}{\partial \gamma_i} = \mu_i. \end{aligned} \tag{1.1.6}$$

Usando (1.1.1) e (1.1.5), temos:

$$\begin{aligned}
\text{Var}\left[\left\{Y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\} \Big/ a(\phi) \Big| \mathbf{u}\right] &= -\text{E}\left[-\frac{1}{a(\phi)}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\Big|\mathbf{u}\right] \\
\text{Var}\left[\frac{Y_i - \mu_i}{a(\phi)}\Big|\mathbf{u}\right] &= \frac{1}{a(\phi)}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} \\
\text{Var}[Y_i|\mathbf{u}] &= a(\phi)\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2} \\
\text{Var}[Y_i|\mathbf{u}] &= a(\phi)\upsilon(\mu_i), \qquad (1.1.7)
\end{aligned}$$

em que $\upsilon(\mu_i)$ é $\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}$, chamada de função de variância, determina univocamente o modelo exponencial com parâmetro dispersão e indica como a variância condicional de $Y_i|\mathbf{u}$ se relaciona com a esperança condicional de $Y_i|\mathbf{u}$.

1.1.2 Consequências da introdução de efeitos aleatórios

É importante notar que as relações estabelecidas em (1.1.6) e (1.1.7) referem-se à distribuição condicional $Y_i|\mathbf{u}$, representando sua esperança e variância, respectivamente. Entretanto, alguns aspectos relacionados à distribuição marginal de Y_i podem ser obtidos de maneira análoga:

$$\begin{aligned}
\text{E}[Y_i] &= \text{E}[\text{E}[Y_i|\mathbf{u}]] \\
&= \text{E}[\mu_i] \\
&= \text{E}[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})], \qquad (1.1.8)
\end{aligned}$$

além disso:

$$\begin{aligned}
\text{Var}[Y_i] &= \text{Var}[\text{E}[Y_i|\mathbf{u}]] + \text{E}[\text{Var}[Y_i|\mathbf{u}]] \\
&= \text{Var}[\mu_i] + \text{E}[a(\phi)\upsilon(\mu_i)] \\
&= \text{Var}[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})] + \text{E}[a(\phi)\upsilon(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}))]. \qquad (1.1.9)
\end{aligned}$$

Para ilustrar os resultados (1.1.8) e (1.1.9), considere os seguintes exemplos. Assumiremos que $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, em que $\mathbf{0}$ é um vetor nulo q-dimensional e $\mathbf{\Sigma}$ é uma matriz $q \times q$ positiva definida.

Distribuição Normal

De acordo com a Tabela 1.1 e os resultados (1.1.6), (1.1.7), (1.1.8) e (1.1.9), valem os seguintes resultados para a função de ligação identidade $g(x) = x$

$$\begin{aligned} \mathrm{E}[Y_i] &= \mathrm{E}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}] = \mathbf{x}_i'\boldsymbol{\beta}, \\ \mathrm{Var}[Y_i] &= \mathrm{Var}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}] + \mathrm{E}[\sigma^2] = \mathbf{z}_i'\mathbf{\Sigma}\mathbf{z}_i + \sigma^2 \end{aligned}$$

e para a ligação logarítimica $g(x) = \log(x)$

$$\begin{aligned} \mathrm{E}[Y_i] &= \mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathrm{E}[\exp\{\mathbf{z}_i'\mathbf{u}\}] = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}M_{\mathbf{U}}(\mathbf{z}_i), \\ \mathrm{Var}[Y_i] &= \mathrm{Var}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] + \mathrm{E}[\sigma^2] = \mathrm{Var}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] + \sigma^2 \\ &= \mathrm{E}[\exp\{2(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})\}] - (\mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}])^2 + \sigma^2 \\ &= \exp\{2\mathbf{x}_i'\boldsymbol{\beta}\}(M_{\mathbf{U}}(2\mathbf{z}_i) - [M_{\mathbf{U}}(\mathbf{z}_i)]^2) + \sigma^2, \end{aligned}$$

em que $M_{\mathbf{U}}(.)$ é a função geradora de momentos de $\mathbf{U}$.

Distribuição Poisson

De acordo com a Tabela 1.1 e os resultados (1.1.6), (1.1.7), (1.1.8) e (1.1.9), valem os seguintes resultados para a função de ligação identidade $g(x) = x$

$$\begin{aligned} \mathrm{E}[Y_i] &= \mathrm{E}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}] = \mathbf{x}_i'\boldsymbol{\beta}, \\ \mathrm{Var}[Y_i] &= \mathrm{Var}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}] + \mathrm{E}[\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}] = \mathbf{z}_i'\mathbf{\Sigma}\mathbf{z}_i + \mathbf{x}_i'\boldsymbol{\beta} \end{aligned}$$

e para a ligação logarítimica $g(x) = \log(x)$

$$\begin{aligned}
\mathrm{E}[Y_i] &= \mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}\mathrm{E}[\exp\{\mathbf{z}_i'\mathbf{u}\}] = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}M_{\mathbf{U}}(\mathbf{z}_i), \\
\mathrm{Var}[Y_i] &= \mathrm{Var}[\mu_i] + \mathrm{E}[\mu_i] \\
&= \mathrm{Var}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] + \mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] \\
&= \mathrm{E}[\exp\{2(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u})\}] - (\mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}])^2 + \mathrm{E}[\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}] \\
&= \exp\{2\mathbf{x}_i'\boldsymbol{\beta}\}[M_{\mathbf{U}}(2\mathbf{z}_i) - [M_{\mathbf{U}}(\mathbf{z}_i)]^2 + \exp\{-\mathbf{x}_i'\boldsymbol{\beta}\}M_{\mathbf{U}}(\mathbf{z}_i)]. \qquad (1.1.10)
\end{aligned}$$

Neste exemplo, se fizermos a suposição adicional que $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_q)$ e que cada linha de $\mathbf{z}_i$ possua apenas uma entrada não nula, igual a 1, então:

$$\begin{aligned}
\mathrm{Var}[Y_i] &= \exp\{2\mathbf{x}_i'\boldsymbol{\beta}\}[\exp\{2\sigma_U^2\} - \exp\{\sigma_U^2\}] + \exp\{2\mathbf{x}_i'\boldsymbol{\beta}\}\exp\{\sigma_U^2/2\} \\
&= \mathrm{E}[Y_i][\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}(\exp\{3\sigma_U^2/2\} - \exp\{\sigma_U^2/2\}) + 1]. \qquad (1.1.11)
\end{aligned}$$

Note que, neste caso, o termo entre parêntesis é maior do que 1 e, portanto, a variância marginal de Y_i é maior que sua média. Logo, muito embora a distribuição condicional dos dados seja Poisson, a distribuição marginal não é. De fato, tal distribuição marginal será caracterizada pela sua *superdispersão*, nos levando a considerar a inclusão dos efeitos aleatórios como uma forma de atribuir superdispersão aos dados.

Outra característica importante da classe dos MLGM (e também dos MLM) é a sua capacidade de modelar e introduzir correlação entre observações que compartilham os mesmos efeitos aleatórios. Considerando independência condicional entre os elementos i e j de $\mathbf{Y}$, temos:

$$\begin{aligned}
\mathrm{Cov}(Y_i, Y_j) &= \mathrm{Cov}(\mathrm{E}[Y_i|\mathbf{u}], \mathrm{E}[Y_j|\mathbf{u}]) + \mathrm{E}[\mathrm{Cov}(Y_i, Y_j|\mathbf{u})] \\
&= \mathrm{Cov}(\mu_i, \mu_j) + \mathrm{E}[0] \\
&= \mathrm{Cov}(g^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}), g^{-1}(\mathbf{x}_j'\boldsymbol{\beta} + \mathbf{z}_j'\mathbf{u})). \qquad (1.1.12)
\end{aligned}$$

Note que a estrutura de variâncias e covariâncias entre as unidades amostrais será um resultado imediato da matriz de variâncias e covariâncias da distribuição de $\mathbf{U}$.

1.2 Discussão sobre teoria genética

Nesta seção, iremos introduzir brevemente alguns conceitos da teoria genética utilizada na aplicação da metodologia desenvolvida neste trabalho. A teoria genética é um vasto campo de pesquisa e, aqui, nos limitaremos a apresentar brevemente somente as definições e conceitos necessários para a construção dos modelos estatísticos que serão ajustados.

Resumidamente, a utilização da metodologia estatística dentro do contexto de genética quantitativa utiliza as estruturas e relações genéticas existentes entre os indivíduos em estudo para relacioná-los por meio do *pedigree*.

1.2.1 Medidas de grau de parentesco

As medidas de parentesco apresentadas nesta seção são base para a construção dos modelos estatísticos que serão utilizados neste trabalho. A definição formal de tais medidas é importante pois são elas que determinam as relações e correlações que entrarão como componentes genéticas em cada modelo estatístico. Todas elas possuem, invariavelmente, ao menos duas características em comum (Lynch e Walsh 1998).

A primeira característica considera que parentesco somente pode ser definido com respeito à alguma referência. Muito embora os membros de uma população sejam relacionados uns aos outros em determinado grau pois estes apresentam cópias de genes[2] que estavam presentes em algum ancestral remoto, por exemplo, todas as medidas de parentesco consideram uma geração de referência em que os indivíduos desta geração são considerados, aproximadamente, não correlacionados e, a partir desta geração, constroem-se os *pedigrees* observados.

[2]Segmento de DNA que contém informações para a síntese de uma ou mais proteínas.

A segunda característica compartilhada entre as medidas de parentesco é que estas são baseadas nos conceitos de identidade por descendência e identidade por estado. Dizemos que dois alelos[3] são idênticos por descendência se estes são cópias idênticas de um mesmo alelo carregado por algum ancestral. Por outro lado, dizemos que dois alelos são idênticos por estado se estes simplesmente expressam o mesmo efeito fenotípico, isto é, possuem sequências idênticas de nucleotídeos. Portanto, todo par de genes idênticos por descendência são necessariamente idênticos por estado mas a recíproca não é verdadeira.

A Figura 1.1 apresenta um exemplo do cálculo do número e da proporção de alelos idênticos por descendência. Neste caso, comparando-se os dois primeiros indivíduos gerados do cruzamento entre genitores com alelos $(1, 2)$ e $(3, 4)$, respectivamente, temos que o número de alelos idênticos por descendência é igual a 2, uma vez que ambos indivíduos herdaram os mesmos alelos paternos e maternos. Entretanto, quando comparamos o segundo e o terceiro indivíduo desta geração, temos que estes compartilham somente um alelo idêntico por descendência, o alelo materno 3. Portanto, temos neste caso que a proporção de alelos identicos por descendencia entre estes dois indivíduos é igual a $1/2$.

Um conceito muito utilizado no contexto de genética quantitativa é o conceito de consanguinidade. A definição deste conceito é importante pois é por meio dele que construiremos a matriz de relação genética entre os indivíduos em estudo.

Definição 1.2.1. Dois indivíduos são considerados *consanguíneos* se estes contêm pares de alelos idênticos por descendência.

Os conceitos de identidade por descendência e identidade por estado não se restringem a um indivíduo. Considerando dois indivíduos diplóides e um único lócus gênico, podemos ter 15 possíveis configurações diferentes de identidade por descendência para os quatro genes envolvidos (Lynch e Walsh 1998, Figura 7.2). Ignorando a distinção entre a contribuição paterna e materna, estas

[3]Uma de duas ou mais versões de uma sequência genética em uma localização particular no genôma.(Feero, Guttmacher e Collins 2010).

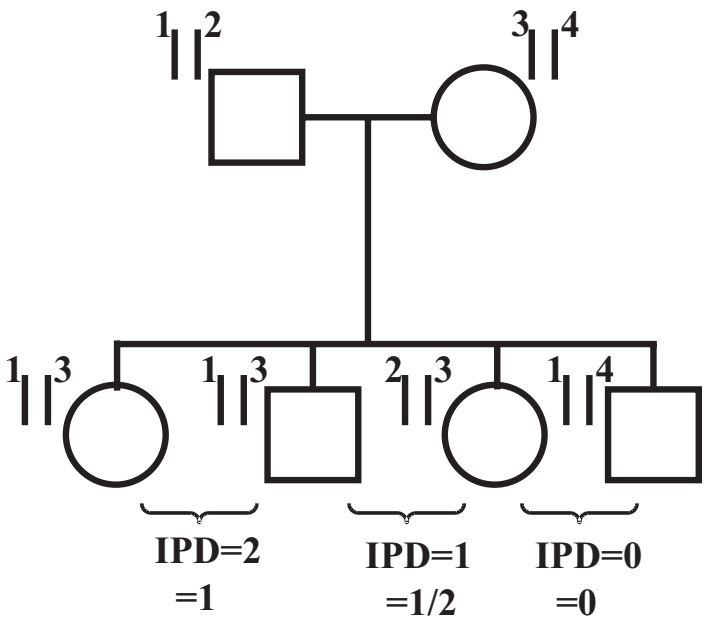


Figura 1.1: Cálculo da identidade por descendência (Andrade e Pinheiro 2002)

quinze configurações podem ser reorganizadas em nove *estados de identidade*. Probabilidades estão associadas a cada um destes estados de identidade e são denotadas por *coeficientes de identidade condensados*.

A partir dos coeficientes de identidade condensados é possível obter o chamado coeficiente de consanguinidade.

Definição 1.2.2. O coeficiente de consanguinidade entre dois indivíduos i e j, denotado por $\Theta_{i,j}$, é definido como a probabilidade de que, dado dois genes escolhidos ao acaso entre os indivíduos, estes sejam idênticos por descendência.

Wright (1922) define o coeficiente de consanguinidade em termos dos genes de uma possível prole k herdados de i e j. Neste caso, f_k (ou $\Theta_{i,j}$) é a probabilidade que, em um dado lócus gênico, os dois genes sejam idênticos por descendência.

1.2.2 Modelos para estimação dos valores genéticos

Os MLGM podem ser definidos de diversas maneiras. Entretanto, no contexto de genética quantitativa, existem basicamente três tipos de especificações destes modelos que permitem a estimação dos valores genéticos em uma amostra de uma população.

Os modelos animais (*animal models*) permitem estimar os valores genéticos medidos em cada indivíduo. Por outro lado, os modelos gaméticos (*gametic models*) permitem a estimação dos valores genéticos de cada indivíduo em termos da contribuição genética de seus antecessores. A classe dos modelos gaméticos é uma generalização dos modelos *sire* e dos modelos *dam*, em que estes ignoram as contribuiçoes da mãe e do pai, respectivamente. Adicionalmente, uma terceira classe de modelos permite a estimação dos valores genéticos das unidades amostrais quando estamos interessados somente nas características herdadas da geração imediatamente anterior. Este último modelo, conhecido como modelo animal reduzido (*reduced animal model*), combina aspectos dos dois modelos previamente definidos.

Para exemplificar a definção destes modelos genéticos, considere o seguinte modelo linear misto com somente um efeito fixo, β (média populacional), e somente um efeito aleatório u_i para cada indivíduo, em que a variável resposta Y_i, do i-ésimo indivíduo, é expressa como

$$Y_i = \beta + u_i + e_i \tag{1.2.1}$$

O modelo animal

No modelo animal, u_i é o valor genético aditivo do indivíduo i e, de acordo com a equação (1.1.2), $\mathbf{x}_i^{'} = 1$, $\mathbf{z}_i^{'}$ é um vetor de ordem $1 \times n$ com entrada igual a um na i-ésima posição e igual a zero nas demais, e $\mathbf{u} = (u_1, \cdots, u_n)^{'}$.

A matriz $\mathbf{\Sigma}$ de covariâncias de $\mathbf{U}$ descreve a estrutura de covariância entre os efeitos aleatórios e é construída de acordo com a covariância genética aditiva entre dois indivíduos i e j que é dada por $2\Theta_{ij}\sigma_A^2$ (Lynch e Walsh 1998). Portanto, no modelo animal temos $\mathbf{\Sigma} = \sigma_A^2\mathbf{A}$, em que a matriz

de relação **A** possui elementos $A_{ij} = 2\Theta_{ij}$.

O modelo gamético

No modelo gamético, o valor genético aditivo de cada indivíduo é representado por meio dos valores genéticos de seus antecessores. Sejam u_{si} e u_{di} os valores genéticos paternos e maternos, respectivamente, para o i-ésimo indivíduo, então

$$u_i = \frac{1}{2}(u_{si} + u_{di}) + e_{ui} \tag{1.2.2}$$

é a média dos valores genéticos paternos e maternos acrescida de um termo de erro aleatório e_{ui} resultante da segregação Mendeliana. O modelo (1.2.1) pode ser então reescrito como

$$\begin{aligned} Y_i &= \beta + u_i + e_i \\ Y_i &= \beta + \frac{1}{2}(u_{si} + u_{di}) + (e_{ui} + e_i). \end{aligned} \tag{1.2.3}$$

Os modelos *sire* e *dam* ignoram a contribuição genética materna e paterna, respectivamente, incorporando esta no termo de erro aleatório.

A variância dos termos de erro aleatório e_{ui} é dada por

$$\mathrm{Var}(e_{ui}) = \left(1 - \frac{f_{si} + f_{di}}{2}\right)\frac{\sigma_A^2}{2} = (1 - \bar{f}_i)\frac{\sigma_A^2}{2} \tag{1.2.4}$$

em que f denota o coeficiente de consanguinidade e $\bar{f}_i$ é a consanguinidade média entre os ancestrais (Dempfle 1990). O coeficiente de consaguinidade pode ser obtido diretamente da matriz de relação **A**. Dado que $A_{ii} = 2\Theta_{ii} = 1 + f_i$ (Lynch e Walsh 1998), temos

$$\bar{f}_i = \frac{f_{si} + f_{di}}{2} = \frac{(A_{si,si} - 1) + (A_{di,di} - 1)}{2} \tag{1.2.5}$$

$$= \frac{A_{si,si} + A_{di,di}}{2} - 1 \tag{1.2.6}$$

O modelo animal reduzido

O modelo animal reduzido pode ser considerado como uma combinação do modelo animal e do modelo gamético no seguinte sentido. Suponha que l indivíduos e seus respectivos k antecessores são mensurados. No modelo animal reduzido, os antecessores são tratados segundo o modelo animal usual, isto é, $Y_i = \beta + u_i + e_i$, enquanto sua prole é tratada de acordo com o modelo gamético. Nesta classe de modelos, os indivíduos da terceira geração em diante são ignorados.

A álgebra desta classe de modelos é construída de maneira análoga aos modelos animal e gamético, somente particionando o vetor de observações e as matrizes de desenho com relação a prole e seus antecessores. O modelo animal reduzido é muitas vezes preferido segundo o ponto de vista computacional pois requer somente o cálculo de matrizes inversas da ordem do número de antecessores enquanto que o modelo animal, por exemplo, requer o cálculo de matrizes inversas da ordem do número total de observações em estudo.

1.3 Definição dos modelos utilizados

Nesta seção iremos introduzir os modelos estatísticos que serão utilizados no decorrer deste trabalho. Apresentamos a classe dos modelos lineares generalizados mistos univariados e bivariados para respostas dicotômicas. Entretanto, a extensão da metodologia apresentada para os modelos multivariados é imediata.

1.3.1 Modelos Univariados

A classe de modelos apresentada nesta seção pode ser utilizada no seguinte cenário. Uma amostra de n indivíduos de uma população é selecionada ao acaso e para cada indivíduo i, $i = 1, \ldots, n$, está associada uma variável aleatória dicotômica Y_i que representa a presença ou a ausência de determinada característica. Covariáveis associadas aos efeitos fixos e covariáveis associadas aos

efeitos aleatórios estão presentes para cada unidade amostral.

$$Y_i = \begin{cases} 1 & , \text{ se o } i\text{-ésimo indivíduo apresenta a característica,} \\ 0 & , \text{ caso contrário.} \end{cases}$$

O modelo, então, será especificado em termos de k variáveis explanatórias (efeitos fixos) e um conjunto de q variáveis explanatórias (efeitos aleatórios). De acordo com a Seção 1.1, iremos modelar uma função da esperança de Y_i, condicionalmente em $\mathbf{U} = \mathbf{u}$, utilizando a função de ligação logito, *i.e.* $g(x) = \log[x/(1-x)]$.

$$\text{logito}[\text{E}(Y_i|\mathbf{U} = \mathbf{u})] = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}. \tag{1.3.1}$$

Outras funções de ligação poderiam ser utilizadas, como a função probito ou a função log-log complementar, por exemplo.

Aqui, assumindo $Y_i|(\mathbf{U} = \mathbf{u}) \sim \text{Bernoulli}(p_i)$ com $p_i \in (0,1)$, temos que $\text{E}(Y_i|\mathbf{U} = \mathbf{u}) = p_i$ é a probabilidade que, condicional aos efeitos aleatórios, o i-ésimo indivíduo possua a característica de interesse.

Iremos assumir durante todo este trabalho que as componentes aleatórias sejam distribuídas segundo a distribuição normal multivariada com vetor de médias nulo e matriz de covariâncias $\boldsymbol{\Sigma}_1$, isto é, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$.

1.3.2 Modelos Bivariados

A classe dos modelos bivariados permite modelar duas características de interesse de forma bivariada. Esta classe de modelos permite estudar e verificar associações existentes entre duas possíveis covariáveis de interesse, por exemplo, uma vez que a estrutura bidimensional do modelo é determinada pela distribuição multivariada dos efeitos aleatórios.

A classe dos modelos bivariados pode ser utilizada no seguinte cenário. Uma amostra de n

indivíduos de uma população é selecionada ao acaso e para cada indivíduo i, $i = 1, \ldots, n$, estão associadas duas variáveis aleatórias dicotômicas $Y_{1,i}$ e $Y_{2,i}$ que representam a presença ou a ausência das características 1 e 2, respectivamente. No momento, iremos supor que, para um mesmo indivíduo, a presença de uma característica seja independente da presença da outra característica

$$Y_{j,i} = \begin{cases} 1 & , \text{ se o } i\text{-ésimo indivíduo apresenta a } j\text{-ésima característica,} \\ 0 & , \text{ caso contrário.} \end{cases}$$

com $j = \{1, 2\}$.

O modelo bivariado, então, será especificado em termos de k variáveis explanatórias (efeitos fixos) e um conjunto de q variáveis explanatórias (efeitos aleatórios) para cada uma das características de interesse. Neste contexto, sejam $\mathbf{x}_{j,i}^{'}$ o vetor de covariáveis associado ao vetor de efeitos fixos $\boldsymbol{\beta}_j$ e $\mathbf{z}_{j,i}^{'}$ o vetor de covariáveis associado ao vetor de efeitos aleatórios $\mathbf{U}_j$ para a j-ésima característica do i-ésimo indivíduo.

