Universidade de Brasília - UnB

Instituto de Ciências Exatas

Departamento de Estatística

Modelos Para Análise De Dados Não-Normais Multivariados Longitudinais

Rubem Kaipper Ceratti

Brasília, DF

2013

Rubem Kaipper Ceratti

Modelos para análise de dados não-normais multivariados longitudinais

Dissertação apresentada no programa de pós-graduação em Estatística, Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de Mestre em Estatística.

Orientador: Prof.º Dr. Afrânio Márcio Corrêa Vieira

Brasília - DF
2013

Agradecimentos

À minha família pelo apoio incondicional neste período de tanto trabalho. À Embrapa recursos Genéticos e Biotecnologia e aos pesquisadores Maria Carolina Moraes, Raul Laumann, Miguel Borges e Joseane Padilha pela motivação do problema estatístico, oferta de dados, além de disponibilidade e paciência para sanar dúvidas.

Sumário

Resumo

Neste trabalho são abordados modelos lineares generalizados de efeitos mistos para análise de dados longitudinais multivariados, no tratamento de dados em que se assume a distribuição Poisson composta, que tem suporte em $[0, +\infty)$ e é um caso particular da família Tweedie de distribuições, também pertencente à família exponencial de dispersão. No ajuste dos modelos mistos multivariados para a distribuição Poisson composta, utiliza-se uma abordagem de pseudo-verossimilhança, estimando modelos par-a-par e reduzindo o tempo computacional. Como aplicação, analisa-se um conjunto de dados provenientes de um experimento agronômico no qual avaliam-se os efeitos de tratamentos, ao longo do tempo, no perfil de 25 compostos químicos de plantas de algodão.

Palavras-chave: Análise multivariada, dados longitudinais, família exponencial, distribuição Poisson composta, modelos lineares generalizados, modelos de efeitos mistos.

Abstract

This work presents generalized linear mixed effects models as a framework to the analysis of longitudinal multivariate data for which the underlying distribution is assumed to follow a compound Poisson distribution, whose support lies in $[0, +\infty)$, and is a particular case of the Tweedie family of distributions, and, also, belongs to the exponential dispersion family. In order to fit multivariate mixed models to the compound Poisson distribution, a pseudo-likelihood approach is used, fitting pairwise models and reducing computational time. As an application, agronomic experiment data is analyzed, estimating the effects of 5 treatments, over different time periods, on the profile of 25 organic compounds of cotton plants.

Keywords: Multivariate data, longitudinal data, exponential family, compound Poisson distribution, generalized linear models, mixed effects models.

Capítulo 1

Introdução

A análise de dados observacionais ou experimentais por meio de modelos estatísticos é sempre acompanhada por um conjunto de suposições a respeito do processo gerador desses dados. Em geral, essas suposições são feitas com base na escala das variáveis observadas – qualitativa, quantitativa contínua, quantitativa discreta – bem como na forma das possíveis estruturas de relação entre as observações – observações independentes, ao longo do tempo, em pontos no espaço, ou, ainda, a observação simultânea de um conjunto de variáveis.

Estas suposições distribucionais e de estrutura, unidas às observações na forma da função de verossimilhança, ou distribuição de probabilidade conjunta dos dados, permitem que sejam realizadas inferências a respeito dos parâmetros da distribuição, que, por sua vez, constituem o objeto de interesse da análise. A partir das estimativas destes parâmetros é feita a generalização de resultados perante a população estudada.

Historicamente, duas suposições comuns na análise estatística são de normalidade da distribuição dos dados e/ou independência entre as observações. Por muito tempo, isso se deveu a limitações computacionais, ainda que sejam, em muitos casos, suposições inadequadas. Entretanto, a necessidade de resultados mais precisos aliada à intensa evolução das ferramentas computacionais nas últimas décadas, permitiu o desenvolvimento, a utilização e a disseminação de modelos mais realistas na análise de dados complexos.

Para aqueles casos em que a suposição de normalidade dos dados não é razoável, apresentam-se modelos aplicáveis quando se têm variáveis de interesse de diferentes tipos (por exemplo, dados contínuos, dicotômicos ou de contagem), abrangendo distribuições pertencentes à família exponencial - Poisson, Binomial, Gama, Normal inversa, entre outras. Esses modelos, que podem ser utilizados na presença de variá-

veis explicativas, fazem parte da classe denominada modelos lineares generalizados (McCullagh; Nelder, 1989).

Por outro lado, quando a independência entre as observações não pode ser garantida – por exemplo, em experimentos com medidas longitudinais ou *cross-over*, nos quais a resposta de interesse é medida em cada indivíduo mais de uma vez –, pode-se utilizar a abordagem de modelos mistos, uma classe de modelos para a qual a correlação entre as respostas é considerada por meio da presença de efeitos ditos aleatórios, que caracterizam os efeitos das estruturas de delineamento que geram dependências, em adição aos efeitos de variáveis explicativas comuns.