De acordo com a Seção 1.1, iremos modelar funções das esperanças de $Y_{1,i}|\mathbf{U}_1 = \mathbf{u}_1$ e $Y_{2,i}|\mathbf{U}_2 = \mathbf{u}_2$ utilizando a função de ligação logito, em que $g(x) = \log[x/(1-x)]$.

$$\begin{aligned} \text{logito}[\text{E}(Y_{1,i}|\mathbf{U}_1 = \mathbf{u}_1)] &= \mathbf{x}_{1,i}^{'}\boldsymbol{\beta}_1 + \mathbf{z}_{1,i}^{'}\mathbf{u}_1, \\ \text{logito}[\text{E}(Y_{2,i}|\mathbf{U}_2 = \mathbf{u}_2)] &= \mathbf{x}_{2,i}^{'}\boldsymbol{\beta}_2 + \mathbf{z}_{2,i}^{'}\mathbf{u}_2. \end{aligned} \tag{1.3.2}$$

Aqui, assumindo $Y_{j,i}|(\mathbf{U}_j = \mathbf{u}_j) \sim \text{Bernoulli}(p_{j,i})$ com $p_{j,i} \in (0, 1)$, temos que $\text{E}(Y_{j,i}|\mathbf{U}_j = \mathbf{u}_j) = p_{j,i}$ é a probabilidade que, condicional aos efeitos aleatórios $\mathbf{U}_j$, o i-ésimo indivíduo possua a j-ésima característica de interesse.

A estrutura bivariada desta classe de modelos será descrita por meio da função densidade de probabilidade atribuída ao vetor aleatório $(\mathbf{U}_1, \mathbf{U}_2)^{'}$. Neste trabalho, iremos assumir que as componentes aleatórias sejam distribuídas segundo a distribuição normal multivariada com vetor de médias nulo e matriz de covariâncias $\boldsymbol{\Sigma}_2$, isto é, $(\mathbf{U}_1, \mathbf{U}_2)^{'} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2)$.

Capítulo 2

Métodos inferenciais

A teoria de verossimilhança para MLGM se baseia basicamente na construção da função de verossimilhança marginal de $\mathbf{Y}$. Esta não é uma tarefa trivial pois, como veremos nesta seção, requer o cálculo de integrais de dimensão tão alta quanto a dimensão do vetor de efeitos aleatórios $\mathbf{U}$. De fato, a função de verossimilhança pode ser escrita da seguinte forma

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \int f_{\mathbf{Y}|\mathbf{U}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, \mathrm{d}\mathbf{u} \\ &= \int \prod_i f_{Y_i|\mathbf{U}}(y_i|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u})\, \mathrm{d}\mathbf{u}. \end{aligned} \tag{2.0.1}$$

Nos casos mais simples em que o logaritmo da função de verossimilhança pode ser expresso em termos de somas de incrementos independentes e cada incremento possui apenas integrais de baixa dimensão, métodos de integração e maximização numérica funcionam de maneira satisfatória e sua implementação é, de certa forma, simples. Entretanto, para modelos com estruturas mais complicadas (modelos com efeitos aleatórios cruzados, por exemplo), tal abordagem se mostra insatisfatória. O Capítulo 3 deste trabalho irá ilustrar, por meio de uma aplicação com dados reais, as dificuldades computacionais que podem ser encontradas no ajuste dos MLGM.

O objetivo desta seção é apresentar os aspectos inferenciais relacionados aos modelos estatísticos

considerados neste trabalho. Além disto, a metodologia desenvolvida por Breslow e Clayton (1993) é apresentada durante esta seção uma vez que esta será a técnica utilizada na estimação dos parâmetros do modelo estatístico ajustado para os dados de metritis *postpartum*. Esta metodologia se mostrou robusta e flexivel frente aos problemas encontrados na estimação dos parâmetros do problema em questão. A justificativa para a escolha da utilização desta metodologia será apresentada no capítulo reservado à análise dos dados.

2.1 Verossimilhança para o modelo univariado

Para construir a função de verossimilhança do modelos logístico univariado, devemos primeiramente encontrar a função de verossimilhança da variável aleatória $\mathbf{Y}$ condicional nos efeitos aleatórios $\mathbf{U} = \mathbf{u}$. Após isto, a função de verossimilhança marginal de $\mathbf{Y}$ será obtida por meio da integração do produto entre a verossimilhança condicional e a função densidade de probabilidade dos efeitos aleatórios. Para melhor visualização, iremos denotar a distribuição condicional $\mathbf{Y}|\mathbf{U} = \mathbf{u}$ como $\mathbf{Y}|\mathbf{u}$, em que $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ e $f_{\mathbf{U}}(\mathbf{u}) = \Phi(\mathbf{u})$.

$$P(\mathbf{Y} = \mathbf{y}) = \int P(\mathbf{Y} = \mathbf{y}|\mathbf{u}) \cdot \Phi(\mathbf{u}) \, \mathrm{d}\mathbf{u}. \tag{2.1.1}$$

Para o i-ésimo indivíduo, a probabilidade da variável aleatória Y_i assumir o valor y_i, em que $y_i = \{0, 1\}$, é dada por

$$\begin{aligned} P(Y_i = y_i|\mathbf{u}) &= p_i^{y_i} \cdot (1 - p_i)^{(1-y_i)} \\ &= \left[\frac{p_i}{1 - p_i}\right]^{y_i} (1 - p_i), \end{aligned} \tag{2.1.2}$$

em que y_i é o valor assumido por Y_i.

Dado o modelo (1.3.1), temos que p_i pode ser escrito como:

$$p_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}}. \tag{2.1.3}$$

Aplicando o logaritmo em (2.1.2) e considerando (2.1.3), temos

$$\begin{aligned}
\log[P(Y_i = y_i|\mathbf{u})] &= y_i \log\left[\frac{p_i}{1-p_i}\right] + \log[1-p_i] \\
&= y_i(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}) + \log\left[\frac{1}{1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}}\right] \\
&= y_i(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}) - \log[1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}].
\end{aligned}$$

Assumindo n observações independentes, o logaritmo da função de verossimilhança condicional é dada por

$$\begin{aligned}
\log[P(\mathbf{Y} = \mathbf{y}|\mathbf{u})] &= \log[\prod_{i=1}^{n} P(Y_i = y_i|\mathbf{u})] \\
&= \sum_{i=1}^{n} \log[P(Y_i = y_i|\mathbf{u})] \\
&= \sum_{i=1}^{n} y_i(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}) \\
&\quad - \sum_{i=1}^{n} \log\left[1 + \exp\{\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u}\}\right].
\end{aligned}$$

A função de verossimilhança marginal de $\mathbf{Y}$ pode ser então obtida.

$$\begin{aligned}
L_{\mathbf{y}}(\boldsymbol{\Sigma}_1, \boldsymbol{\beta}) &= P(\mathbf{Y} = \mathbf{y}) \\
&= \int P(\mathbf{Y} = \mathbf{y}|\mathbf{u})\Phi(\mathbf{u})\,\mathrm{d}\mathbf{u} \\
&= \int \exp\{\log[P(\mathbf{Y} = \mathbf{y}|\mathbf{u})]\}\Phi(\mathbf{u})\,\mathrm{d}\mathbf{u}.
\end{aligned} \tag{2.1.4}$$

2.2 Verossimilhança para o modelo bivariado

Para construir a função de verossimilhança dos modelos logísticos bivariados, devemos primeiramente encontrar a função de verossimilhança do vetor aleatório de respostas dos n indivíduos $\mathbf{Y}$, condicional nos efeitos aleatórios $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)^{'}$. Após isto, a função de verossimilhança marginal de $\mathbf{Y}$ será obtida por meio da integração do produto entre a verossimilhança condicional e a distribuição conjunta dos efeitos aleatórios.

Denote o vetor de resposta do i-ésimo indivíduo por $\mathbf{Y}_i = (Y_{1,i}, Y_{2,i})$, em que $Y_{1,i}$ e $Y_{2,i}$ são as duas variáveis respostas associadas às duas características de interesse do i-ésimo indivíduo e, para $j = \{1, 2\}$, $Y_{j,i}|(\mathbf{U}_j = \mathbf{u}_j) \sim$ Bernoulli$(p_{j,i})$, de acordo com a Seção 1.3.2. A extensão para o caso com mais do que duas características de interesse é imediata. Assumiremos aqui que $Y_{1,i}$ e $Y_{2,i}$ são condicionalmente independentes dado $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)^{'}$ e que as variáveis respostas para diferentes unidades amostrais são independentes. Para melhor visualização, iremos denotar a distribuição condicional $\mathbf{Y}|\mathbf{U} = \mathbf{u}$ como $\mathbf{Y}|\mathbf{u}$, em que $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)^{'} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_2)$ e $f_{\mathbf{U}}(\mathbf{u}) = \Phi(\mathbf{u})$. Assim,

$$P(\mathbf{Y} = \mathbf{y}) = \int P(\mathbf{Y} = \mathbf{y}|\mathbf{u}) \cdot \Phi(\mathbf{u}) \, \mathrm{d}\mathbf{u}.$$

Para o i-ésimo indivíduo, a probabilidade da variável aleatória $Y_{j,i}$ assumir o valor $y_{j,i}$, em que $y_{j,i} = \{0, 1\}$, com $j = \{1, 2\}$, é dada por

$$\begin{aligned} P(Y_{j,i} = y_{j,i}|\mathbf{u}) &= p_{j,i}^{y_{j,i}} \cdot (1 - p_{j,i})^{(1-y_{j,i})} \\ &= \left[\frac{p_{j,i}}{1 - p_{j,i}}\right]^{y_{j,i}} (1 - p_{j,i}), \end{aligned} \tag{2.2.1}$$

em que $y_{j,i}$ é o valor assumido por $Y_{j,i}$.

Dado o modelo (1.3.2), temos que $p_{j,i}$ pode ser escrito como:

$$p_{j,i} = \frac{\exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}}{1 + \exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}} \tag{2.2.2}$$

Aplicando o logaritmo em (2.2.1) e considerando (2.2.2), temos

$$\begin{aligned}
\log[P(Y_{j,i} = y_{j,i}|\mathbf{u})] &= y_{j,i} \log\left[\frac{p_{j,i}}{1-p_{j,i}}\right] + \log[1-p_{j,i}] \\
&= y_{j,i}(\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j) + \log\left[\frac{1}{1+\exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}}\right] \\
&= y_{j,i}(\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j) - \log[1+\exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}].
\end{aligned}$$

Assumindo n unidades amostrais independentes, o logaritmo da função de verossimilhança condicional é dada por

$$\begin{aligned}
\log[P(\mathbf{Y} = \mathbf{y}|\mathbf{u})] &= \log[\prod_{i=1}^{n} P(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{u})] \\
&= \log[\prod_{i=1}^{n} P(Y_{1,i} = y_{1,i}|\mathbf{u}) \cdot P(Y_{2,i} = y_{2,i}|\mathbf{u})] \\
&= \sum_{i=1}^{n} \log[P(Y_{1,i} = y_{1,i}|\mathbf{u}) \cdot P(Y_{2,i} = y_{2,i}|\mathbf{u})] \\
&= \sum_{i=1}^{n} \{\log[P(Y_{1,i} = y_{1,i}|\mathbf{u})] + \log[P(Y_{2,i} = y_{2,i}|\mathbf{u})]\} \\
&= \sum_{i=1}^{n} \{y_{1,i}(\mathbf{x}_{1,i}^{'}\boldsymbol{\beta}_1 + \mathbf{z}_{1,i}^{'}\mathbf{u}_1) - \log[1+\exp\{\mathbf{x}_{1,i}^{'}\boldsymbol{\beta}_1 + \mathbf{z}_{1,i}^{'}\mathbf{u}_1\}] \\
&\quad + y_{2,i}(\mathbf{x}_{2,i}^{'}\boldsymbol{\beta}_2 + \mathbf{z}_{2,i}^{'}\mathbf{u}_2) - \log[1+\exp\{\mathbf{x}_{2,i}^{'}\boldsymbol{\beta}_2 + \mathbf{z}_{2,i}^{'}\mathbf{u}_2\}]\}.
\end{aligned}$$

A função de verossimilhança marginal de $\mathbf{Y}$ pode ser então obtida.

$$\begin{aligned}
L_{\mathbf{y}}(\boldsymbol{\Sigma}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= P(\mathbf{Y} = \mathbf{y}) \\
&= \int P(\mathbf{Y} = \mathbf{y}|\mathbf{u})\Phi(\mathbf{u})\,\mathrm{d}\mathbf{u} \\
&= \int \exp\{\log[P(\mathbf{Y} = \mathbf{y}|\mathbf{u})]\}\Phi(\mathbf{u})\,\mathrm{d}\mathbf{u}. \qquad (2.2.3)
\end{aligned}$$

2.3 Quase-verossimilhança penalizada

A primeira suposição feita na construção dos MLGM é sobre a distribuição condicional dos dados. Em alguns problemas tal distribuição pode ser perfeitamente conhecida. Porém, na maioria dos casos, é impossível o conhecimento exato a priori acerca de tal distribuição. Métodos inferenciais que não necessitam de suposições específicas sobre a distribuição dos dados existem e, como mostraremos no Capítulo 3, muitas vezes solucionam os problemas encontrados no processo de estimação do método de máxima verossimilhança usual. Uma destas metodologias, o método de *quase-verossimilhança*, constrói uma pseudo verossimilhança sem a necessidade de supor uma classe de distribuições para os dados.

Nesta seção, iremos introduzir o conceito de quase-verossimilhança e apresentar a relação desta metodologia com os MLGM, introduzida por Breslow e Clayton (1993).

O conceito de quase-verossimilhança se inicia com a busca de uma quantidade que possua as mesmas características de uma função de verossimilhança, ou ainda, de sua função escore. Resultados análogos a (1.1.4) e (1.1.5) podem ser obtidos derivando o logaritmo da função de verossimilhança com relação a μ_i (1.1.2) ao invés de γ_i (1.1.1):

$$\mathrm{E}\left[\frac{\partial \log f_{Y_i|\mathbf{U}}(y_i|\mathbf{u})}{\partial \mu_i}\middle|\mathbf{u}\right] = 0 \tag{2.3.1}$$

$$\mathrm{Var}\left[\frac{\partial \log f_{Y_i|\mathbf{U}}(y_i|\mathbf{u})}{\partial \mu_i}\middle|\mathbf{u}\right] = -\mathrm{E}\left[\frac{\partial^2 \log f_{Y_i|\mathbf{U}}(y_i|\mathbf{u})}{\partial \mu_i^2}\middle|\mathbf{u}\right]. \tag{2.3.2}$$

É possível verificar que a seguinte quantidade

$$q_i = \frac{Y_i - \mu_i}{\phi a_i \upsilon(\mu_i)} \tag{2.3.3}$$

satisfaz às relações (2.3.1) e (2.3.2), em que assumimos que $\mathrm{E}(Y_i|\mathbf{u}) = \mu_i$ e $\mathrm{Var}(Y_i|\mathbf{u}) \propto \upsilon(\mu_i)$. A constante ϕa_i que aparece em (2.3.3) é somente uma constante de proporcionalidade que relaciona $\mathrm{Var}(Y_i|\mathbf{u})$ com $\upsilon(\mu_i)$, e não é necessariamente a mesma que aparece na densidade (1.1.1).

Entretanto, uma vez que ambas possuem o mesmo papel, usaremos a mesma notação.

Definimos o logaritmo da função de quase-verossimilhança por:

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi a_i \upsilon(t)} \, \mathrm{d}t, \tag{2.3.4}$$

em que, por definição, sua derivada com respeito a μ_i é igual a q_i. Note que $\mathrm{E}(Y_i|\mathbf{u}) = \mu_i$ e $\mathrm{Var}(Y_i|\mathbf{u}) \propto \upsilon(\mu_i)$ são as únicas suposições feitas até o momento. Finalmente, definimos o logaritmo da função de quase-verossimilhança correspondente à distribuição conjunta como sendo $\sum Q_i$. Para encontrar os estimadores de máxima quase-verossimilhança (MQV), devemos resolver o seguinte sistema:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum Q_i = \mathbf{0}. \tag{2.3.5}$$

Suponha, por exemplo, que estejamos interessados em assumir que a média e a variância em um determinado problema são iguais, isto é, $\upsilon(\mu_i) = \mu_i$. Note que esta suposição permite, na teoria de quase-verossimilhança, que a variância seja proporcional à media, tal que:

$$\begin{aligned} Q_i &= \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi a_i t} \, \mathrm{d}t \\ &= \frac{1}{\phi a_i} (y_i \log t - t) \Big|_{y_i}^{\mu_i} \\ &= \frac{1}{\phi a_i} (y_i \log \mu_i - \mu_i - y_i \log y_i + y_i) \end{aligned} \tag{2.3.6}$$

e as equações de MQV para $\boldsymbol{\beta}$ são:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum (y_i \log \mu_i - \mu_i) = \mathbf{0}, \tag{2.3.7}$$

eliminando os termos que não dependem de $\boldsymbol{\beta}$. Suponha agora, em um novo problema, que $Y_i \sim$ Poisson(μ_i), supondo implicitamente que $\mathrm{Var}(Y_i) = \mu_i$. Então $\log f_{Y_i}(y_i) = y_i \log \mu_i - \mu_i - \log(y_i!)$,

e as equações de máxima verossimilhança são:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum (y_i \log \mu_i - \mu_i) = \mathbf{0}, \tag{2.3.8}$$

que são exatamente as mesmas equações de MQV (2.3.7). Neste caso, os estimadores de MQV e de MV seriam exatamente iguais e a teoria de quase-verossimilhança seria completamente eficiente. Muito embora este seja um exemplo particular, a teoria de MQV possui certas vantagens sobre a teoria de MV. Na prática, frequentemente nos deparamos com situações em que a variância é maior do que a média. Se a variância é proporcional à media, então a especificação do modelo sob a teoria de quase-verossimilhança estaria correta, o que não ocorreria sob o modelo que considera a distribuição de Poisson para os dados.

McCulloch e Searle (2001) afirmam que a teoria de quase-verossimilhança é robusta em dois sentidos. Primeiro, não precisamos fazer suposições sobre a distribuição dos dados. Segundo, devemos especificar somente a relação média-variância por meio de uma constante de proporcionalidade que poderá ser estimada pelo modelo estatístico.

Breslow e Clayton (1993) utilizaram o conceito de quase-verossimilhança para estimar os parâmetros dos MLGM por meio do uso da aproximação de Laplace e da suposição que $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. A função de quase-verossimilhança integrada utilizada para estimar $(\boldsymbol{\beta}, \boldsymbol{\theta})'$, em que $\boldsymbol{\beta}$ é o vetor de parâmetros dos efeitos fixos e $\boldsymbol{\theta}$ é o vetor de parâmetros associados à matriz de covariâncias de $\mathbf{U}$, é dada por:

$$\begin{aligned} e^{ql(\boldsymbol{\beta},\boldsymbol{\theta})} &\propto |\boldsymbol{\Sigma}|^{-1/2} \int \exp\left\{ -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i) - \frac{1}{2} \mathbf{u}' \boldsymbol{\Sigma}^{-1} \mathbf{u} \right\} \mathrm{d}\mathbf{u} \\ &= C \cdot |\boldsymbol{\Sigma}|^{-1/2} \cdot \int e^{-\kappa(\mathbf{u})} \, \mathrm{d}\mathbf{u} \\ &= C \cdot |\boldsymbol{\Sigma}|^{-1/2} \cdot I, \end{aligned} \tag{2.3.9}$$

em que

$$d_i(y_i, \mu_i) = -2 \int_{y_i}^{\mu_i} \frac{y_i - t}{a_i v(t)} \, \mathrm{d}t \tag{2.3.10}$$

denota a medida *deviance* do ajuste. Se, condicionalmente em $\mathbf{u}$, as observações pertencem à família exponencial com função de variância $v(.)$, então a *deviance* do modelo é dada pela diferença $2\phi\{l(y; y, \phi) - l(y; \mu, \phi)\}$, em que $l(y; \mu, \phi)$ denota a verossimilhança condicional de y dada sua média μ.

Seja $\boldsymbol{\kappa}^{'}$ o vetor q-dimensional e $\boldsymbol{\kappa}^{''}$ a matriz $q \times q$ dimensional de derivadas parciais de primeira e segunda ordem, respectivamente, de κ com respeito a $\mathbf{u}$. Aproximando ,primeiramente, a função $\kappa(\mathbf{u})$ por meio de séries de Taylor em torno de $\tilde{\mathbf{u}} = \arg\max_{\mathbf{u}} \kappa(\mathbf{u})$, temos:

$$\kappa(\mathbf{u}) \approx \kappa(\tilde{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^{'} \boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}}) \tag{2.3.11}$$

Logo, a integral I presente na equação (2.3.9) pode ser aproximada por:

$$\begin{aligned}
I &\approx \int \exp\{-[\kappa(\tilde{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^{'} \boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}})]\} \, \mathrm{d}\mathbf{u} \\
&\approx \exp\{-\kappa(\tilde{\mathbf{u}})\} \int \exp\{-\frac{1}{2}(\mathbf{u} - \tilde{\mathbf{u}})^{'} \boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})(\mathbf{u} - \tilde{\mathbf{u}})\} \, \mathrm{d}\mathbf{u} \\
&\approx \exp\{-\kappa(\tilde{\mathbf{u}})\}(2\pi)^{q/2} \left|\boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})^{-1}\right|^{1/2} \underbrace{N_q(\tilde{\mathbf{u}}, \boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})^{-1})}_{=1} \\
&\approx \exp\{-\kappa(\tilde{\mathbf{u}})\}(2\pi)^{q/2} \left|\boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})^{-1}\right|^{1/2}.
\end{aligned} \tag{2.3.12}$$

Ignorando a constante multiplicativa C, a equação (2.3.9) pode ser, então, aproximada da seguinte maneira:

$$\begin{aligned}
e^{ql(\boldsymbol{\beta}, \boldsymbol{\theta})} &\approx C \cdot |\boldsymbol{\Sigma}|^{-1/2} \cdot \exp\{-\kappa(\tilde{\mathbf{u}})\}(2\pi)^{q/2} \left|\boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})^{-1}\right|^{1/2} \\
ql(\boldsymbol{\beta}, \boldsymbol{\theta}) &\approx -\frac{1}{2} \log |\boldsymbol{\Sigma}| - \kappa(\tilde{\mathbf{u}}) - \frac{1}{2} \log \left|\boldsymbol{\kappa}^{''}(\tilde{\mathbf{u}})\right|.
\end{aligned} \tag{2.3.13}$$

Em (2.3.13), $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ denota a solução de:

$$\begin{aligned}
\boldsymbol{\kappa}^{'}(\mathbf{u}) &= \frac{1}{2\phi}\sum_{i=1}^{n}\frac{\partial d_i}{\partial \mathbf{u}} + \frac{1}{2}2\boldsymbol{\Sigma}^{-1}\mathbf{u} = \mathbf{0} \Rightarrow \\
\boldsymbol{\kappa}^{'}(\mathbf{u}) &= \frac{1}{2\phi}\sum_{i=1}^{n} -2\frac{(y_i - \mu_i)}{a_i\upsilon(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \mathbf{u}} + \boldsymbol{\Sigma}^{-1}\mathbf{u} = \mathbf{0} \Rightarrow \\
\boldsymbol{\kappa}^{'}(\mathbf{u}) &= -\sum_{i=1}^{n}\frac{(y_i - \mu_i)\mathbf{z}_i}{\phi a_i \upsilon(\mu_i) g^{'}(\mu_i)} + \boldsymbol{\Sigma}^{-1}\mathbf{u} = \mathbf{0}, \qquad (2.3.14)
\end{aligned}$$

pois

$$\begin{aligned}
\frac{\partial \mu_i}{\partial \mathbf{u}} &= \frac{\partial \mu_i}{\partial g(\mu_i)} \cdot \frac{\partial g(\mu_i)}{\partial \mathbf{u}} \\
&= \left[\frac{\partial g(\mu_i)}{\partial \mu_i}\right]^{-1} \cdot \mathbf{z}_i \\
&= \frac{1}{g^{'}(\mu_i)} \cdot \mathbf{z}_i.
\end{aligned}$$

Diferenciando novamente com relação a $\mathbf{u}$, temos:

$$\begin{aligned}
\boldsymbol{\kappa}^{''}(\mathbf{u}) &= \frac{\partial}{\partial \mathbf{u}^{'}}\left[-\sum_{i=1}^{n}\frac{(y_i - \mu_i)\mathbf{z}_i}{\phi a_i \upsilon(\mu_i) g^{'}(\mu_i)} + \boldsymbol{\Sigma}^{-1}\mathbf{u}\right] \\
&= -\sum_{i=1}^{n}\frac{1}{\phi a_i \upsilon(\mu_i) g^{'}(\mu_i)} \cdot \frac{\partial (y_i - \mu_i)\mathbf{z}_i}{\partial \mathbf{u}^{'}} - \sum_{i=1}^{n}(y_i - \mu_i)\mathbf{z}_i \frac{\partial}{\partial \mathbf{u}^{'}}\left[\frac{1}{\phi a_i \upsilon(\mu_i) g^{'}(\mu_i)}\right] + \boldsymbol{\Sigma}^{-1} \\
&= \sum_{i=1}^{n}\frac{\mathbf{z}_i \mathbf{z}_i^{'}}{\phi a_i \upsilon(\mu_i)[g^{'}(\mu_i)]^2} + \mathbf{R} + \boldsymbol{\Sigma}^{-1} \\
&\approx \mathbf{Z}^{'}\mathbf{W}\mathbf{Z} + \boldsymbol{\Sigma}^{-1} \qquad (2.3.15)
\end{aligned}$$

em que $\mathbf{W}$ é uma matriz diagonal de ordem $n \times n$ com elementos $w_i = \left\{\phi a_i \upsilon(\mu_i)[g^{'}(\mu_i)]^2\right\}^{-1}$. O termo restante:

$$\mathbf{R} = -\sum_{i=1}^{n}(y_i - \mu_i)\mathbf{z}_i \frac{\partial}{\partial \mathbf{u}}\left[\frac{1}{\phi a_i \upsilon(\mu_i) g^{'}(\mu_i)}\right] \qquad (2.3.16)$$

possui esperança nula e é igual a zero para funções de ligações canônicas.

Combinando (2.3.13) e (2.3.15), e ignorando $\mathbf{R}$, temos:

$$ql(\boldsymbol{\beta},\boldsymbol{\theta}) \approx -\frac{1}{2}\log\left|\mathbf{I}+\mathbf{Z}'\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma}\right| - \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i,\mu_i) - \frac{1}{2}\tilde{\mathbf{u}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{u}} \tag{2.3.17}$$

em que $\tilde{\mathbf{u}}$ maximiza a soma dos últimos dois termos.

Assumindo que $\mathbf{W}$ varia pouco como função da média, ignora-se o primeiro termo de (2.3.17) e escolhe-se $\boldsymbol{\beta}$ que maximiza o segundo. Então, $(\widehat{\boldsymbol{\beta}},\hat{\mathbf{u}}) = (\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\hat{\mathbf{u}}(\boldsymbol{\theta}))$, em que $\hat{\mathbf{u}}(\boldsymbol{\theta}) = \tilde{\mathbf{u}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))$, conjuntamente maximizam a quase-verossimilhança penalizada:

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i,\mu_i) - \frac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}^{-1}\mathbf{u}. \tag{2.3.18}$$

Diferenciando com respeito a $\boldsymbol{\beta}$ e $\mathbf{u}$, obtemos o seguinte sistemas de equações para estimação dos parâmetros:

$$\begin{aligned}\sum_{i=1}^{n}\frac{(y_i-\mu_i)\mathbf{x}_i}{\phi a_i \upsilon(\mu_i)g'(\mu_i)} &= \mathbf{0},\\ \mathbf{X}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) &= \mathbf{0},\end{aligned} \tag{2.3.19}$$

em que $\boldsymbol{\Delta}$ é matrix diagonal com elementos $\left\{g'(\mu_i)\right\}$, e:

$$\begin{aligned}\sum_{i=1}^{n}\frac{(y_i-\mu_i)\mathbf{z}_i}{\phi a_i \upsilon(\mu_i)g'(\mu_i)} &= \boldsymbol{\Sigma}^{-1}\mathbf{u},\\ \mathbf{Z}'\mathbf{W}\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\mu}) &= \boldsymbol{\Sigma}^{-1}\mathbf{u}.\end{aligned} \tag{2.3.20}$$

As soluções das equações (2.3.19) e (2.3.20) podem ser obtidas por meio de um algoritmo Escore de Fisher iterativo desenvolvido por Green (1987). Considere o vetor de trabalho $\mathbf{Y}^*$ com componentes $Y_i^* = \eta_i + (y_i-\mu_i)g'(\mu_i)$, em que η_i é o preditor linear do MLGM. As soluções de

(2.3.19) e (2.3.20) podem ser obtidas por meio da solução do seguinte sistema de equações:

$$\begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{I}+\mathbf{Z}'\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{Y}^* \\ \mathbf{Z}'\mathbf{W}\mathbf{Y}^* \end{bmatrix}, \tag{2.3.21}$$

em que $\mathbf{u} = \boldsymbol{\Sigma}\boldsymbol{\nu}$.

Para o modelo linear misto normal, Harville (1977) mostrou que as soluções do sistema (2.3.21) são *BLUE* (*best linear unbiased estimator*) para $\boldsymbol{\beta}$ e $\mathbf{u}$. Para esta classe de modelos, seja lo logaritmo de sua função de verossimilhança, então:

$$\begin{aligned} l &\propto -\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu}) - \frac{1}{2}\log|\mathbf{V}|, \\ \frac{\partial l}{\partial \boldsymbol{\beta}} &= \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \end{aligned} \tag{2.3.22}$$

em que $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}'$. Equivalentemente, poderíamos primeiramente resolver (2.3.22) para $\boldsymbol{\beta}$:

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

e, então, fazer:

$$\hat{\mathbf{u}} = \boldsymbol{\Sigma}\hat{\boldsymbol{\nu}} = \boldsymbol{\Sigma}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{2.3.23}$$

Este procedimento sugere que a matriz de covariâncias para $\boldsymbol{\beta}$ possa ser aproximada por $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. De fato, $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ é a verdadeira matriz de covariâncias para o estimador de $\boldsymbol{\beta}$ sob a suposição de normalidade quando $\boldsymbol{\theta}$ é conhecido. Erros padrões para $\hat{\mathbf{u}}$ podem ser obtidos de (2.3.23). É importante notar que ambas matrizes de covariâncias ignoram a variabilidade adicional existente da necessidade em se estimar $\boldsymbol{\theta}$.

Substituindo as soluções obtidas de (2.3.18) em (2.3.17) e avaliando $\mathbf{W}$ em $(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \hat{\mathbf{u}}(\boldsymbol{\theta}))$, temos

uma quase-verossimilhança perfilada aproximada para fazer inferência sobre $\boldsymbol{\theta}$. Aproximações adicionais são feitas para motivar a estimação deste parâmetro em termos do vetor de trabalho $\mathbf{Y}^*$, a matriz de pesos $\mathbf{W}$ e as matrizes de desenho $\mathbf{X}$ e $\mathbf{Z}$. Ignorando a dependência de $\mathbf{W}$ sobre $\boldsymbol{\theta}$ e substituindo a medida *deviance* pela estatística chi-quadrado de Pearson, temos a seguinte quase-verossimilhança perfilada aproximada (a menos de uma constante de proporcionalidade):

$$ql(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{2.3.24}$$

A demostração formal destas quantidades podem ser encontradas em Harville (1977).

A forma quadrática presente em (2.3.24) contém o termo $\widehat{\boldsymbol{\beta}}$ ao invés do termo $\boldsymbol{\beta}$. Para fazer os ajustes necessários para compensar este fato, é utilizada a versão REML (*restricted estimation maximum likelihood*) de Patterson e Thompson (1971):

$$ql_1(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right| - \frac{1}{2}(\mathbf{Y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{2.3.25}$$

Tais aproximações são completamente justificadas quando $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ são ortogonais e a matriz de informação de $\widehat{\boldsymbol{\beta}}$ é dada por $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$.

Definindo $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$ e diferenciando (2.3.25) com respeito aos componentes de $\boldsymbol{\theta}$, temos o seguinte sistema de equações para estimação dos parâmetros da matriz de covariâncias

$$-\frac{1}{2}\left[(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}) - \mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\right)\right] = 0. \tag{2.3.26}$$

A correspondente matriz de informação de Fisher $\mathcal{I}$ possui componentes:

$$\mathcal{I}_{jk} = -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_k}\right). \tag{2.3.27}$$

Devido às inúmeras aproximações feitas na construção das equações de quase-verossimilhança penalizada, tal metodologia não apresenta resultados satisfatórios em determinados tipos de problemas. Breslow e Lin (1995) e Lin e Breslow (1996) mostraram que o método de quase-verossimilhança penalizada pode levar a estimadores assintoticamente viesados e, portanto, inconsistentes. Sua performance melhora à medida que a distribuição condicional de $\mathbf{Y}|\mathbf{U}$ se aproxima da distribuição normal. Entretanto, do ponto de vista prático, é preferível a aplicação de transformações nos dados para deixá-los aproximadamente normais e fazer uso da teoria de Modelo Lineares Mistos (MLM).

Recentemente, técnicas mais acuradas para a construção da teoria de quase-verossimilhança penalizada foram desenvolvidas utilizando, por exemplo, expansões de Taylor de ordem maior (Raudenbush, Yang e Yosef 2000). Entretanto, tais técnicas ainda não foram completamente testadas.

2.4 Testes para efeitos fixos

Testes de hipóteses para os efeitos fixos do modelo são realizados para verificar se uma ou mais funções dos parâmetros de interesse são iguais a uma certa constante. Consideraremos aqui somente o caso de hipóteses lineares. O modelo sujeito à hipotese nula será denotado como modelo restrito e o modelo completo será aquele em que a restrição da hipótese nula não se aplica. Neste contexto, o teste de s hipóteses lineares é da forma

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi} \quad vs. \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\xi}, \tag{2.4.1}$$

em que a matriz de coeficientes $\mathbf{C}$ tem posto completo $s \leq p$, em que p é a dimensão do vetor $\boldsymbol{\beta}$.

Neste contexto, o teste da razão de verossimilhança, o teste de Wald e o teste de escore podem ser utilizados para testar hipóteses do tipo (2.4.1). Assintoticamente e sob a hipótese nula, todos os três testes possuem distribuição Qui-Quadrado com s graus de liberdade.

O teste da razão de verossimilhança compara o logaritmo da função de verossimilhança avaliada

na estimativa de máxima verossimilhança $\widetilde{\boldsymbol{\beta}}$ obtida sob o modelo reduzido com o logaritmo da função de verossimilhança avaliada na estimativa de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$ obtida sob o modelo completo.

$$TRV = -2[l(\widetilde{\boldsymbol{\beta}}|\mathbf{y}) - l(\widehat{\boldsymbol{\beta}}|\mathbf{y})]$$

Por outro lado, o teste de Wald compara a diferença ponderada entre a estimativa de máxima verossimilhança $\mathbf{C}\widehat{\boldsymbol{\beta}}$, obtida sob o modelo completo, com o valor hipotético $\boldsymbol{\xi}$, sob a hipótese nula. A ponderação é feita por meio de $[\mathbf{C}\mathbf{F}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{C}']^{-1}$, a inversa da matriz de variâncias e covariâncias assintótica de $\mathbf{C}\widehat{\boldsymbol{\beta}}$, em que $\mathbf{F}_{\widehat{\boldsymbol{\beta}}}$ é a matriz de informação esperada de Fisher. Neste caso, a estatística do teste de Wald é dada por

$$W = [\mathbf{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}]'[\mathbf{C}\mathbf{F}_{\widehat{\boldsymbol{\beta}}}^{-1}\mathbf{C}']^{-1}[\mathbf{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}].$$

A estatística do teste de Escore é dada por

$$E = \mathbf{U}(\widetilde{\boldsymbol{\beta}})'\mathbf{F}_{\widetilde{\boldsymbol{\beta}}}^{-1}\mathbf{U}(\widetilde{\boldsymbol{\beta}}), \tag{2.4.2}$$

em que $\mathbf{U}(\widetilde{\boldsymbol{\beta}})$ é a função escore avaliada na estimativa de máxima verossimilhança obtida sob o modelo reduzido.

As estatísticas dos testes de Wald e Escore utilizam aproximações do logaritmo da função de verossimilhança via séries de Taylor de segunda ordem e, portanto, são aproximações da estatística do teste da razão de verossimilhanças. Entretanto, os testes de Wald e Escore possuem a vantagem de não precisarem do cálculo das estimativas de máxima verossimilhança obtidas sob o modelo restrito e completo, respectivamente.

Capítulo 3

Análise dos dados de metritis

Este capítulo apresenta uma aplicação em dados reais da teoria dos Modelos Lineares Generalizados Mistos (MLGM) desenvolvida durante este trabalho. Parte dos resultados apresentados aqui são baseados no trabalho desenvolvido por Labouriau et al. 2014, cujo principal objetivo foi detectar a presença de mecanismos genéticos associados à susceptibilidade ou resistência à metritis *postpartum* em vacas leiteiras sujeitas a sistemas de produção de grande porte.

Metritis é uma inflamação encontrada na parede do útero das vacas leiteiras e é causada, em geral, pelas bactérias *Escherichia coli, Trueperella pyogenes* e *Fusobacterium necrophorum*, podendo também ser causada pelo vírus BoHV-4 (Sheldon et al. 2009). Quando esta doença é diagnosticada após a parição, sua denominação é metritis *postpartum*.

Neste estudo, a avaliação da metritis foi feita de maneira uniforme em todas as fazendas e foi realizada segundo os padrões determinados pela legislação dinamarquesa. Tal avaliação foi feita utilizando uma escala de pontos de 0 a 9, de acordo com o grau de severidade da doença, em que a categoria 0 representa a ausência de sintomas da doença e a categoria 9 representa a doença em seu nível mais severo (Tabela 1, Elkjær et al. 2013).

Adicionalmente, as ferramentas estatísticas consideradas neste trabalho permitiram verificar se os programas de seleção genética aos quais a população de animais em produção está sujeita

contribuem para a melhora ou piora da resistência à metritis *postpartum*. A classe de modelos considerada, os MLGM, é relativamente bem estudada em sistemas de pequeno porte, porém esta abordagem é inovadora neste cenário. Extensões multivariadas também foram avaliadas neste estudo, isto é, extensões que permitem a análise simultânea do valor genético em várias características simultaneamente. Procedimentos de validação cruzada foram realizados com o intuito de verificar a consistência do processo de estimação dos modelos ajustados.

Tais extensões foram fundamentais para avaliar o impacto de programas de seleção genética na resistência a doenças. Estudos similares podem eventualmente ser desenvolvidos com dados brasileiros no futuro próximo.

3.1 Motivação

Registros detalhados de doenças são mantidos rotineiramente em sistemas de produção animal, como os registros de gado leiteiro Dinamarquês e os registros de produção de gado leiteiro compilados pela Embrapa no Brasil, por exemplo. Tais registros frequentemente contêm informações sobre a ocorrência de doenças, como mastitis e metritis, por exemplo, bem como informações genéticas, como o *pedigree* ou mesmo marcadores genéticos de DNA, para um grande número de animais. Isto permite detectar e caracterizar possíveis componentes genéticas relacionadas à resistência ou susceptibilidade à uma série de doenças. Esta caracterização não é uma tarefa trivial pois envolve o uso de modelos estatísticos complexos e o manuseio de conjuntos de dados de grande porte (centenas de milhares de observações, pedigrees profundos e centenas de milhares de marcadores de DNA). O desenvolvimento de ferramentas estatísticas adequadas e eficientes para este tipo de estudo é uma área de pesquisa em franco desenvolvimento.

Um exemplo concreto do tipo de problema descrito acima é o estudo desenvolvido por Elkjær *et al.* (2013) sobre a ocorrência de metritis em gado leiteiro na Dinamarca. Neste estudo, dados de 282.099 vacas sujeitas a exames rotineiros de metritis *postpartum* foram utilizados para mostrar que

a ocorrência desta doença tem efeitos deletérios na fertilidade destes animais, o que tem consequências econômicas e biológicas significativas. Estes dados estão disponíveis para pesquisa no Centro de Extensão Agrícola (*Knowledge Center for Agriculture*), Skejby, Dinamarca. A abrangência e excelente qualidade destes dados faz com que este problema seja ideal para desenvolver e testar novos procedimentos estatísticos para estudos de mecanismos genéticos ligados à susceptibilidade ou resistência de doenças.

Um outro aspecto importante é que populações de animais domésticos em produção estão sujeitas a uma forte pressão de seleção. Existe atualmente a preocupação de que esta seleção, tipicamente voltada diretamente ao aumento da produtividade comercial, possa estar deteriorando a resistência dos animais a uma série de doenças, o que toca na questão do bem-estar dos animais em produção, além de ter consequências econômicas claras. O tipo de modelo utilizado neste trabalho permitiu avaliar e quantificar a ocorrência desta possível degenerescência genética.

Os modelos estatísticos considerados permitem a modelagem simultânea de diversas características de interesse, como a incidência de doenças, produção, crescimento, tempo de vida produtiva com ou sem censura, por exemplo, usando-se diferentes tipos de distribuições de probabilidade para cada característica. Os tipos de modelos considerados permitem também a modelagem de efeitos genéticos aditivos, que são transmitidos regularmente para as próximas gerações, e efeitos genéticos passageiros, como efeitos de consanguinidade, mistura de raças e heterosis, por exemplo. Estes modelos foram implementados de forma eficiente no software DMU (`http://www.dmu.agrsci.dk/`), construído na Universidade de Aarhus, o que permitiu a análise imediata proposta. Entretanto, diversos avanços teóricos e computacionais no processo de inferência destes modelos ainda são necessários.

3.2 Análise descritiva

Para este trabalho, foram utilizados registros de 832.124 parições de 472.290 vacas leiteiras dinamarquesas da raça Holstein de 1.760 fazendas, durante os anos de 2006 a 2013 do registro de gado dinamarquês. Todas as fazendas incluídas neste estudo realizaram avaliações mensais e participaram do programa de avaliação dinamarquês "NySR". Neste programa, todas as vacas são examinadas por um veterinário uma vez no período de cinco a vinte dias após cada parição e, portanto, nosso objeto de estudo será a metritis *postpartum*.

O tempo mediano do exame para este conjunto de dados foi igual a 9 dias após a parição (média 9.8 dias e desvio padrão 3.9 dias). Todo o diagnóstico de metritis foi feito utilizando a mesma escala de avaliação recomendada pela legislação dinamarquesa em todas as fazendas. Tal escala assume valores discretos de 0 a 9 e considera a quantidade, cor e cheiro da secreção uterina retirada manualmente da vaca (Elkjær et al. 2013).

A Tabela 3.1 apresenta uma descrição geral de todas as variáveis disponíveis e utilizadas durante a análise dos dados de metritis.

Tabela 3.1: Descrição das variáveis

Variável	Descrição
Vaca	Número de identificação da vaca (unidade amostral)
Touro	Número de identificação do touro (genitor)
Fazenda	Número de identificação da fazenda
Parição	Número da parição
Ano	Ano de diagnóstico da metritis
Estação	Estação do ano de diagnóstico da metritis
Metritis	Nível da metritis

No conjunto de dados analisado, unidades amostrais foram observadas ao longo do tempo e registros foram feitos sempre com até vinte dias após a respectiva parição. É importante ressaltar a existência de vacas que entraram no estudo ao longo do tempo. As primeiras observações destes animais não são necessariamente referentes à primeira parição. Consideraremos aqui somente observações de parições menores ou iguais a 9 (Figura 3.1). Parições maiores que 9 são raras e

consideradas atípicas.

Uma característica interessante dos dados é que o nível médio de metritis na população é decrescente ao longo dos anos (Coeficiente de correlação de Spearman: $\hat{\rho} = -0.98$, p-valor = 0.0004), como podemos observar através da Figura 3.2. Este fato leva a questionamentos sobre a causa deste comportamento. O melhoramento genético voltado para a produção econômica poderia estar causando um efeito positivo na resistência à metritis ou também a utilização de novos tratamentos para o controle de outras doenças como a mastite, por exemplo, poderia estar contribuindo indiretamente para a resistência à metritis.

Outra característica importante neste conjunto de dados é a assimetria presente na escala de diagnóstico de metritis. Como podemos observar na Tabela 3.2, a maior parte das observações foram diagnosticadas sem a presença de qualquer sintoma de metritis ou mesmo com um grau de severidade baixo desta doença. Consideraremos a presença de metritis em uma observação de uma unidade amostral se esta foi diagnosticada com nível maior ou igual a 4. Conhecimentos empíricos sugerem que se uma vaca é diagnosticada com nível de metritis maior ou igual a 4, esta apresenta padrões de reprodução anormais e/ou deficientes (Elkjær et al. 2013). Utilizando esta definição operacional de metritis, observamos 167.027 casos desta doença, levando a um percentual de prevalência geral de 20.1%. Pontos de corte adicionais irão ser utilizados para caracterizar a metritis segundo os diferentes níveis de severidade.

Tabela 3.2: Distribuição das observações de acordo com o nível de metritis

Nível	# Obs.	%
0	290.622	34,9%
1	148.321	17,8%
2	133.195	16,0%
3	92.959	11,2%
4	54.403	6,5%
5	55.396	6,7%
6	22.826	2,7%
7	18.153	2,2%
8	9.088	1,1%
9	7.161	0,9%

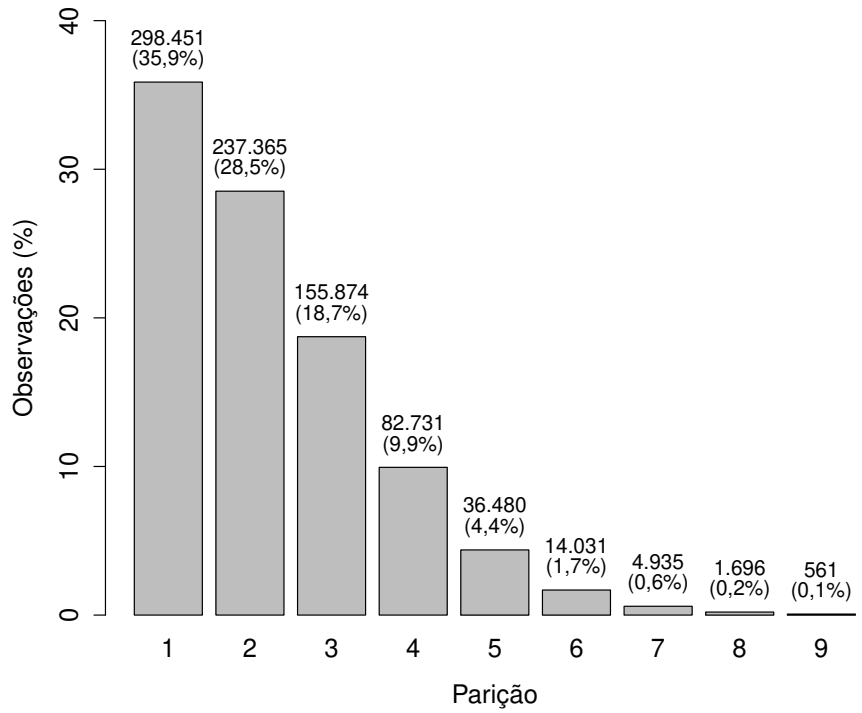


Figura 3.1: Distribuição das observações de acordo com a parição.

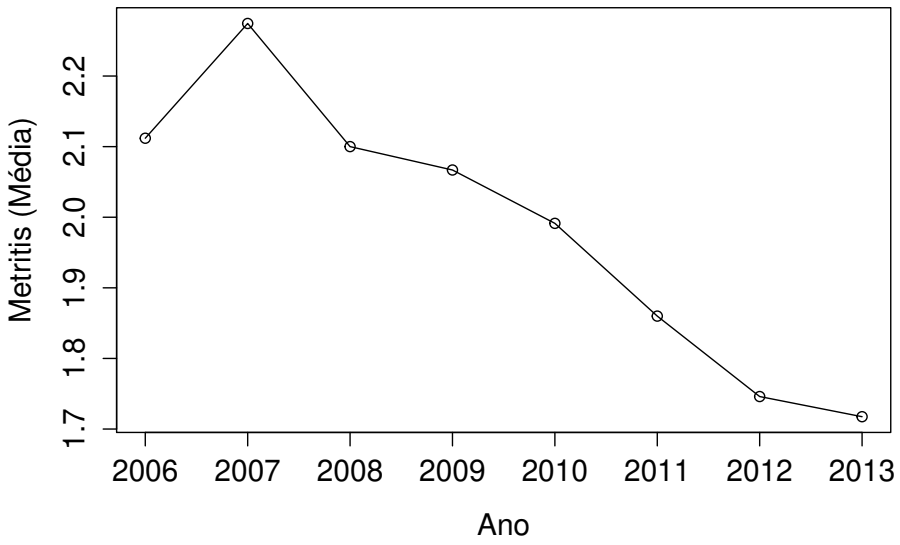


Figura 3.2: Nível médio amostral de metritis de acordo com o ano.

O número de genitores (touros) presentes na base de dados analisada é igual a 8.846 e, para quantificar as relações genéticas destes animais, foi criada uma base de dados de *pedigree* referente a três gerações anteriores destes touros, resultando em uma base de *pedigree* contendo 45.603 animais. Por meio destes dados, foi construída toda a relação genética entre as unidades amostrais em estudo.

3.3 Metodologia

Nesta seção iremos detalhar os modelos estatísticos utilizados na análise dos dados de metritis *postpartum*, apresentando a teoria dos MLGM desenvolvida neste trabalho dentro do contexto de genética quantitativa e melhoramento animal.

Primeiramente, é importante ressaltar que todos os modelos estatísticos utilizados aqui são modelos do tipo *sire*, isto é, modelos que permitem a estimação dos valores genéticos de cada vaca em termos da contribuição genética do touro. Neste tipo de modelo, vacas filhas de um mesmo touro compartilham entre si determinada estrutura de correlação. A necessidade de verificar e quantificar a existência de mecanismos genéticos aditivos relacionados com o diagnóstico da doença justifica a escolha de tal modelo. De acordo com a seção 1.2.2, o efeito genético paterno é referenciado no modelo estatístico por meio da inclusão de efeitos aleatórios. Neste sentido, para cada touro presente na base de dados está associado um, e somente um, elemento no vetor de efeitos aleatórios relacionados às componentes genéticas. Assim, o vetor aleatório terá dimensão igual ao número de touros presentes no conjunto de dados e unidades amostrais referentes a estes animais compartilharão o mesmo elemento do vetor aleatório.

3.3.1 Modelo univariado

De acordo com a seção 1.3.1 deste trabalho, para cada unidade amostral iremos modelar a variável resposta dicotômica Y_i que representa a ocorrência ou não ocorrência de metritis *postpartum*

na i-ésima vaca. Informações genéticas na forma de *pedigrees* profundos e covariáveis explanatórias estão presentes para cada unidade amostral.

O interesse em aplicar a classe de modelos a ser descrita nesta seção é modelar a presença/ausência de metritis em uma amostra de uma população de vacas leiteiras dinamarquesas e com este modelo conseguir descrever e quantificar estruturas genéticas associadas à susceptibilidade de tal doença. A presença de uma estrutura genética poderia, por exemplo, responder algumas perguntas naturais que surgem no contexto de genética quantitativa e que são de interesse imediato. Um questionamento natural que surge neste tipo de problema é sobre a existência de variabilidade genética aditiva na probabilidade de ocorrência desta doença e, em caso afirmativo, qual a magnitude relativa desta variação.

A presença ou ausência de metritis pode ser operacionalmente mensurada por meio de uma variável aleatória discreta assumindo valores em $\{0, 1\}$:

$$Y_i = \begin{cases} 1 & \text{, se a } i\text{-ésima vaca possui a doença,} \\ 0 & \text{, caso contrário.} \end{cases}$$

O modelo univariado ajustado ao conjunto de dados foi especificado em termos de um conjunto de cováriaveis explanatórias, associadas aos efeitos fixos, e um conjunto de componentes aleatórias, associadas aos efeitos aleatórios.

As componentes aleatórias utilizadas no modelo estatístico e denotadas por $\mathbf{U}$ e $\mathbf{V}$, representam, respectivamente, os efeitos genéticos aditivos determinados por meio do *pedigree* dos animais em estudo, e os efeitos ambientais. Pela facilidade de notação, iremos descrever o modelo em função de somente duas componentes aleatórias. Entretanto, todos os modelos estatísticos ajustados incluem exatamente três componentes aleatórias: uma componente aleatória associada aos fatores genéticos (*pedigree*) e duas componentes aleatórias associadas aos fatores ambientais (uma componente representando a fazenda na qual o animal foi diagnosticado e outra componente representando uma interação entre a fazenda, ano e estação do diagnóstico). A generalização teórica

para este caso é imediata. Neste contexto, vacas filhas de um mesmo touro e vacas que foram diagnosticadas no mesmo ambiente compartilham determinada estrutura de correlação que será especificada adiante.

Iremos assumir aqui que as componentes aleatórias associadas aos efeitos genéticos e as componentes aleatórias associadas aos efeitos ambientais são independentes. Esta restrição pode ser eliminada se quisermos descrever possíveis interações entre efeitos genéticos e ambientais, porém este problema não será estudado neste trabalho.

De acordo com o modelo geral descrito na seção 1.3.1, iremos modelar uma função da esperança de $Y_i|(\mathbf{U} = \mathbf{u}, \mathbf{V} = \mathbf{v})$, através de

$$\text{logito}(p_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} + \mathbf{z}_i^{'}\mathbf{u} + \mathbf{w}_i^{'}\mathbf{v}. \tag{3.3.1}$$

Aqui, a quantidade p_i é a probabilidade de ocorrência da doença na i-ésima vaca de acordo com o modelo $Y_i|(\mathbf{U} = \mathbf{u}, \mathbf{V} = \mathbf{v}) \sim \text{Bernoulli}(p_i)$, em que $p_i \in (0, 1)$. Na equação (3.3.1), o parâmetro $\boldsymbol{\beta} \in \mathbb{R}^k$ (finito dimensional) representa os efeitos fixos, $\mathbf{x}_i^{'}, \mathbf{z}_i^{'}$ e $\mathbf{w}_i^{'}$ são vetores de desenho para os efeitos fixos, genéticos e ambientais, respectivamente, da i-ésima unidade amostral.

A definição da estrutura de covariância das componentes aleatórias é parte crucial no desenvolvimento do modelo estatístico e no contexto geral de modelos lineares generalizados mistos, pois é por meio dela que iremos estudar as componentes genéticas e ambientais do nosso problema. Assumiremos aqui que as componentes aleatórias sejam distribuídas segundo a distribuição normal multivariada, *i.e.*, $(\mathbf{U}, \mathbf{V})^{'} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ em que

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_u^2\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \sigma_v^2\mathbf{I} \end{bmatrix}. \tag{3.3.2}$$

Na equação (3.3.2), $\mathbf{I}$ é a matriz identidade e $\mathbf{A}$ é a matriz de relação genética de ordem igual ao número de touros presentes no conjunto de dados, neste caso de dimensão 8.846×8.846. Cada

entrada da matriz **A** representa a proporção esperada de genes compartilhados entre dois touros da amostra em estudo e é construída de acordo com o grau de relação r entre eles. $(A_{ii}) = 1$ para todo $i = 1, \ldots, 8.846$ e $(A_{ij}) = 0$ se o par de indivíduos (i, j) não possui qualquer tipo de relação. Se o grau de relação entre dois indivíduos $(i, j), i \neq j$, é r, então $(A_{ij}) = (1/2)^r$. Parentes de primeiro grau possuem grau de relação r igual a um, parentes de segundo grau possuem grau r de relação igual a dois, e assim sucessivamente.

3.3.2 Modelo bivariado (severidade)

Nesta seção, iremos utilizar a teoria dos MLGM para modelar, de forma bivariada, o grau de severiade de metritis *postpartum* em cada unidade amostral. Neste contexto, questionamentos adicionais podem ser feitos com relação aos mecanismos genéticos e ambientais associados aos diferentes níveis de manifestação da doença em estudo. A classe de modelos bivariados permite verificar e quantificar marginalmente a variação genética aditiva e ambiental presente em cada nível da doença. Além disso, é possível verificar a existência de mecanismos genéticos/ambientais comuns e mecanismos genéticos/ambientais independentes em cada uma das dimensões e quantificar suas respectivas ordens de magnitude.

Para construir a teoria de verossimilhança deste tipo de modelo por meio da utilização do MLGM com respostas dicotômicas, primeiramente note que para cada unidade amostral está associada uma variável resposta multinomial, que será condicionada nos efeitos aleatórios, representando os níveis de metritis *postpartum*, que iremos agrupar em três categorias distintas. Seja Y_i uma variável aleatória quantitativa discreta representando os níveis de metritis da i-ésima unidade amostral em três categorias.

$$Y_i = \begin{cases} 0 & , \text{se nível} = \{0, 1, 2, 3\}, \\ 1 & , \text{se nível} = \{4, 5\}, \\ 2 & , \text{se nível} = \{6, 7, 8, 9\}. \end{cases}$$

Neste tipo de modelo, para cada observação definimos duas variáveis aleatórias $Y_{1,i}$ e $Y_{2,i}$ que representam a ocorrência ou não ocorrência de metritis *postpartum* em um grau não muito forte, com classificação empírica igual a 4 ou 5, e a ocorrência ou não ocorrência da mesma doença em um grau mais elevado, com classificação empírica igual ou maior que 6.

A variável aleatória Y_i pode ser representada por meio do vetor bidimensional de variáveis aleatórias $(Y_{1,i}, Y_{2,i})$, em que $Y_{1,i} = \mathbb{1}(Y_i = 1)$ e $Y_{2,i} = \mathbb{1}(Y_i = 2)$. Tal vetor bidimensional assume valores em $\{(0,0),(0,1),(1,0)\}$ e sua distribuição de probabilidades pode ser representada através de uma tabela de probabilidades (Tabela 3.3).

Tabela 3.3: Tabela de probabilidades - Modelo bivariado (severidade)

$Y_{1,i} \backslash Y_{2,i}$	0	1	
0	p_{00i}	p_{01i}	$(p_{00i} + p_{01i})$
1	p_{10i}	0	p_{10i}
	$(p_{00i} + p_{10i})$	p_{01i}	

É possível verificar que $Y_{1,i}|(Y_{2,i} = 0) \sim$ Bernoulli$(p_{1,i})$ e que $Y_{2,i}|(Y_{1,i} = 0) \sim$ Bernoulli$(p_{2,i})$, em que $p_{1,i} = \frac{p_{10i}}{p_{00i}+p_{10i}}$ e $p_{2,i} = \frac{p_{01i}}{p_{00i}+p_{01i}}$.

$$
\begin{aligned}
P(Y_{1,i} = 0|Y_{2,i} = 0) &= \frac{P(Y_{1,i} = 0, Y_{2,i} = 0)}{P(Y_{2,i} = 0)} = \frac{p_{00i}}{p_{00i} + p_{10i}} \\
P(Y_{1,i} = 1|Y_{2,i} = 0) &= \frac{P(Y_{1,i} = 1, Y_{2,i} = 0)}{P(Y_{2,i} = 0)} = \frac{p_{10i}}{p_{00i} + p_{10i}} \\
P(Y_{2,i} = 0|Y_{1,i} = 0) &= \frac{P(Y_{2,i} = 0, Y_{1,i} = 0)}{P(Y_{1,i} = 0)} = \frac{p_{00i}}{p_{00i} + p_{01i}} \\
P(Y_{2,i} = 1|Y_{1,i} = 0) &= \frac{P(Y_{2,i} = 1, Y_{1,i} = 0)}{P(Y_{1,i} = 0)} = \frac{p_{01i}}{p_{00i} + p_{01i}}
\end{aligned}
$$

Além disso, $Y_{1,i}|(Y_{2,i} = 1)$ e $Y_{2,i}|(Y_{1,i} = 1)$ possuem distribuições degeneradas em 0.

$$
\begin{aligned}
P(Y_{1,i} = 0|Y_{2,i} = 1) &= \frac{P(Y_{1,i} = 0, Y_{2,i} = 1)}{P(Y_{2,i} = 1)} = \frac{p_{01i}}{p_{01i}} = 1 \\
P(Y_{2,i} = 0|Y_{1,i} = 1) &= \frac{P(Y_{2,i} = 0, Y_{1,i} = 1)}{P(Y_{1,i} = 1)} = \frac{p_{10i}}{p_{10i}} = 1
\end{aligned}
$$

O modelo bivariado para diferentes níveis de metritis foi especificado em termos de um conjunto de covariáveis explanatórias, associadas aos efeitos fixos, e um conjunto de componentes aleatórias, associadas aos efeitos aleatórios.

As componentes aleatórias utilizadas no modelo estatísticos são denotadas por $\mathbf{U}_1$ e $\mathbf{U}_2$, representando os efeitos genéticos aditivos determinados pelos touros em cada uma das duas dimensões do modelo estatístico, e por $\mathbf{V}_1$ e $\mathbf{V}_2$, representando os efeitos ambientais em cada uma das duas dimensões. Similarmente ao modelo univariado, todos os modelos ajustados incluem três componentes aleatórias (neste caso para cada uma das dimensões): uma componente aleatória associada aos fatores genéticos (*pedigree*) e duas componentes aleatórias associadas aos fatores ambientais (uma componente representando a fazenda na qual o animal foi diagnosticado e outra componente representando uma interação entre a fazenda, ano e estação do diagnóstico). Entretanto, para facilitar a notação, iremos especificar o modelo como função de somente duas componentes. A generalização teórica é imediata. Note que, muito embora as componentes aleatórias associadas aos efeitos genéticos e as componentes aleatórias associadas aos efeitos ambientais estejam presentes em ambas as dimensões do modelo, a construção do modelo permite que seus efeitos sejam distintos em cada uma das dimensões. Esta é uma característica importante desta classe de modelos, pois permite avaliar e quantificar a existência de mecanismos genéticos e ambientais distintos em diferentes níveis da doença.

Para a construção do modelo estatístico, todas as distribuições de probabilidade são condicionadas aos efeitos aleatórios.

$$
\begin{aligned}
Y_i|(\mathbf{U}_1 = \mathbf{u}_1, \mathbf{U}_2 = \mathbf{u}_2, \mathbf{V}_1 = \mathbf{v}_1, \mathbf{V}_2 = \mathbf{v}_2) &\sim \text{Multinomial}(1, p_{00i}, p_{10i}, p_{01i}) \\
Y_{1,i}|(Y_{2,i} = 0, \mathbf{U}_1 = \mathbf{u}_1, \mathbf{V}_1 = \mathbf{v}_1) &\sim \text{Bernoulli}(p_{1,i}) \\
Y_{2,i}|(Y_{1,i} = 0, \mathbf{U}_2 = \mathbf{u}_2, \mathbf{V}_2 = \mathbf{v}_2) &\sim \text{Bernoulli}(p_{2,i})
\end{aligned}
$$

Assumiremos aqui que as componentes associadas aos efeitos genéticos e as componentes as-

sociadas aos efeitos ambientais são independentes. Esta restrição pode ser eliminada se quisermos descrever possíveis interações entre efeitos genéticos e ambientais, porém este problema não será estudado neste trabalho.

De acordo com o modelo geral descrito na seção 1.3.2, iremos modelar funções da esperança de $Y_{1,i}|(Y_{2,i}=0)$ e $Y_{2,i}|(Y_{1,i}=0)$, condicionalmente a $\mathbf{U}_1=\mathbf{u}_1$, $\mathbf{U}_2=\mathbf{u}_2$, $\mathbf{V}_1=\mathbf{v}_1$ e $\mathbf{V}_2=\mathbf{v}_2$.

$$\begin{aligned}\text{logito}(p_{1,i}) &= \mathbf{x}_{1,i}^{'}\boldsymbol{\beta}_1+\mathbf{z}_{1,i}^{'}\mathbf{u}_1+\mathbf{w}_{1,i}^{'}\mathbf{v}_1,\\ \text{logito}(p_{2,i}) &= \mathbf{x}_{2,i}^{'}\boldsymbol{\beta}_2+\mathbf{z}_{2,i}^{'}\mathbf{u}_2+\mathbf{w}_{2,i}^{'}\mathbf{v}_2.\end{aligned} \tag{3.3.3}$$

Na equação (3.3.3), os parâmetros $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathbb{R}^k$ (finito dimensionais) representam os efeitos fixos, $\mathbf{x}_{j,i}^{'},\mathbf{z}_{j,i}^{'}$ e $\mathbf{w}_{j,i}^{'}$, para $j=\{1,2\}$ e $i=\{1,\ldots,n\}$, são vetores de desenho para os efeitos fixos, genéticos e ambientais, respectivamente, para a j-ésima característica da i-ésima unidade amostral.

Note que as equações (3.3.3) modelam simultaneamente as probabilidades dos níveis $Y_i=1$ e $Y_i=2$ com relação a categoria de referência $Y_i=0$.

$$\begin{aligned}\text{logito}(p_{1,i}) &= \log\left[\frac{p_{10i}/(p_{00i}+p_{10i})}{1-p_{10i}/(p_{00i}+p_{10i})}\right]=\log\left[\frac{p_{10i}}{p_{00i}}\right],\\ \text{logito}(p_{2,i}) &= \log\left[\frac{p_{01i}/(p_{00i}+p_{01i})}{1-p_{01i}/(p_{00i}+p_{01i})}\right]=\log\left[\frac{p_{01i}}{p_{00i}}\right].\end{aligned} \tag{3.3.4}$$

A definição da estrutura de covariância das componentes aleatórias é parte crucial no desenvolvimento do modelo e no contexto geral de modelos lineares generalizados mistos, pois é por meio dela que iremos estudar as componentes genéticas e ambientais do nosso problema. Assumiremos aqui que as componentes aleatórias sejam distribuídas segundo a distribuição normal multivariada, *i.e.* $(\mathbf{U}_1,\mathbf{U}_2,\mathbf{V}_1,\mathbf{V}_2)^{'}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_2)$ em que

$$\boldsymbol{\Sigma}_2=\begin{bmatrix}\boldsymbol{\Sigma}_u\otimes\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_v\otimes\mathbf{I}\end{bmatrix}. \tag{3.3.5}$$

Na equação (3.3.5), $\mathbf{I}$ é a matriz identidade e $\mathbf{A}$ é a matriz de relação, de ordem igual ao número de touros presentes no conjunto de dados, e é construída de maneira análoga ao caso univariado. A estrutura de covariância entre as componentes aleatórias associadas aos fatores genéticos é dada por

$$\boldsymbol{\Sigma}_u = \begin{bmatrix} \sigma^2_{u,1} & \sigma_{u,12} \\ \sigma_{u,12} & \sigma^2_{u,2} \end{bmatrix}. \tag{3.3.6}$$

A correlação genética entre dois níveis da doença é dada por $\rho_g = \sigma_{u,12}/(\sigma_{u,1} \cdot \sigma_{u,2})$. Finalmente, a estrutura de covariância entre as componentes aleatórias associadas aos fatores ambientais é dada por

$$\boldsymbol{\Sigma}_v = \begin{bmatrix} \sigma^2_{v,1} & \sigma_{v,12} \\ \sigma_{v,12} & \sigma^2_{v,2} \end{bmatrix} \tag{3.3.7}$$

Para construir a função de verossimilhança do modelo logístico bivariado para diferentes níveis de metritis devemos primeiramente encontrar a função de verossimilhança da variável aleatória $\mathbf{Y} = (Y_1, \ldots, Y_n)$ condicional nos efeitos aleatórios. Após isto, a função de verossimilhança marginal de $\mathbf{Y}$ será obtida por meio da integração do produto entre a verossimilhança condicional e a função densidade de probabilidade dos efeitos aleatórios. Para melhor visualização, iremos denotar a distribuição condicional de $\mathbf{Y}$ dadas as realizações dos efeitos aleatórios $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2)'$ simplesmente como $\mathbf{Y}|\mathbf{u}$, em que $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}_1, \mathbf{V}_2)' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2)$ e $f_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{V}_1,\mathbf{V}_2}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{v}_1, \mathbf{v}_2) = \Phi(\mathbf{u})$. Logo,

$$P(\mathbf{Y} = \mathbf{y}) = \int P(\mathbf{Y} = \mathbf{y}|\mathbf{u}) \cdot \Phi(\mathbf{u}) \, \mathrm{d}\mathbf{u}. \tag{3.3.8}$$

Para a i-ésima unidade amostral, a probabilidade da variável aleatória Y_i assumir o valor y_i,

em que $y_i = \{0, 1, 2\}$, é dada por

$$P(Y_i = y_i|\mathbf{u}) = p_{10i}^{y_{1i}} \cdot p_{01i}^{y_{2i}} \cdot p_{00i}^{(1-y_{1i}-y_{2i})}, \tag{3.3.9}$$

em que y_{1i} e y_{2i} são os valores assumidos por Y_{1i} e Y_{2i}, respectivamente.

Note que, dada a restrição $p_{10i} + p_{01i} + p_{00i} = 1$ e o modelo (3.3.3), p_{00i} pode ser escrito como

$$p_{00i} = \frac{1}{1 + \sum_{j=1}^{2} \exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}}. \tag{3.3.10}$$

Aplicando o logaritmo em (3.3.9) e considerando (3.3.10), temos

$$\begin{aligned}
\log[P(Y_i = y_i|\mathbf{u})] &= y_{1i}\log[p_{10i}] + y_{2i}\log[p_{01i}] + (1 - y_{1i} - y_{2i})\log[p_{00i}] \\
&= y_{1i}\log\left[\frac{p_{10i}}{1 - p_{10i} - p_{01i}}\right] + y_{2i}\log\left[\frac{p_{01i}}{1 - p_{10i} - p_{01i}}\right] \\
&\quad + \log[1 - p_{10i} - p_{01i}] \\
&= y_{1i}[\mathbf{x}_{1,i}^{'}\boldsymbol{\beta}_1 + \mathbf{z}_{1,i}^{'}\mathbf{u}_1] \\
&\quad + y_{2i}[\mathbf{x}_{2,i}^{'}\boldsymbol{\beta}_2 + \mathbf{z}_{2,i}^{'}\mathbf{u}_2] \\
&\quad - \log\left[1 + \sum_{j=1}^{2} \exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}\right].
\end{aligned}$$

Assumindo n observações independentes, o logaritmo da função de verossimilhança condicional

é dado por

$$\begin{aligned}
\log[P(\mathbf{Y}=\mathbf{y}|\mathbf{u})] &= \log[\prod_{i=1}^{n} P(Y_i=y_i|\mathbf{u})] \\
&= \sum_{i=1}^{n} \log[P(Y_i=y_i|\mathbf{u})] \\
&= \sum_{i=1}^{n} y_{1i}[\mathbf{x}_{1,i}^{'}\boldsymbol{\beta}_1 + \mathbf{z}_{1,i}^{'}\mathbf{u}_1] \\
&\quad + \sum_{i=1}^{n} y_{2i}[\mathbf{x}_{2,i}^{'}\boldsymbol{\beta}_2 + \mathbf{z}_{2,i}^{'}\mathbf{u}_2] \\
&\quad - \sum_{i=1}^{n} \log\left[1 + \sum_{j=1}^{2} \exp\{\mathbf{x}_{j,i}^{'}\boldsymbol{\beta}_j + \mathbf{z}_{j,i}^{'}\mathbf{u}_j\}\right].
\end{aligned}$$

A função de verossimilhança marginal de $\mathbf{Y}$ pode ser, então, obtida por

$$\begin{aligned}
L_{\mathbf{y}}(\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_v, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= P(\mathbf{Y}=\mathbf{y}) \\
&= \int P(\mathbf{Y}=\mathbf{y}|\mathbf{u})\Phi(\mathbf{u})\,\mathrm{d}\mathbf{u} \\
&= \int \exp\{\log[P(\mathbf{Y}=\mathbf{y}|\mathbf{u})]\}\Phi(\mathbf{u})\,\mathrm{d}\mathbf{u}. \qquad (3.3.11)
\end{aligned}$$

3.3.3 Modelo bivariado - parições

Com o intuito de verificar e quantificar os possíveis mecanismos genéticos associados à presença de metritis *postpartum* em cada uma das parições, foram ajustados modelos bivariados em que a primeira dimensão do modelo estatístico considera somente as observações referentes à primeira parição e a segunda dimensão do modelo considera observações de todas as demais parições.

3.4 Resultados dos ajustes

Nesta seção iremos apresentar os principais resultados obtidos nas etapas de modelagem dos dados de metritis.

Em um primeiro momento foram ajustados modelos univariados e modelos bivariados somente para as observações relativas à primeira parição dos animais. Nesta etapa, foi possível verificar e quantificar o efeito genético existente na presença da doença nos animais e também verificar a possível existência de diferentes mecanismos genéticos associados aos diferentes níveis da doença.

Em um segundo momento, considerando todas as observações, foram ajustados dois modelos estatísticos bivariados. Nesta etapa, consideramos os diferentes níveis de metritis, categorizados em nível 'baixo' e nível 'alto', e as diferentes parições, categorizadas em 'parição 1' e 'parições ≥ 2', como dimensões do modelo estatístico. Neste sentido, além de responder questionamentos sobre os mecanismos genéticos atuantes nos diferentes níveis de severidade da doença, foi possível também responder questionamentos sobre os mecanismos genéticos atuantes nas diferentes parições dos animais. É importante notar que tal modelo é, de fato, condicional às parições dos animais.

Por fim, foi ajustado um modelo bivariado para os diferentes níveis da doença considerando somente as observações referentes às parições diferentes das primeiras.

Todos os modelos foram ajustados utilizando o *software* DMU, desenvolvido na Universidade de Aarhus, cujo processo de estimação dos parâmetros é realizado utilizando a teoria de quase verossimilhança penalizada desenvolvida por Breslow e Clayton (1993). O parâmetro de dispersão associado a componente de erro aleatória será denotado por ϕ, de acordo com a teoria desenvolvida neste trabalho.

É importante ressaltar que todos os modelos ajustados contêm exatamente três componentes associadas aos efeitos aleatórios: uma componente genética que contém informação genética dos touros por meio do pedigree, de acordo com o modelo *sire*, uma componente ambiental considerando o efeito da fazenda na qual o animal foi diagnosticado, e outra componente ambiental formada pela interação entre a fazenda, ano e estação no qual o animal foi diagnosticado.

Todas as componentes aleatórias foram consideradas, por construção, independentes entre si. Além disso, o efeito do ano e da estação também foram incluídos como efeitos fixos em todos os modelos ajustados. O efeito da parição foi considerado como fixo nos modelos e nas dimensões

que consideram as observações de diferentes parições. O efeito aleatório relacionado à fazenda foi considerado como sendo o mesmo em ambas as dimensões de todos os modelos bivariados ajustados. Durante a apresentação dos resultados, iremos denotar as componentes de variância dos três efeitos aleatórios como σ_g^2 (*genetic*), σ_h^2 (*herd*) e σ_{hys}^2 (*herd,year,season*).

O principal objetivo nestas análises foi verificar e quantificar a existência de componentes genéticas associadas à ocorrência de metritis nos animais em estudo. Este fato pode ser verificado por meio do estudo das componentes de variâncias relacionadas aos efeitos aleatórios genéticos, nos modelos univariado e bivariado, e da componente de correlação genética, nos modelos bivariados.

Muito embora o efeito da parição e do ano/estação estejam incluídos como efeitos fixos nos modelos ajustados, o interesse nos valores e interpretações destes parâmetros é secundário. A significância de tais parâmetros pode levar a questionamentos interessantes do ponto de vista ambiental, por exemplo. Entretanto, tais variáveis foram consideradas no modelo com o intuito de controlar seus respectivos efeitos. Devido à complexidade dos modelos ajustados, as interpretações das razões de chances são feitas de forma aproximada, uma vez que as covariáveis ano e estação, associadas aos efeitos fixos, também entram como componentes na interação entre fazenda, ano e estação associada à componente aleatória.

3.4.1 Primeira parição

Modelo univariado

A Tabela 3.4 apresenta os resultados das estimativas dos parâmetros do modelo univariado que considera somente observações da primeira parição.

Como podemos observar, a estimativa da variância associada à componente genética do touro, $\hat{\sigma}_g^2$, é positiva e este fato sugere a existência de fatores genéticos associados à susceptibilidade do desenvolvimento de metritis *postpartum*.

As estimativas das componentes de variância associadas à fazenda e à interação entre fazenda,

Tabela 3.4: Resultados do modelo univariado - primeira parição

Parâmetro	Estimativa	Erro padrão
σ_g^2	0.07019	0.00612
σ_h^2	0.97055	0.04035
σ_{hys}^2	0.19763	0.00659
ϕ	0.90289	0.00241
Ano: 2006	-1.14263	0.07928
Ano: 2007	-1.13086	0.07171
Ano: 2008	-1.24888	0.07033
Ano: 2009	-1.35974	0.06935
Ano: 2010	-1.44769	0.06847
Ano: 2011	-1.61270	0.06819
Ano: 2012	-1.75899	0.06823
Ano: 2013	-1.78750	0.06935
Estação: Dez/Fev	0.18494	0.01668
Estação: Mar/Mai	0.16824	0.01714
Estação: Jun/Ago	0.07733	0.01696

"Estação: Set/Nov" como referência

Tempo de execução: 1h40min

ano e estação indicam que fatores ambientais exercem influência no desenvolvimento e diagnóstico de metritis *postpartum*.

As estimativas dos parâmetros relacionados à covariável Ano são decrescentes, com exceção de 2007, e negativas, indicando que a chance de diagnóstico de metritis *postpartum* ao longo dos anos é, em média, menor ano após ano, mantendo os outros níveis constantes. A razão de chance de diagnóstico desta doença entre os anos 2013 e 2006 é aproximadamente $\exp(-1.78750+1.14263) \approx 0.52$. Este fato indica, por exemplo, que a chance de diagnóstico é, em média, aproximadamente 48% menor quando comparamos os anos de 2013 e 2006, levando a questionamentos sobre a causa desta melhora.

Por outro lado, quando analisamos o efeito da estação temos que a razão de chance de diagnóstico de metritis entre o inverno (Dez/Fev) e o verão (Jun/Ago) é aproximadamente $\exp(0.18494 - 0.07733) \approx 1.11$. Este fato indica que a chance de diagnóstico no período de inverno é, em média, aproximadamente 11% maior do que a chance de diagnóstico no verão.

Modelo bivariado (severidade)

A Tabela 3.5 apresenta os resultados das estimativas dos parâmetros do modelo bivariado (severidade) que considera somente observações da primeira parição. Neste modelo, questionamentos sobre a existência de componentes genéticas, distintas e comuns, associadas à presença de diferentes níveis da doença podem ser respondidos.

De acordo com a Tabela 3.5, a estimativa da variância associada a componente genética do touro, $\hat{\sigma}_g^2$, é positiva e este fato sugere a existência de fatores genéticos associados à susceptibilidade do desenvolvimento de metritis *postpartum* em cada uma das dimensões do modelo. Entretanto, a alta correlação estimada $(0.9654, \text{E.P.} = 0.0121)$ revela essencialmente que os mesmos mecanismos genéticos atuam nos diferentes níveis desta doença, aqui categorizados em baixo (4-5) e alto (6-9).

Tabela 3.5: Resultados do modelo bivariado (severidade) - primeira parição

	Nível baixo: 4-5				Nível alto: 6-9	
Parâmetros	Estimativa	E.P.	Correlação	E.P.	Estimativa	E.P.
σ_g^2	0.09190	0.00703	0.96540	0.01210	0.23998	0.01556
σ_h^2	1.04211	0.04281	–	–	1.04211	0.04281
σ_{hys}^2	0.28834	0.00948	0.37340	0.01890	0.76557	0.01791
ϕ	0.87545	0.00244	–	–	0.67305	0.00195
Ano: 2006	-1.68639	0.08374	–	–	-2.17058	0.11958
Ano: 2007	-1.68810	0.07391	–	–	-2.17353	0.10873
Ano: 2008	-1.77767	0.07212	–	–	-2.33776	0.10687
Ano: 2009	-1.85776	0.07095	–	–	-2.51640	0.10557
Ano: 2010	-1.93892	0.06984	–	–	-2.57676	0.10420
Ano: 2011	-2.05807	0.06953	–	–	-2.81746	0.10391
Ano: 2012	-2.16792	0.06961	–	–	-3.02931	0.10411
Ano: 2013	-2.18424	0.07100	–	–	-3.05809	0.10585
Estação: Dez/Fev	0.17412	0.01976	–	–	0.20603	0.02689
Estação: Mar/Mai	0.18211	0.02019	–	–	0.14066	0.02784
Estação: Jun/Ago	0.07550	0.02009	–	–	0.08691	0.02743

"Estação: Set/Nov" como referência

Tempo de execução: 23h36min

As estimativas dos parâmetros de variância associados aos efeitos da fazenda e da interação

entre fazenda, ano e estação, sugerem a existência de mecanismos ambientais atuantes na susceptibilidade da doença, tanto para o nível baixo ($\hat{\sigma}^2_h = 1.04211, \text{E.P.}(\hat{\sigma}^2_h) = 0.04281$ e $\hat{\sigma}^2_{hys} = 0.28834, \text{E.P.}(\hat{\sigma}^2_{hys}) = 0.00948$) quanto para o nível alto ($\hat{\sigma}^2_h = 1.04211, \text{E.P.}(\hat{\sigma}^2_h) = 0.04281$ e $\hat{\sigma}^2_{hys} = 0.76557, \text{E.P.}(\hat{\sigma}^2_{hys}) = 0.01791$). A correlação estimada entre as duas dimensões do modelo para a interação entre fazenda, ano e estação (0.3734, E.P. = 0.0189) sugere a existência de mecanismos ambientais comuns e mecanismos ambientais independentes atuantes em cada um dos níveis de severidade da doença.

Assim como no modelo univariado, as estimativas dos parâmetros associados ao efeito fixo do Ano foram negativas. Comparando, por exemplo, os anos de 2013 e 2006, temos que a razão de chance de desenvolvimento de metritis *postpartum* no nível baixo é aproximadamente $\exp(-2.18424+1.68639) \approx 0.61$ e no nível alto é aproximadamente $\exp(-3.05809+2.17058) \approx 0.41$, indicando que a chance de diagnóstico desta doença é aproximadamente 39% e 59% menor, respectivamente para os níveis baixo e alto, de um ano para o outro.

O efeito da estação também pode ser verificado no modelo bivariado de acordo com a severidade. Neste caso, a razão de chance de diagnóstico de metritis em um nível baixo é aproximadamente $\exp(0.17412 - 0.07550) \approx 1.10$ e em nível alto é aproximadamente $\exp(0.20603 - 0.08691) \approx 1.12$, comparando o inverno (Dez/Fev) com o verão (Jun/Ago). Este resultado indica que a chance de diagnóstico desta doença é, em média, 10% e 12% maior no inverno, para os níveis baixo e alto, respectivamente.

3.4.2 Todas as parições

Modelo bivariado (parições)

O modelo bivariado para as parições considerou observações da primeira parição como a primeira dimensão do modelo estatístico e observações das parições subsequentes como a segunda dimensão. Por meio da análise deste modelo, é possível verificar e quantificar a existência de

componentes genéticas distintas atuantes nas diferentes parições dos animais.

A Tabela 3.6 apresenta os principais resultados do ajuste.

Aqui, as estimativas dos parâmetros das componentes de variâncias associadas aos fatores genéticos sugerem a existência de mecanismos atuantes no desenvolvimento de metritis *postpartum* durante a primeira parição e durante as parições subsequentes. A correlação genética estimada entre a componente genética do touro para a primeira parição e a componente genética do touro para as demais parições foi igual a 0.6737 (E.P. = 0.0437), indicando a presença de mecanismos genéticos comuns e mecanismos genéticos independentes para ambas as dimensões do modelo.

As estimativas dos parâmetros das componentes de variâncias associadas aos fatores ambientais, aqui representados pela fazenda e pela interação entre fazenda, ano e estação, sugerem a existência de mecanismos ambientais atuantes no desenvolvimento da doença em estudo para ambas as dimensões do modelo estatístico. Unidades amostrais pertencentes ao mesmo ambiente respondem homogeneamene quanto à susceptibilidade da doença, fato caracterizado pelo alto valor estimado das componentes de variância associadas a estes efeitos, quando comparados com $\hat{\sigma}_g^2$ e $\hat{\phi}$.

Comparando os anos de 2013 e 2006, observamos que a razão de chance de diagnóstico da doença é aproximadamente $\exp(-1.73072 + 1.11000) \approx 0.53$, para a primeira parição, e aproximadamente $\exp(-1.39992 + 0.86681) \approx 0.58$, para as parições subsequentes, indicando que a chance de diagnóstico de metritis *postpartum* é, em média, 47% e 42% menor em 2013, para as respectivas dimensões do modelo estatístico, mantendo todos os outros níveis constantes.

Aqui, o efeito da estação do ano difere em ambas as dimensões do modelo ajustado. A razão de chances de diagnóstico de metritis entre o inverno e o verão é aproximadamente $\exp(0.18524 - 0.08086) \approx 1.11$ e aproximadamente $\exp(0.07457 - 0.07756) \approx 0.99$, para as respectivas dimensões do modelo, indicando que, em média, a chance de diagnóstico é 11% maior no inverno, considerando a primeira parição, e 1% maior no verão, considerando as demais parições, mantendo todos os outros níveis constantes.

Tabela 3.6: Resultados do modelo bivariado (parições) - todas as parições

	Parição 1				Parição ≥ 2	
Parâmetros	Estimativa	E.P.	Correlação	E.P.	Estimativa	E.P.
σ^2_g	0.07484	0.00606	0.67370	0.04370	0.07194	0.00526
σ^2_h	0.96000	0.03700	–	–	0.96000	0.03700
σ^2_{hys}	0.22187	0.00688	0.61760	0.01810	0.17944	0.00488
ϕ	0.90677	0.00242	–	–	0.92799	0.00183
Ano: 2006	-1.11000	0.07690	–	–	-0.86681	0.12615
Ano: 2007	-1.08131	0.06957	–	–	-0.78418	0.12259
Ano: 2008	-1.20858	0.06822	–	–	-0.91396	0.12189
Ano: 2009	-1.31754	0.06728	–	–	-1.07991	0.12127
Ano: 2010	-1.40043	0.06646	–	–	-1.16302	0.12076
Ano: 2011	-1.55656	0.06620	–	–	-1.27171	0.12061
Ano: 2012	-1.70432	0.06625	–	–	-1.39764	0.12055
Ano: 2013	-1.73072	0.06739	–	–	-1.39992	0.12076
Estação: Dez/Fev	0.18524	0.01685	–	–	0.07457	0.01395
Estação: Mar/Mai	0.16855	0.01728	–	–	0.08192	0.01425
Estação: Jun/Ago	0.08086	0.01712	–	–	0.07756	0.01376
Parição: 2	–	–	–	–	-0.63301	0.10353
Parição: 3	–	–	–	–	-0.49720	0.10346
Parição: 4	–	–	–	–	-0.39570	0.10351
Parição: 5	–	–	–	–	-0.34402	0.10383
Parição: 6	–	–	–	–	-0.26566	0.10492
Parição: 7	–	–	–	–	-0.17313	0.10842
Parição: 8	–	–	–	–	-0.23613	0.11931

"Estação: Set/Nov"e "‘Parição: 9"como referência

Tempo de execução: 20h28min

Modelo bivariado (severidade)

A Tabela 3.7 apresenta os resultados das estimativas dos parâmetros do modelo bivariado que considera níveis distintos da doença e observações de todas as parições. Neste modelo, o efeito da parição é controlado por meio da inclusão desta covariável como um efeito fixo.

As estimativas dos parâmetros de variância associados às componentes genéticas sugerem a existência de mecanismos genéticos que influenciam a susceptibilidade de metritis nos animais observados. Este mecanismo, entretato, aparenta ser o mesmo para ambos os níveis da doença, uma vez que a estimativa da correlação genética é alta e próxima de 1 $(0.9929, \text{E.P.} = 0.0038)$.

Tabela 3.7: Resultados do modelo bivariado (severidade) - todas as parições

	Nível baixo: 4-5				Nível alto: 6-9	
Parâmetros	Estimativa	E.P.	Correlação	E.P.	Estimativa	E.P.
σ^2_g	0.07403	0.00419	0.99290	0.00380	0.15896	0.00837
σ^2_h	0.98241	0.03767	–	–	0.98241	0.03767
σ^2_{hys}	0.21644	0.00511	0.22200	0.01460	0.57484	0.01097
ϕ	0.91735	0.00149	–	–	0.77447	0.00131
Ano: 2006	0.45276	0.03816	–	–	0.72567	0.05216
Ano: 2007	0.47840	0.02584	–	–	0.84710	0.03579
Ano: 2008	0.38354	0.02313	–	–	0.64446	0.03268
Ano: 2009	0.24224	0.02112	–	–	0.46432	0.03014
Ano: 2010	0.16791	0.01896	–	–	0.37907	0.02733
Ano: 2011	0.08270	0.01778	–	–	0.18823	0.02604
Ano: 2012	-0.00281	0.01738	–	–	-0.02112	0.02593
Estação: Dez/Fev	0.11205	0.01392	–	–	0.11915	0.02010
Estação: Mar/Mai	0.12298	0.01422	–	–	0.09104	0.02070
Estação: Jun/Ago	0.08373	0.01390	–	–	0.06566	0.02022
Parição: 1	-2.08654	0.05684	–	–	-2.90661	0.07841
Parição: 2	-2.38370	0.05690	–	–	-3.34116	0.07851
Parição: 3	-2.27247	0.05705	–	–	-3.16487	0.07865
Parição: 4	-2.17640	0.05752	–	–	-3.05505	0.07920
Parição: 5	-2.12763	0.05880	–	–	-3.01393	0.08069
Parição: 6	-2.04483	0.06205	–	–	-2.94201	0.08441
Parição: 7	-1.95604	0.07059	–	–	-2.82851	0.09370
Parição: 8	-1.96297	0.09086	–	–	-3.02802	0.12244
Parição: 9	-1.91993	0.13831	–	–	-2.48155	0.15579

"Ano: 2013"e "Estação: Set/Nov"como referência

Tempo de execução: 60h42min

As estimativas dos parâmetros de variância associados aos efeitos da fazenda e da interação entre fazenda, ano e estação, sugerem a existência de mecanismos ambientais atuantes na susceptibilidade da doença, tanto para o nível baixo ($\hat{\sigma}^2_h = 0.98241, \text{E.P.}(\hat{\sigma}^2_h) = 0.03767$ e $\hat{\sigma}^2_{hys} = 0.21644, \text{E.P.}(\hat{\sigma}^2_{hys}) = 0.00511$) quanto para o nível alto ($\hat{\sigma}^2_h = 0.98241, \text{E.P.}(\hat{\sigma}^2_h) = 0.03767$ e $\hat{\sigma}^2_{hys} = 0.57484, \text{E.P.}(\hat{\sigma}^2_{hys}) = 0.01097$).

Comparando os anos de 2013 e 2006, a razão de chances de diagnóstigo de metritis é aproximadamente $\exp(-0.45276) \approx 0.63$, para o nível baixo, e aproximadamente $\exp(-0.72567) \approx 0.48$, para o nível alto. Este fato indica que a chance de diagnóstico da doença é, em média, 37% e 52%

menor para os níveis baixo e alto, respectivamente, comparando o ano de 2013 com o ano de 2006, mantendo todos os outros níveis constantes.

A interpretação para o efeito da estação é similar aos outros modelos ajustados: a chance de diagnóstico de metritis é, em média, maior no inverno quando comparado com os meses de verão. Aqui, a razão de chances é aproximadamente $\exp(0.11205 - 0.08373) \approx 1.03$ para o nível baixo e aproximadamente $\exp(0.11915 - 0.06566) \approx 1.05$ para o nível alto, indicando que, em média, a chance de diagnóstico da doença é 3% e 5% maior no inverno do que no verão, respectivamente para os níveis baixo e alto, mantendo todos os outros níveis constantes.

Comparando as parições 9 e 1, a razão de chances de diagnóstico de metritis é aproximadamente $\exp(-1.91993 + 2.08654) \approx 1.18$ para o nível baixo, indicando que, em média, a chance de diagnóstico desta doença é aproximadamente 18% maior para a nona parição, mantendo todos os outros níveis constantes. Por outro lado, a razão de chances de diagnóstico de metritis é aproximadamente $\exp(-2.48155 + 2.90661) \approx 1.53$ para o nível alto, indicando que, em média, a chance de diagnóstico desta doença é aproximadamente 53% maior para a nona parição, mantendo todos os outros níveis constantes. Algumas possíveis explicações para este fato são que os animais são naturalmente e artificialmente selecionados ao longo de suas vidas e que componentes genéticas, comuns e independentes, estão associadas à susceptibilidade da doença para as diferentes parições.

Modelo bivariado (severidade) - parições ≥ 2

Com o intuito de controlar os diferentes mecanismos genéticos atuantes na susceptibilidade de metritis *postpartum* nas diferentes parições dos animais, foi ajustado um modelo bivariado de acordo com o grau de severidade da doença somente para as observações referentes às parições diferentes das primeiras. A Tabela 3.8 apresenta os resultados das estimativas dos parâmetros.

As estimativas dos parâmetros de variância associados as componentes genéticas sugerem a existência de mecanismos genéticos que influenciam a susceptibilidade de metritis nos animais observados em ambas as dimensões do modelo. Este mecanismo, entretato, aparenta ser o mesmo

Tabela 3.8: Resultados do modelo bivariado (severidade) - parições ≥ 2

	Nível baixo: 4-5				Nível alto: 6-9	
Parâmetros	Estimativa	E.P.	Correlação	E.P.	Estimativa	E.P.
σ^2_g	0.08890	0.00565	0.99470	0.00390	0.21420	0.01252
σ^2_h	0.98221	0.03862	–	–	0.98221	0.03862
σ^2_{hys}	0.23489	0.00662	0.29980	0.01690	0.64760	0.01381
ϕ	0.89701	0.00183	–	–	0.72224	0.00153
Ano: 2006	0.44190	0.04492	–	–	0.61782	0.06155
Ano: 2007	0.47382	0.03009	–	–	0.81383	0.04141
Ano: 2008	0.38307	0.02686	–	–	0.60407	0.03774
Ano: 2009	0.20822	0.02449	–	–	0.44853	0.03458
Ano: 2010	0.14705	0.02191	–	–	0.35084	0.03129
Ano: 2011	0.07794	0.02049	–	–	0.19657	0.02965
Ano: 2012	0.00296	0.02011	–	–	-0.02516	0.02964
Estação: Dez/Fev	0.07677	0.01622	–	–	0.06516	0.02309
Estação: Mar/Mai	0.08654	0.01654	–	–	0.06341	0.02366
Estação: Jun/Ago	0.08147	0.01598	–	–	0.04671	0.02294
Parição: 2	-2.33766	0.06428	–	–	-3.28383	0.09400
Parição: 3	-2.22438	0.06434	–	–	-3.10581	0.09401
Parição: 4	-2.12638	0.06471	–	–	-2.99589	0.09440
Parição: 5	-2.07511	0.06580	–	–	-2.95084	0.09557
Parição: 6	-1.98917	0.06868	–	–	-2.88265	0.09858
Parição: 7	-1.90559	0.07639	–	–	-2.77025	0.10632
Parição: 8	-1.90211	0.09518	–	–	-2.95982	0.13120
Parição: 9	-1.86251	0.14047	–	–	-2.39316	0.16071

"Ano: 2013"e "Estação: Set/Nov"como referência

Tempo de execução: 73h28min

para ambos os níveis da doença, uma vez que a estimativa da correlação genética é alta e próxima de 1 ($0.9947, \text{E.P.} = 0.0039$).

As estimativas dos parâmetros de variância associados aos efeitos da fazenda e da interação entre fazenda, ano e estação, sugerem a existência de mecanismos ambientais atuantes na susceptibilidade da doença, tanto para o nível baixo ($\hat{\sigma}^2_h = 0.98221, \text{E.P.}(\hat{\sigma}^2_h) = 0.03862$ e $\hat{\sigma}^2_{hys} = 0.23489, \text{E.P.}(\hat{\sigma}^2_{hys}) = 0.00662$) quanto para o nível alto ($\hat{\sigma}^2_h = 0.98221, \text{E.P.}(\hat{\sigma}^2_h) = 0.03862$ e $\hat{\sigma}^2_{hys} = 0.64760, \text{E.P.}(\hat{\sigma}^2_{hys}) = 0.01381$).

Comparando os anos de 2013 e 2006, a razão de chances de diagnóstigo de metritis é aproxi-

madamente $\exp(-0.44190) \approx 0.64$, para o nível baixo, e aproximadamente $\exp(-0.61782) \approx 0.53$, para o nível alto. Este fato indica que a chance de diagnóstico da doença é, em média, 36% e 47% menor para os níveis baixo e alto, respectivamente, comparando o ano de 2013 com o ano de 2006, mantendo todos os outros níveis constantes.

Comparando as parições 9 e 2, a razão de chances de diagnóstico de metritis é aproximadamente $\exp(-1.86251 + 2.33766) \approx 1.62$ para o nível baixo, indicando que, em média, a chance de diagnóstico desta doença é aproximadamente 62% maior para a nona parição, mantendo todos os outros níveis constantes. Por outro lado, a razão de chances de diagnóstico de metritis é aproximadamente $\exp(-2.39316 + 3.28383) \approx 2.43$ para o nível alto, indicando que, em média, a chance de diagnóstico desta doença é aproximadamente 143% maior para a nona parição, mantendo todos os outros níveis constantes.

3.5 Tendência genética

Um questionamento natural que surge no contexto de genética quantitativa é se os programas de melhoramento animal influenciam, de maneira direta ou indireta, na susceptibilidade do desenvolvimento de doenças. O tipo de influência exercida por estes programas pode trazer consequências econômicas claras.

Nesta seção, apresentaremos uma breve análise da tendência genética, ou *genetic trend*, para todos os modelos ajustados. A tendência genética é uma medida de controle populacional utilizada para verificar, em nosso contexto, a influência genética do touro ao longo do tempo na susceptibilidade no desenvolvimento de metritis *postpartum*. Neste caso, utilizaremos como medida de controle a média dos valores preditos das componentes aleatórias de todos os touros nascidos em determinado ano.

Seja $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_q)'$ o conjunto dos valores preditos dos efeitos aleatórios resultantes do processo de estimação (2.3.21), em que cada elemento do vetor representa o valor genético

estimado de cada um dos touros. No contexto de genética quantitativa, este valor é chamado de valor genético estimado (*estimated breeding value*), e denotaremos aqui simplesmente pela sigla EBV. Considere $\{\hat{\mathbf{u}}_t\}$ o conjunto de todos os EBV's dos touros que nasceram em determinado ano t. A tendência genética é construída de tal forma que a média dos EBV's é calculada para cada ano t e, por meio da análise gráfica destes valores, é possível verificar a melhora ou piora do efeito genético associado à susceptibilidade da doença.

Dado o modelo estatístico (3.3.1), por exemplo, em que p_i é a probabilidade de desenvolvimento da doença, a uma possível tendência positiva dos EBV's médios ao longo dos anos, associa-se uma piora do ponto de vista genético pois a chance de desenvolvimento de metritis está, em média, cada vez maior. Por outro lado, uma tendência negativa dos EBV's médios ao longo dos anos sugere uma melhora quanto ao diagnóstico da doença.

No conjunto de dados analisado, temos observações provenientes de inseminações artificiais de touros nascidos, por exemplo, antes de 1980. Entretanto, devido à pequena quantidade de observações, os gráficos apresentados aqui apresentam as médias dos EBV's dos touros nascidos a partir de 1996.

3.5.1 Primeira parição

As Figuras 3.3 e 3.4 apresentam os gráficos de tendência genética dos modelos univariado e bivariado que consideram somente observações da primeira parição.

Analisando a Figura 3.3 podemos observar uma ligeira deterioração genética no período anterior a 2002 e uma acentuada melhora entre os anos de 2003 a 2009, cujo coeficiente de correlação de Spearman entre o EBV médio e o ano foi igual a $\hat{\rho} = -0.97$ (p-valor = 0.0004). Este fato indica basicamente que o programa de melhoramento genético no qual os animais em estudo estão inseridos causou uma diminuição na susceptibilidade do desenvolvimento de metritis *postpartum* após o ano de 2002. Questionamentos adicionais podem ser feitos para descobrir as causas da mudança de comportamento da curva de tendência genética para os períodos pré e pós 2002.

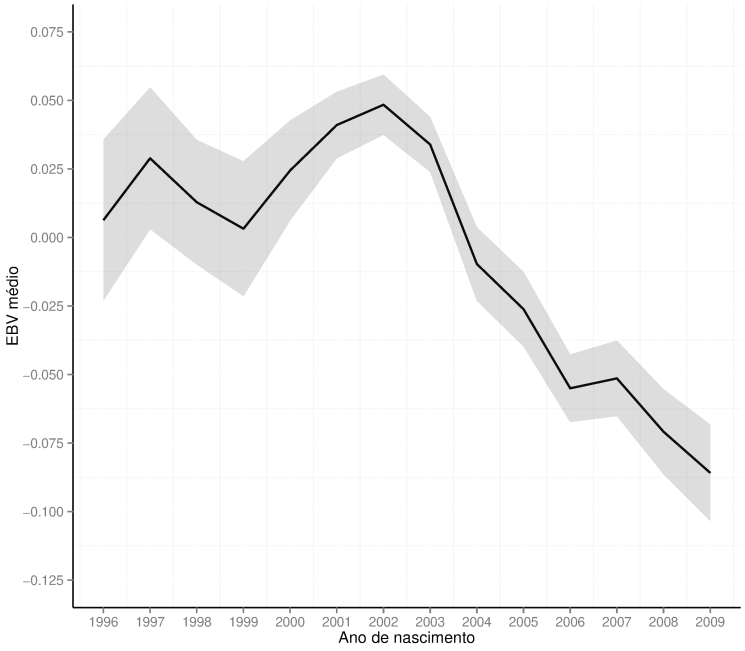


Figura 3.3: EBV's médios e IC's(95%) para média, modelo univariado - primeira parição.

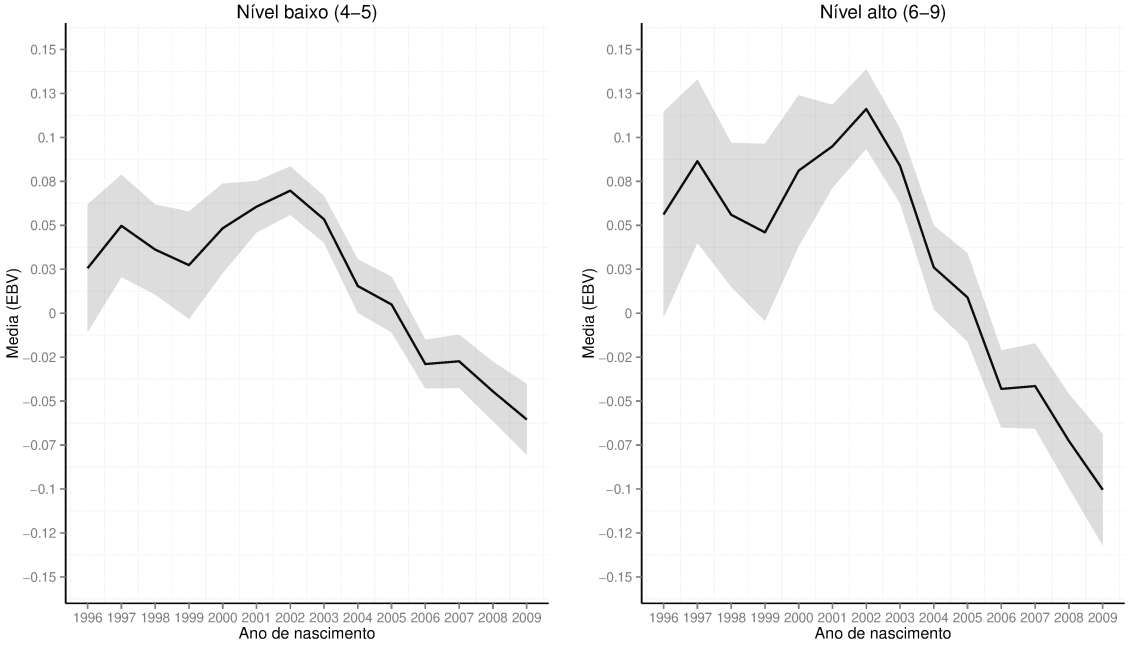


Figura 3.4: EBV's médios e IC's(95%) para média, modelo bivariado (severidade) - primeira parição.

De maneira análoga, a Figura 3.4 apresenta o gráfico de tendência genética com comportamento similiar ao apresentado na Figura 3.3, porém aqui com a curva graficada para ambas as dimensões do modelo bivariado. É possível observar uma melhora acentuada do ponto de vista genético após o ano de 2002, cujo coeficiente de correlação de Spearman estimado foi $\hat{\rho} = -0.97$ (p-valor $= 0.0004$), idêntico ao modelo univariado, para ambas as dimensões do modelo.

3.5.2 Todas as parições

As Figuras 3.5, 3.6 e 3.7 apresentam os gráficos de tendência genética para o modelo bivariado (parições), modelo bivariado (severidade) e o modelo bivariado (severidade) que não considera observações da primeira parição, respectivamente.

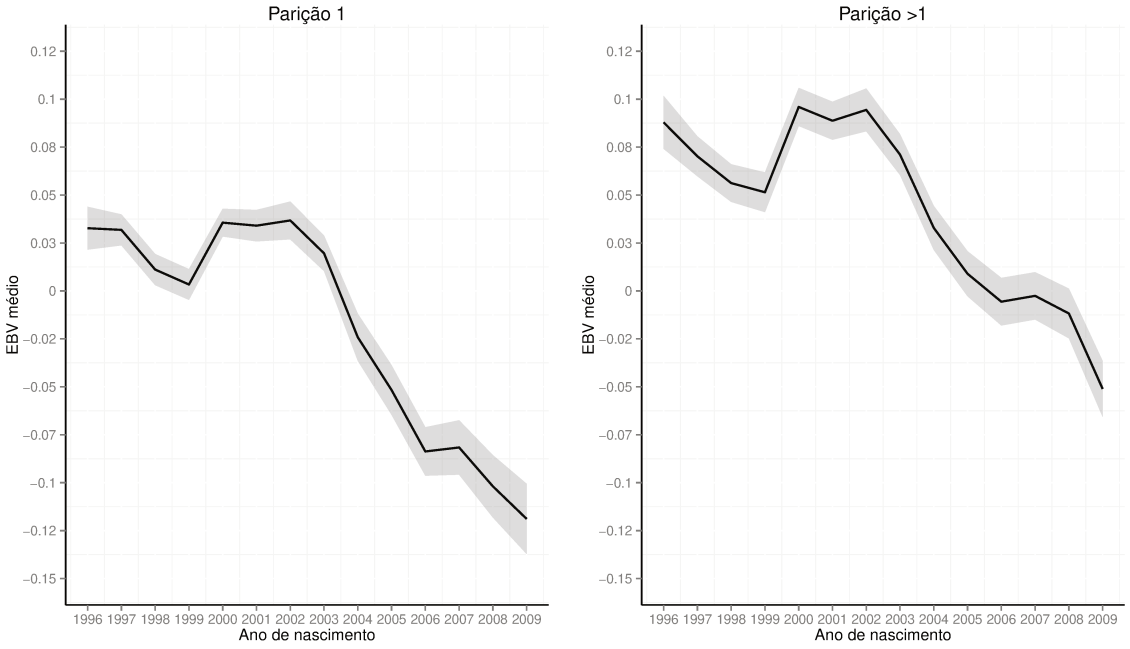


Figura 3.5: EBV's médios e IC's(95%) para média, modelo bivariado (parições) - todas as parições.

Em todos os modelos ajustados que consideram observações de todas as parições, a tendência genética apresenta comportamento similar. Após o ano de 2002, o programa de melhoramento no qual as unidades amostrais estão inseridas tem causado uma melhora quanto ao diagnóstico de

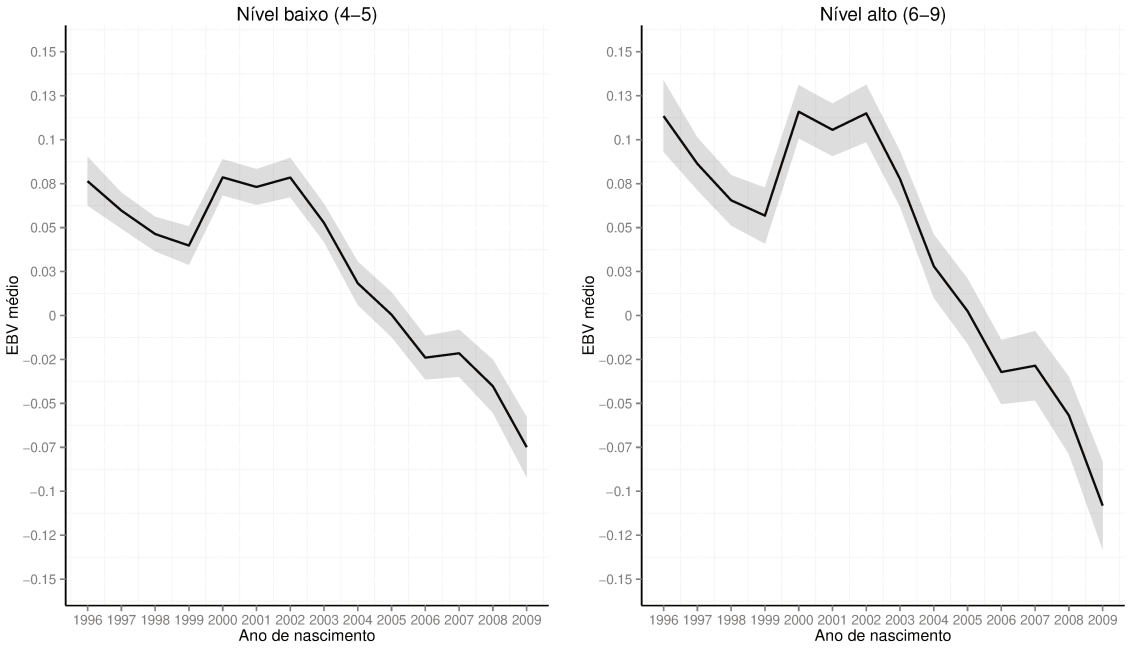


Figura 3.6: EBV's médios e IC's(95%) para média, modelo bivariado (severidade) - todas as parições.

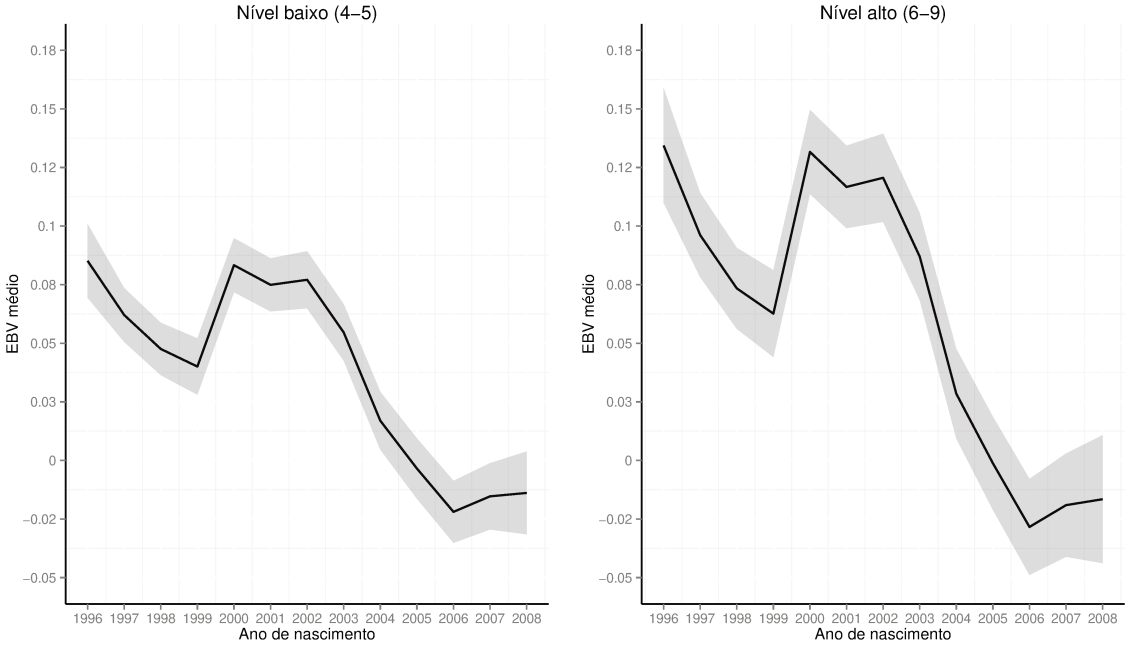


Figura 3.7: EBV's médios e IC's(95%) para média, modelo bivariado (severidade) - parições ≥ 2.

metritis, cujo coeficiente de correlação de Spearman estimado foi $\hat{\rho} = -0.97$ (p-valor = 0.0004), idêntico ao coeficiente de correlação estimado para os modelos que consideram observações da primeira parição. Questionamentos sobre a causa deste comportamento são necessários para entender o motivo pelo qual as unidades amostrais estão respondendo de maneira mais satisfatória à infecção desta doença.

3.6 Análise de resíduos

Uma das etapas mais importantes no processo de modelagem estatística é a análise de resíduos. Nesta etapa é possível verificar se os modelos ajustados estão, de fato, conseguindo predizer de maneira satisfatória os dados coletados. Neste trabalho, um dos questionamentos possíveis a respeito da qualidade do modelo ajustado é se este consegue predizer, por exemplo, a probabilidade de que um determinado touro tenha uma filha diagnosticada com metritis em alguma parição. Uma medida similar seria verificar se o modelo consege predizer o número de filhas de cada touro diagnosticadas com a doença.

Nesta seção, apresentaremos uma breve análise da qualidade dos modelos estatísticos ajustados, tanto para aqueles que consideram observações somente da primeira parição quanto para aqueles que consideram observações de todas as parições.

3.6.1 Primeira parição

As Figuras 3.8, 3.9 e 3.10 apresentam os gráficos de dispersão dos valores observados e esperados, de acordo com o modelo estatístico, do número de observações diagnosticadas com metritis *postpartum* para cada touro da amostra. É esperado que, para um bom modelo, tais valores se distribuam em torno da reta com intercepto igual a zero e coeficiente de inclinação igual a um. Além disso, as figuras supracitadas também apresentam gráficos de caixa dos valores da diferença entre o valor observado e o valor ajustado.

Como podemos observar, com exceção de algumas observações atípicas, o modelo estatístico conseguiu prever de maneira satisfatória o número de observações diagnosticadas com a doença para cada touro da amostra.

No modelo bivariado que considera diferentes níveis da doença, o modelo ajustado conseguiu prever de maneira mais consistente as observações diagnosticadas com a doença em um nível mais severo. Este fato pode ser verificado por meio dos gráficos de caixa que, para observações com

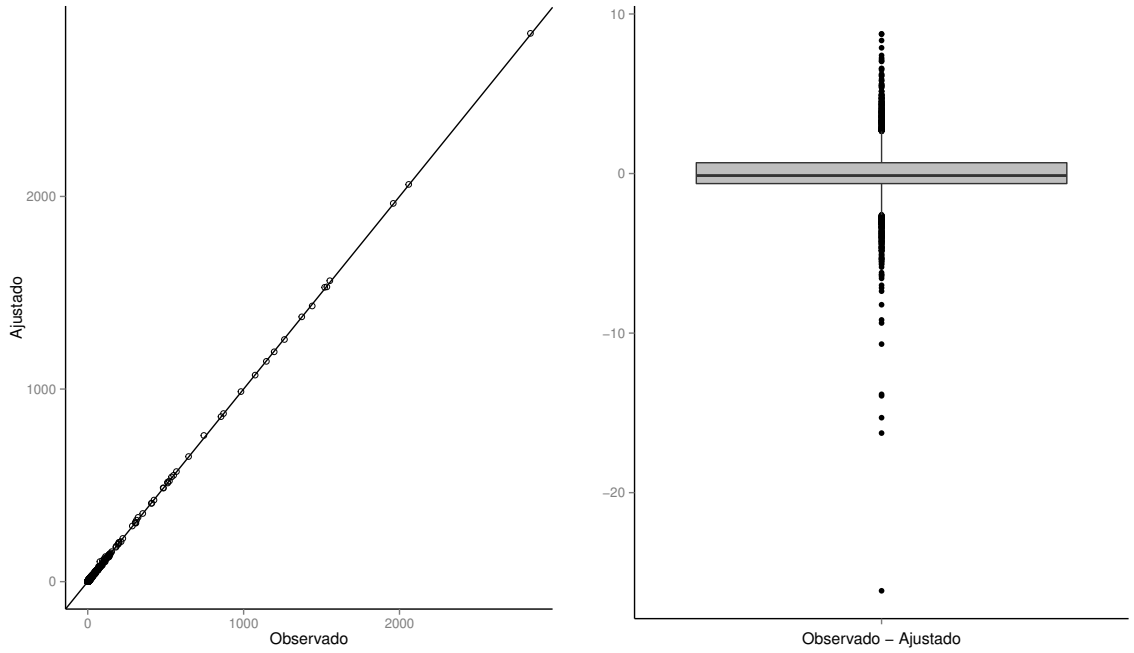


Figura 3.8: Número esperado/observado de filhas com metritis - gráfico de dispersão e caixa - modelo univariado.

nível de severidade alto, a variabilidade da diferença entre o valor observado e o valor esperado do número de filhas com metritis para cada touro é menor. Este fato é justificado pela raridade da doença em seu nível mais severo, levando a uma menor variabilidade da variável aleatória dicotômica que representa esta característica.

Análises adicionais foram feitas para verificar a qualidade dos ajustes quanto ao número esperado e ajustado do número de observações diagnosticadas com metritis para cada parição. Entretanto, uma vez que o modelo estatístico conseguiu prever corretamente o número de observações diagnosticadas com a doença em todas as nove parições, não exibiremos estes gráficos neste trabalho.

3.6.2 Todas as parições

As Figuras 3.11 e 3.12 apresentam os gráficos dos resíduos para o modelo bivariado que considera as diferentes parições como dimensões do modelo.

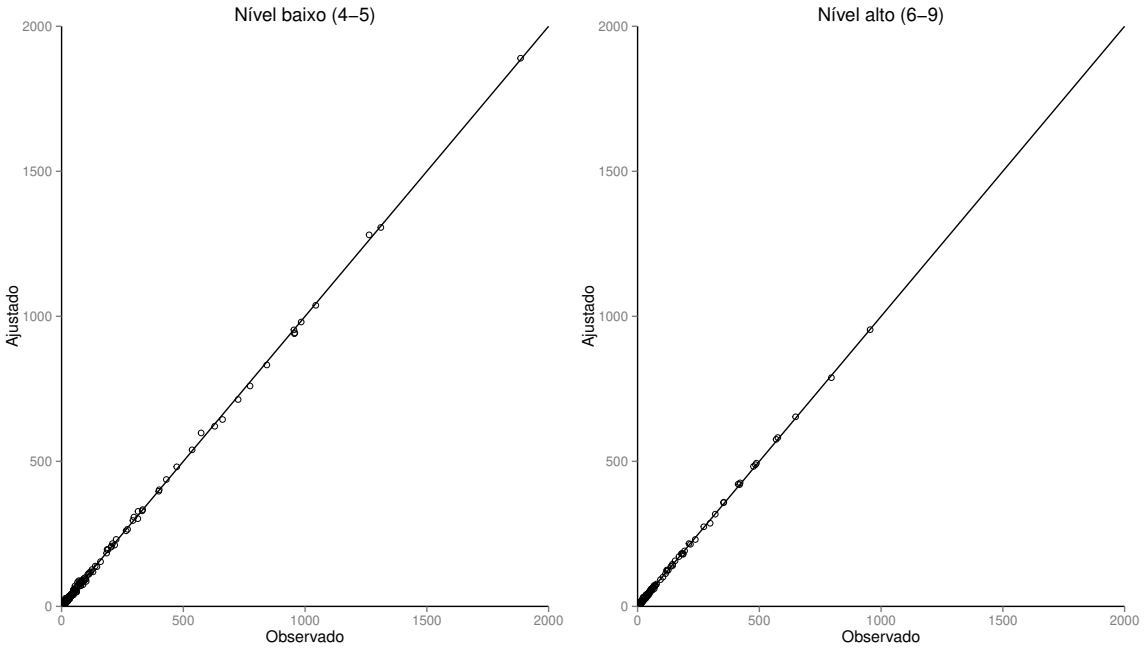


Figura 3.9: Número esperado/observado de filhas com metritis - gráficos de dispersão - modelo bivariado (primeira parição)

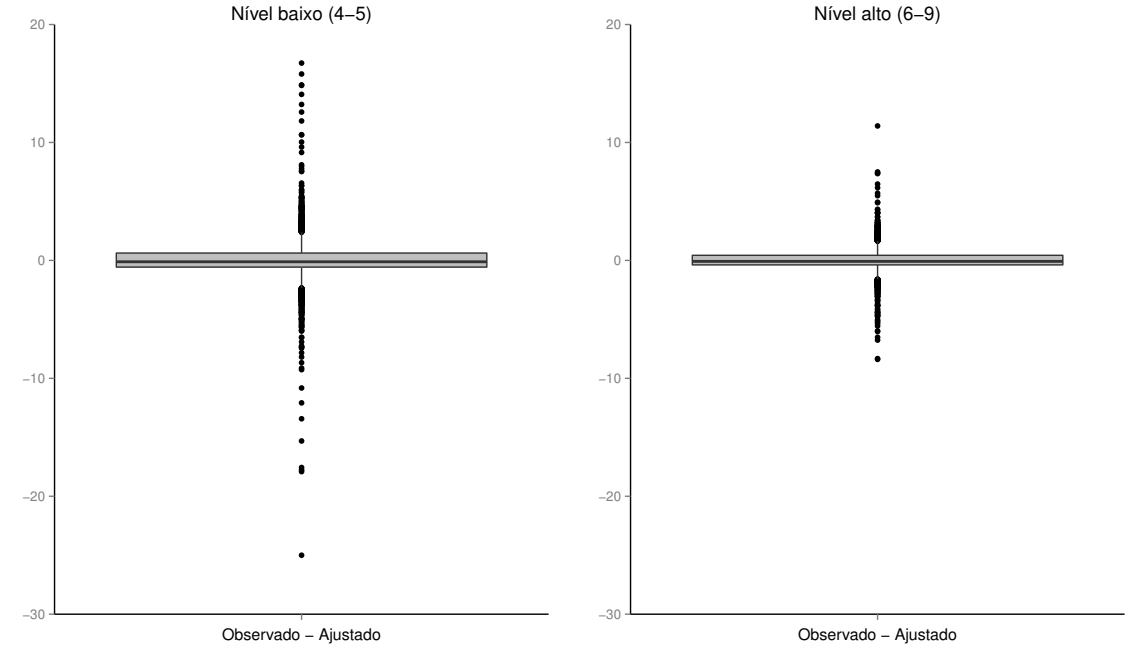


Figura 3.10: Diferença entre número esperado/observado de filhas com metritis - boxplot - modelo bivariado (primeira parição).

Por meio da análise gráfica das Figuras 3.11 e 3.12 podemos verificar que o modelo se ajustou bem aos dados. Novas covariáveis podem ser incluidas no modelo estatístico para conseguir explicar a variabilidade inerente aos dados e, assim, melhorar ainda mais a qualidade dos ajustes.

As Figuras 3.13 e 3.14 apresentam os gráficos dos resíduos para o modelo bivariado que considera os diferentes níveis da doença.

De maneira análoga, ambos os gráficos apresentam resultados similares ao modelo bivariado que considera as diferentes parições como dimensões do modelo estatístico. Com exceção de algumas observações atípicas, este modelo conseguiu, de maneira geral, prever satisfatoriamente bem o número de observações diagnosticadas com a doença para cada touro.

As Figuras 3.15 e 3.16 apresentam os gráficos dos resíduos para o modelo bivariado que considera os diferentes níveis da doença. Ambas as figuras supracitadas apresentam interpretações similares àquelas dos modelos anterioriormente apresentadas.

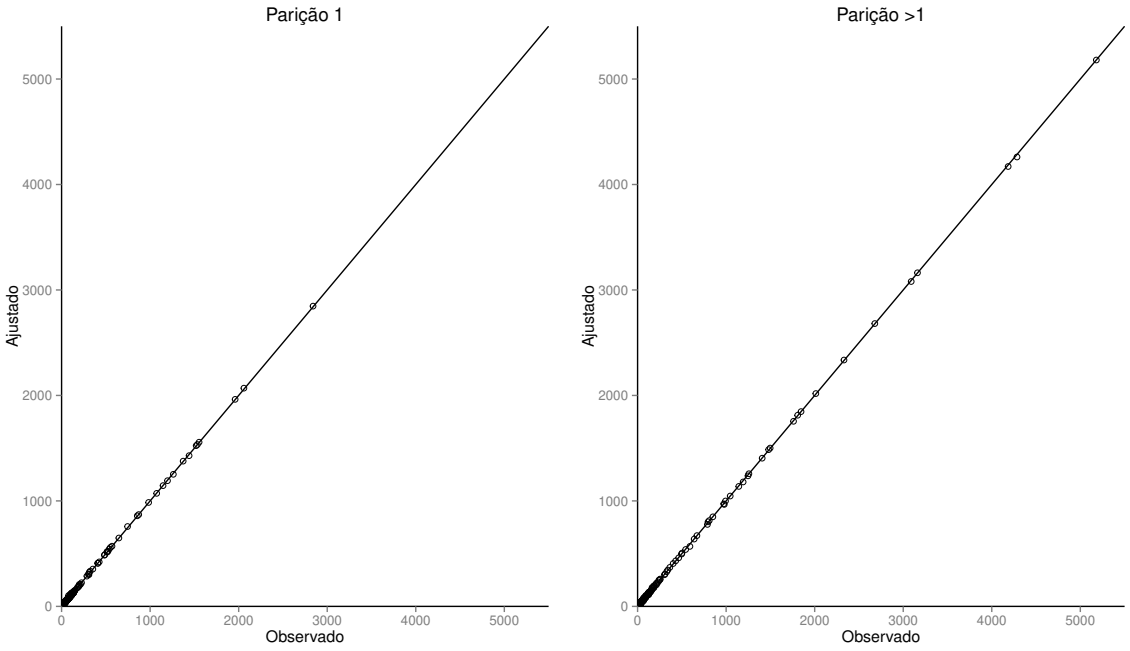


Figura 3.11: Número esperado/observado de filhas com metritis - gráficos de dispersão, modelo bivariado (parição) - todas as parições.

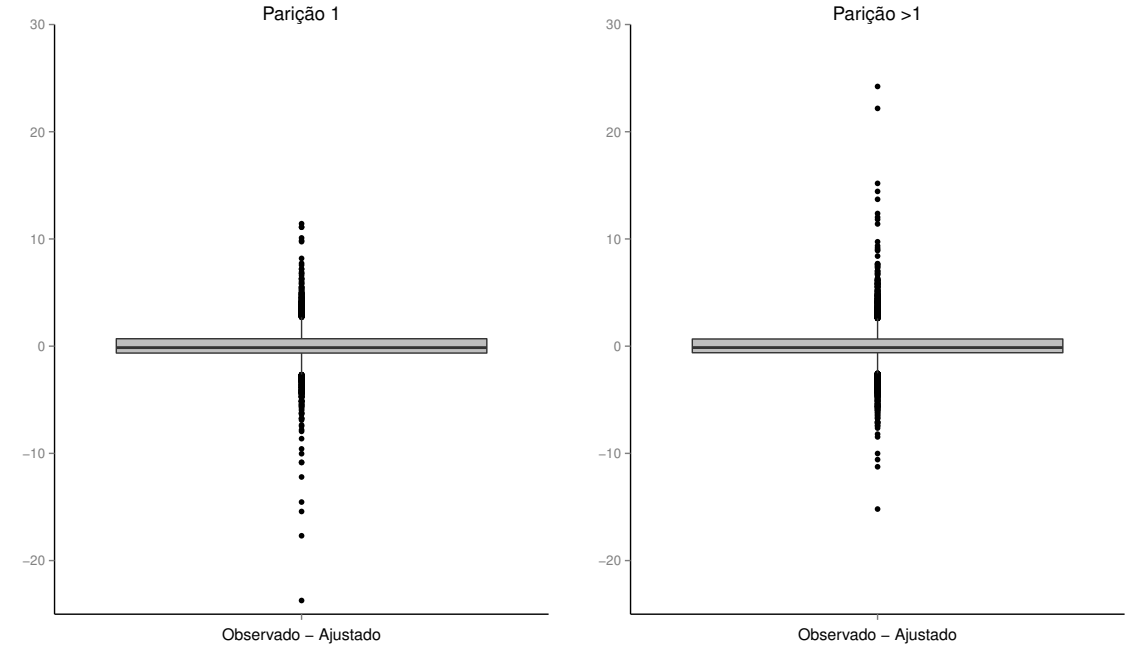


Figura 3.12: Diferença entre número esperado/observado de filhas com metritis - boxplot, modelo bivariado (parição) - todas as parições.

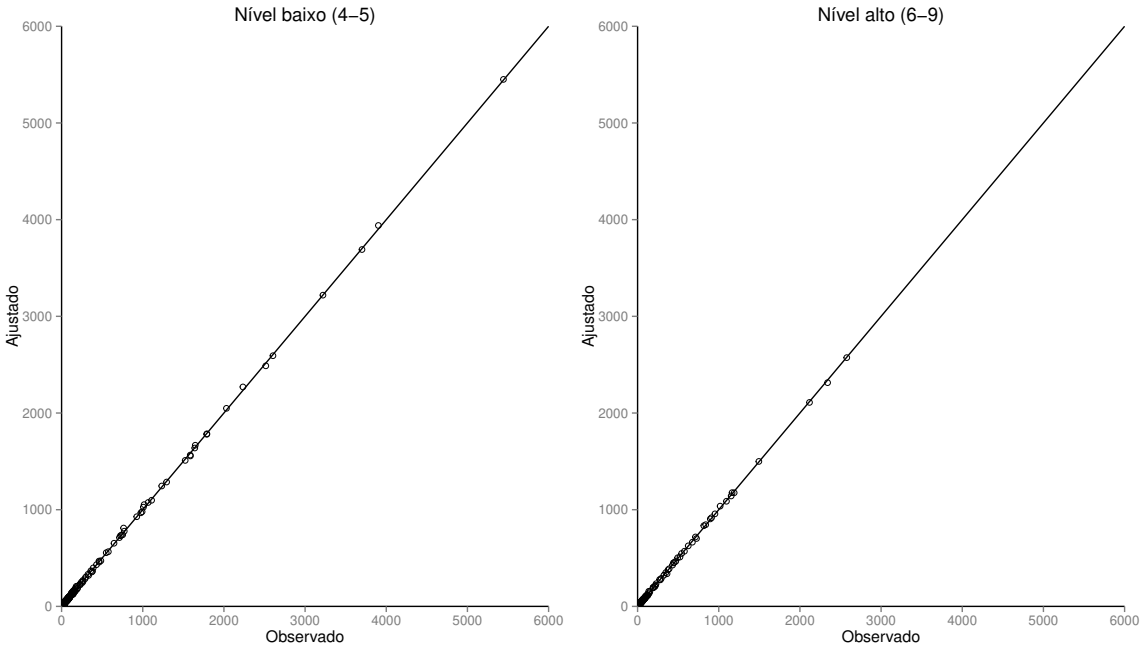


Figura 3.13: Número esperado/observado de filhas com metritis - gráficos de dispersão, modelo bivariado (severidade) - todas as parições.

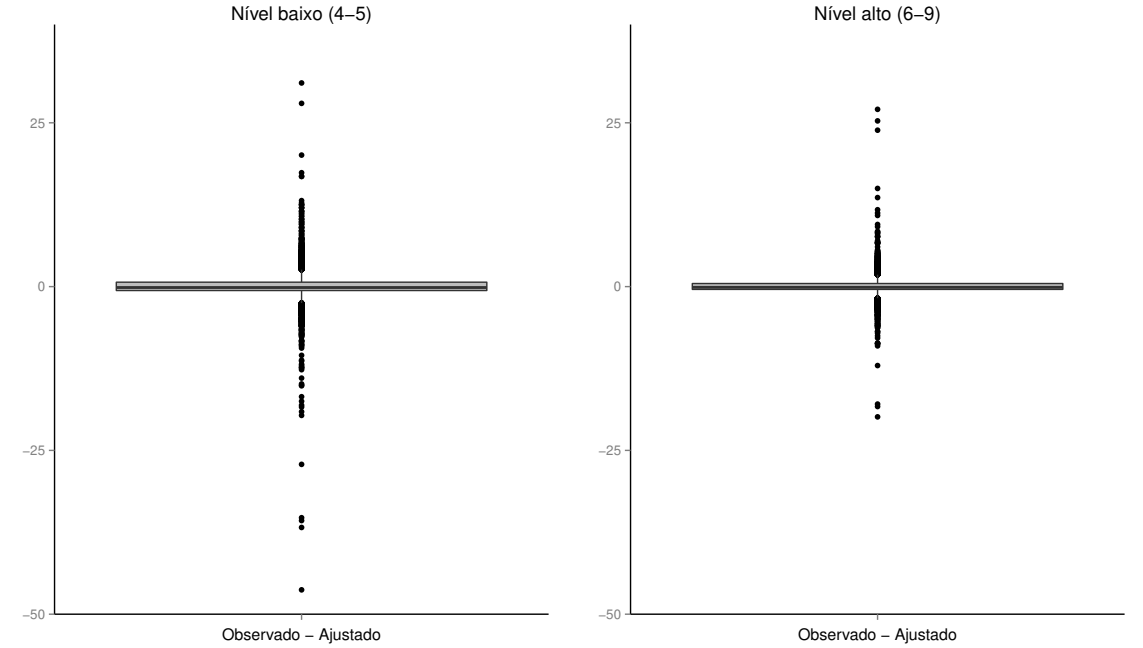


Figura 3.14: Diferença entre número esperado/observado de filhas com metritis - boxplot, modelo bivariado (severidade) - todas as parições.

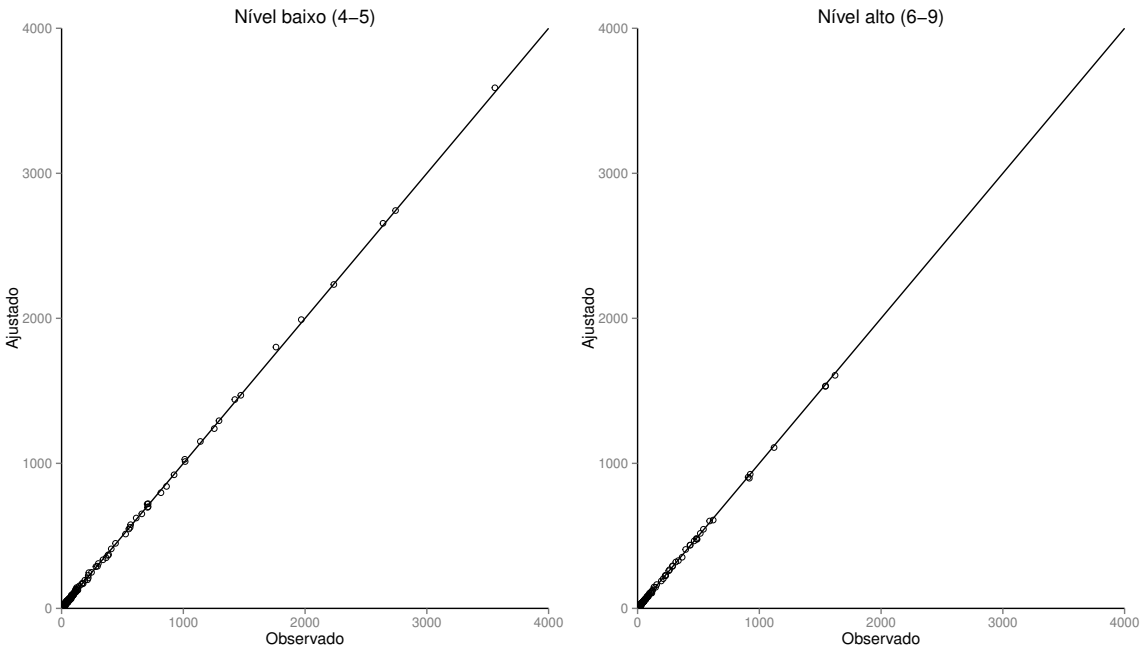


Figura 3.15: Número esperado/observado de filhas com metritis - gráficos de dispersão, modelo bivariado (severidade) - parições ≥ 2

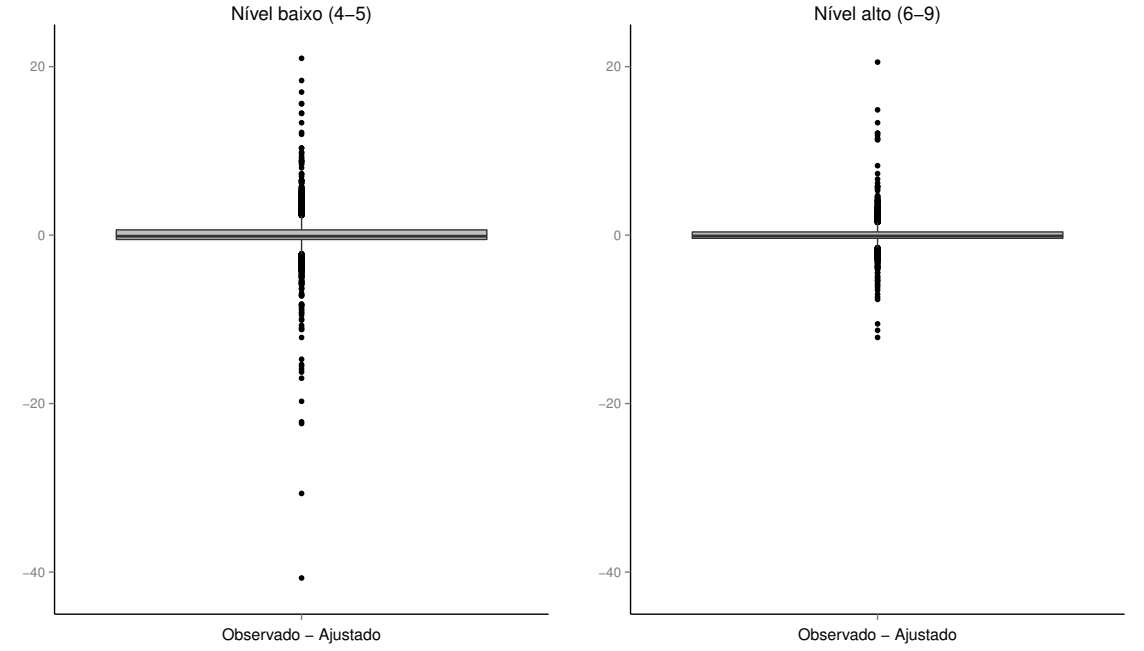


Figura 3.16: Diferença entre número esperado/observado de filhas com metritis - boxplot, modelo bivariado (severidade) - parições ≥ 2

3.7 Validação cruzada

Com o intuito de verificar a consistência do processo de estimação dos modelos ajustados, etapas de validação cruzada foram realizadas considerando amostras da população de touros em estudo.

Primeiramente, os EBV's de todos os touros presentes nos dados foram estimados no ajuste do modelo completo. Em seguida, foram ajustados diversos modelos semelhantes, em relação aos parâmetros, que consideravam somente observações de uma amostra dos touros em estudo. O processo de amostragem realizado foi sem reposição e considerou percentuais da população de touros de 10% a 90%, com incrementos de 10%. Finalmente, os EBV's dos touros não selecionados foram preditos de acordo com o modelo reduzido por meio dos dados de *pedigree*. Para cada percentual de amostragem foram ajustados 100 modelos reduzidos.

Para medir a consistência dos modelos, foi calculado o coeficiente de correlação de Spearman entre os EBV's preditos pelo modelo reduzido e os EBV's estimados pelo modelo completo. Além disto, as estimativas dos parâmetros das componentes de variância dos efeitos aleatórios e do parâmetro de dispersão foram armazenados para cada um dos cem modelos ajustados, para cada um dos nove percentuais de amostragem utilizados.

As Figuras 3.17 e 3.18 apresentam os gráficos dos resultados da validação cruzada para os modelos univariados que consideram observações somente da primeira parição e das demais parições, respectivamente. Do lado esquerdo são graficados a mediana das correlações de Spearman entre os EBV's preditos e os EBV's estimados para cada um dos nove percentuais de amostragem bem como os respectivos intervalos de confiança quantílicos (0.025 e 0.975). Do lado direito são graficados a mediana das estimativas das componentes de variância ($\sigma^2_g, \sigma^2_h, \sigma^2_{hys}$ e ϕ) bem como os respectivos intervalos de confiança quantílicos (0.025 e 0.975).

Interpretações semelhantes podem ser feitas para ambos os modelos. Podemos observar que a correlação de Spearman entre os EBV's estimados pelo modelo completo e os EBV's preditos pelo modelo reduzido decresce rapidamente a medida que selecionamos um menor número de touros

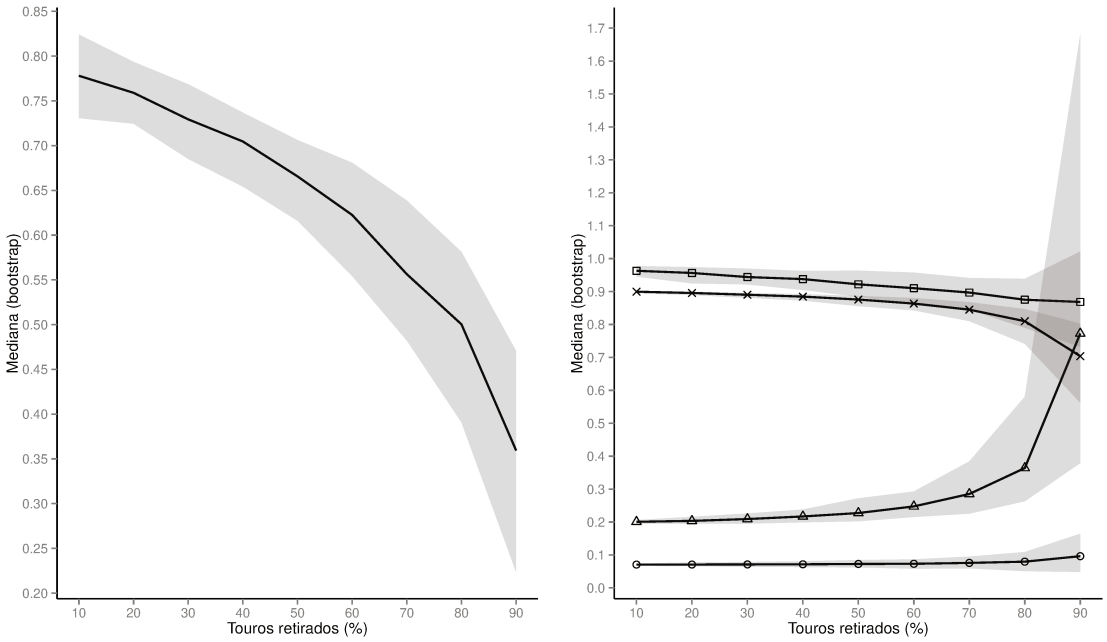


Figura 3.17: Estimativa bootstrap e IC(95%) da correlação entre os EBV's dos touros estimados e preditos (esquerda) e das estimativas das componentes de variância (direita): $\bigcirc = \sigma_g^2$, $\square = \sigma_h^2$, $\triangle = \sigma_{hys}^2$ e $\times = \phi$ - Modelo univariado (primeira parição)

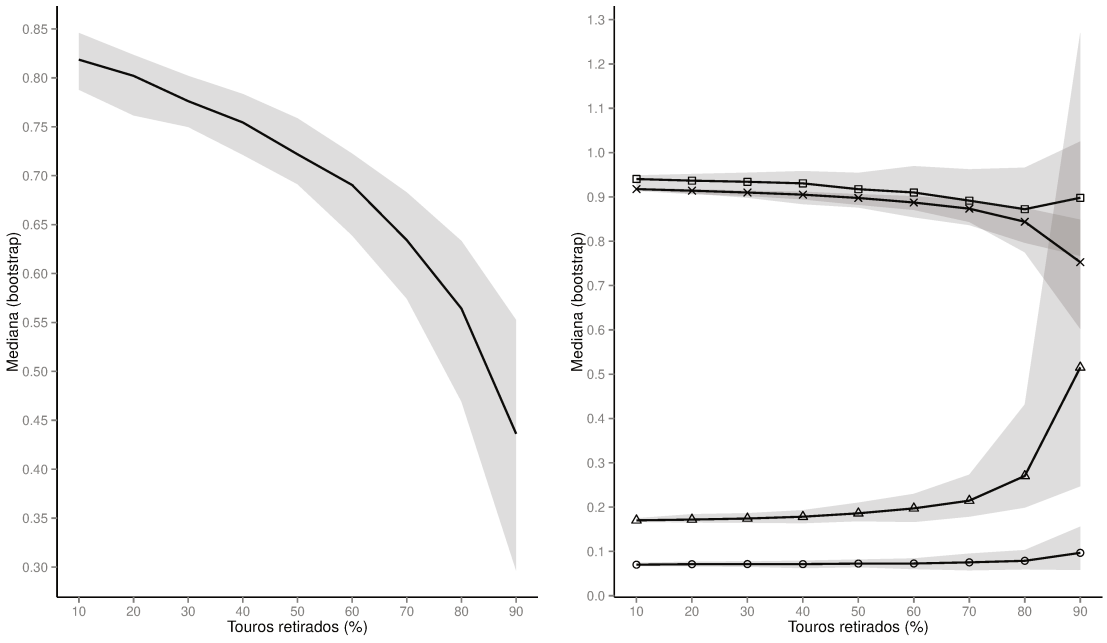


Figura 3.18: Estimativa bootstrap e IC(95%) da correlação entre os EBV's dos touros estimados e preditos (esquerda) e das estimativas das componentes de variância (direita): $\bigcirc = \sigma_g^2$, $\square = \sigma_h^2$, $\triangle = \sigma_{hys}^2$ e $\times = \phi$ - Modelo univariado (parições ≥ 2)

para o ajuste dos modelos. Com 50% dos touros selecionados, a correlação mediana de Spearman entre os EBV's preditos e estimados foi igual a 0.66 (IC 95%: 0.62;0.71) para o modelo univariado que considera observações somente da primeira parição, e igual a 0.72 (IC 95%: 0.69;0.76) para o modelo univariado que considera as demais parições.

As estimativas da componente de variância associada aos efeitos genéticos permanecem praticamente constante à medida que selecionamos menores amostras de touros para os ajustes dos modelos. Para o modelo univariado que considera somente observações da primeira parição, a estimativa mediana da componente de variância genética com somente 20% dos touros selecionados foi igual a 0.08 (IC 95%: 0.05;0.11), enquanto que para o modelo univariado que considera observações das demais parições, a estimativa mediana deste parâmetro com apenas 20% dos touros foi igual a 0.08 (IC 95%: 0.06;0.10). Quando selecionamos apenas 10% dos touros, as estimativas medianas da componente de variância associada aos fatores genéticos para os modelos supracitados são iguais a 0.10 (IC 95%: 0.05;0.16) e 0.10 (IC 95%: 0.06;0.15), respectivamente. Este fato nos diz basicamente que mesmo considerando uma pequena amostra de touros é possível estimar de maneira satisfatória a componente de variância associada aos fatores genéticos.

Por outro lado, as Figuras 3.17 e 3.18 nos dizem que, para ambos os modelos, a componente de variância associada a interação entre fazenda, ano e estação não é bem estimada quando consideramos somente 20% dos touros, por exemplo. A estimativa mediana da componente de variância associada à interação com somente 20% dos touros selecionados foi igual a 0.36 (IC 95%: 0.26;0.58), enquanto que para o modelo univariado que considera observações das demais pariçoes, a estimativa mediana deste parâmetro com apenas 20% dos touros foi igual a 0.27 (IC 95%: 0.20;0.43).

3.8 Discussão

As análises dos resultados dos ajustes dos modelos apresentados neste capítulo permitiram responder algumas perguntas pertinentes, dentro do contexto de genética quantitativa, do problema

em questão.

A existência de mecanismos genéticos associados à susceptibilidade de metritis *postpartum* pôde ser verificada por meio da análise dos resultados do modelo univariado. Além disso, os modelos bivariados (severidade) sugeriram que efeitos genéticos estão associados ao desenvolvimento da doença, tanto para o nível baixo quanto para o nível alto de severidade. Entretanto, a alta correlação genética estimada neste modelo (próxima de 1) mostrou que tais mecanismos são praticamente os mesmos. Os resultados do ajuste do modelo bivariado (parições) sugeriram a existência de mecanismos genéticos atuantes no desenvolvimento de metritis, tanto para as observações da primeira parição quanto para as demais parições. Além disso, a correlação genética estimada neste modelo (0.67, E.P. = 0.04) sugere a existência de mecanismos genéticos independentes e comuns atuantes em cada uma das dimensões do modelo.

A análise da tendência genética mostrou, em todos os modelos ajustados, que o programa de seleção no qual os animais em estudo estão inseridos tem causado uma melhora, do ponto de vista genético, na resistência dos animais quanto à metritis *postpartum*. Tal efeito pôde ser verificado pela tendência negativa apresentada nos gráficos dos EBV's médios após o ano de 2002. Questionamentos adicionais podem e devem ser feitos para saber qual a real causa deste fenômeno. A utilização de tratamentos para o combate de outras doenças ou a seleção dos animais para a produção de leite poderiam causar, de maneira direta e indireta, tal melhoria.

As simulações realizadas nas etapas de validação cruzada mostraram que os modelos univariados, ajustados por meio do método de quase-verossimilhança penalizada, são relativamente consistentes no sentido que as estimativas dos parâmetros das componentes de variância associadas aos fatores genéticos permanecem aproximadamente iguais com 20% dos touros retirados. Entretanto, o mesmo fato não ocorre com $\hat{\sigma}^2_{hys}$. Além disso, a estimativa do coeficiente de correlação de Spearman para cada etapa de validação mostrou que a correlação entre os EBV's preditos pelo modelo reduzido e os EBV's estimados pelo modelo completo decresce rapidamente à medida que retiramos um maior número de touros da amostra.

Capítulo 4

Considerações finais

Neste trabalho, pudemos estudar os Modelos Lineares Generalizados Mistos (MLGM) por meio de uma revisão da teoria desenvolvida para esta classe de modelos e de uma aplicação em dados reais. Tais modelos possuem a vantagem de conseguir modelar a superdispersão frequentemente observada nos dados, por exemplo, além de estabelecer uma estrutura de correlação entre as unidades amostrais através da inclusão de efeitos aleatórios.

A análise dos dados de metritis *postpartum* mostrou, de maneira ilustrativa, uma aplicação real desta classe de modelos. A genética quantitativa é uma área de pesquisa em franco desenvolvimento e a utilização de modelos estatísticos cada vez mais detalhados se faz necessária. Neste sentido, alguns trabalhos futuros podem ser desenvolvidos dentro do problema analisado com o intuito de melhorar ainda mais o processo de estimação dos parâmetros no problema em questão.

A otimização e o desenvolvimento de pacotes computacionais robustos para a aplicação de outras metodologias de estimação, como o algoritmo EM ou algoritmos MCMC, são temas de pesquisa que poderiam ser desenvolvidos futuramente. Tais avanços poderiam, por exemplo, permitir o uso de outras distribuições de probabilidade para os efeitos aleatórios e permitiria também a utilização de abordagens bayesianas.

Um outro aspecto a ser estudado futuramente é a interação existente entre genética e ambi-

ente. Por meio da utilização de dados brasileiros, por exemplo, seria possível verificar e estudar o comportamento dos mecanismos genéticos atuantes na susceptibilidade de metritis *postpartum* nos animais diagnosticados em diferentes ambientes. Neste sentido, a construção do modelo estatístico é feita de tal forma que as componentes aleatórias associadas aos fatores genéticos e aos fatores ambientais compartilhem determinada estrutura de correlação. Consequentemente, o processo de estimação de tais modelos se tornaria muito mais complexo devido ao aumento do número de parâmetros e, além disso, questionamentos mais detalhados poderiam ser feitos.

Referências

Andrade, M. e H.P. Pinheiro (2002). *Métodos estatísticos aplicados em genética humana*. 1a. ed. Vol. 1. São Paulo: ABE - Associacao Brasileira de Estatística, 180p.

Besag, Julian, Jeremy York e Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". Em: *Annals of the Institute of Statistical Mathematics* 43.1, pp. 1–20.

Breslow, Norman E (1984). "Extra-Poisson variation in log-linear models". Em: *Applied Statistics*, pp. 38–44.

Breslow, Norman E e David G Clayton (1993). "Approximate inference in generalized linear mixed models". Em: *Journal of the American Statistical Association* 88.421, pp. 9–25.

Breslow, Norman E e Xihong Lin (1995). "Bias correction in generalised linear mixed models with a single component of dispersion". Em: *Biometrika* 82.1, pp. 81–91.

Casella, George e Roger L Berger (1990). *Statistical inference*. Vol. 70. Duxbury Press Belmont, CA.

Dempfle, L (1990). "Problems in the use of the relationship matrix in animal breeding". Em: *Advances in statistical methods for genetic improvement of livestock*. Springer, pp. 454–473.

Elkjær, K et al. (2013). "Large-scale study on effects of metritis on reproduction in Danish Holstein cows". Em: *Journal of Dairy Science* 96.1, pp. 372–377.

Feero, W. Gregory, Alan E. Guttmacher e Francis S. Collins (2010). "Genomic Medicine - An Updated Primer". Em: *New England Journal of Medicine* 362.21. PMID: 20505179, pp. 2001–2011.

Green, Peter J (1987). "Penalized likelihood for general semi-parametric regression models". Em: *International Statistical Review/Revue Internationale de Statistique*, pp. 245–259.

Harville, David A (1977). "Maximum likelihood approaches to variance component estimation and to related problems". Em: *Journal of the American Statistical Association* 72.358, pp. 320–338.

Jorgensen, Bent, Rodrigo Labouriau e Soren Lundbye-Christensen (1996). "Linear Growth Curve Analysis Based on Exponential Dispersion Models". Em: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.3, pp. 573–592.

Labouriau, Rodrigo et al. (2014). "Genetic determination of susceptibility for metritis in Danish Holstein cows". Em: *To be submitted*.

Lin, Xihong e Norman E Breslow (1996). "Bias correction in generalized linear mixed models with multiple components of dispersion". Em: *Journal of the American Statistical Association* 91.435, pp. 1007–1016.

Lynch, Michael e Bruce Walsh (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland.

McCullagh, Peter e John A Nelder (1989). "Generalized linear models (Monographs on statistics and applied probability 37)". Em: *Chapman Hall, London*.

McCulloch, Charles E (1997). "Maximum likelihood algorithms for generalized linear mixed models". Em: *Journal of the American statistical Association* 92.437, pp. 162–170.

McCulloch, Charles E e Shayle R Searle (2001). "Generalized linear mixed models (GLMMs)". Em: *Generalized, Linear, and Mixed Models*, pp. 220–246.

Patterson, H Desmond e Robin Thompson (1971). "Recovery of inter-block information when block sizes are unequal". Em: *Biometrika* 58.3, pp. 545–554.

Raudenbush, Stephen W, Meng-Li Yang e Matheos Yosef (2000). "Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation". Em: *Journal of computational and Graphical Statistics* 9.1, pp. 141–157.

Sheldon, I Martin et al. (2009). "Defining postpartum uterine disease and the mechanisms of infection and immunity in the female reproductive tract in cattle". Em: *Biology of reproduction* 81.6, pp. 1025–1032.

Stiratelli, Robert, Nan Laird e James H Ware (1984). "Random-effects models for serial observations with binary response". Em: *Biometrics*, pp. 961–971.

Williams, David A (1982). "Extra-binomial variation in logistic linear models". Em: *Applied statistics*, pp. 144–148.

Wright, Sewall (1922). "Coefficients of inbreeding and relationship". Em: *The American Naturalist* 56.645, pp. 330–338.