Quando se tem a medição de uma variável em diversos tempos em um mesmo indivíduo, pode-se considerar que cada indivíduo possui um vetor de respostas associado a ele e que, portanto, tem-se uma resposta multivariada. De forma mais geral, pode-se pensar que o vetor de variáveis respostas é constituído por diferentes variáveis, possivelmente correlacionadas, que são medidas simultaneamente em cada indivíduo. Entretanto, é possível que se tenha interesse em medir um certo conjunto de variáveis resposta ao longo do tempo. Dados com este tipo de estrutura também podem ser analisados por meio de modelos de efeitos mistos, nos quais os efeitos longitudinal e multivariado são tratados por meio da introdução de efeitos aleatórios multivariados.

De particular interesse neste trabalho é o uso de extensões dos modelos lineares generalizados mistos para análise de dados não gaussianos longitudinais multivariados – mais especificamente, o caso em que se supõe que a variável resposta segue uma distribuição Poisson composta. Além da forma dos modelos em si, um ponto chave deste trabalho é a abordagem da estimação via pseudo-verossimilhança, ou ajuste de modelos por pares, que oferece redução do custo computacional. Para isso, foi implementada a abordagem de ajuste par-a-par de modelos lineares mistos para a distribuição Poisson composta baseado no pacote *cplm* (Zhang, 2012a) da linguagem de computação estatística *R* (R Development Core Team, 2012).

Com isso, apresenta-se no capítulo 2 uma revisão metodológica de modelos lineares generalizados e modelos lineares generalizados mistos para o caso univariado, seguido, no capítulo 3, pela revisão de MLG multivariado. No capítulo 4, motiva-se a aplicação dos modelos discutidos nos capítulos 2 e 3 por meio da análise de um subconjunto de dados de experimento agronômico – são analisadas as massas de 3 compostos químicos liberados ao longo do tempo por plantas de algodão submetidas à diferentes tratamentos comparando-se o ajuste de modelos com diferentes graus de complexidade. No capítulo 5, exploram-se aspectos da implementação da abordagem de modelos par-a-par. No capítulo 6, o conjunto de dados com todos

os compostos é analisado via abordagem de pseudo-verossimlhança. Por fim, no capítulo 7, apresenta-se as conclusões do trabalho.

Capítulo 2

Revisão Metodológica

2.1 Modelos Lineares Generalizados

2.1.1 Introdução

Durante muito tempo, o modelo linear clássico, baseado na suposição de distribuição Normal da variável resposta, foi quase que exclusivamente o único modelo utilizado na descrição de dados, mesmo quando a suposição de normalidade não era realista. Para estes casos, transformações como a de Box e Cox (Box; Cox, 1964), efetuadas sobre a variável resposta, foram propostas para que as condições requeridas fossem aproximadamente atendidas.

Por vezes, porém, as transformações propostas não eram únicas ou mesmo adequadas. Dessa forma, Nelder e Wedderburn (1972) propuseram uma nova classe de modelos denominada Modelos Lineares Generalizados (MLG), que permite que a variável resposta assuma distribuição não somente Normal, mas qualquer distribuição pertencente à família exponencial - Poisson, Binomial, Gama, Normal inversa, dentre outras - permitindo maior flexibilidade na relação entre variável resposta e o preditor linear.

Outro propósito igualmente importante, foi integrar diferentes metodologias que eram tratadas separadamente, mas que guardavam similaridades: teste t, ANOVA, ANCOVA, modelo linear geral, regressão logística, regressão Poisson, modelos log-lineares, análise de sobrevivência, entre outros.

2.1.2 Família exponencial e MLG

Seja uma amostra de tamanho n de pares de observações (x_i, y_i), compondo uma matriz $\mathbf{X}$ de dimensão $n \times p$ com $p-1$ variáveis explicativas e $\mathbf{y}$ um vetor de

observações da variável resposta, em que cada elemento é uma realização da variável aleatória Y. Assume-se que os y_i são independentes e Y_i tem densidade dada por

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{1}{a_i(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\right] \tag{2.1}$$

com

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{2.2}$$

e

$$Var(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i) \tag{2.3}$$

sendo, pela notação de Smyth (1989), θ_i o parâmetro canônico, $a_i(\phi) = \phi/w_i$, ϕ o parâmetro de dispersão, w_i um peso "a priori"e $V(\mu_i)$ a função de variância dada por $V(\mu_i) = d\mu_i/d\theta_i$. A Tabela 2.1 apresenta os termos da família exponencial para as principais distribuições.

Os fatores e covariáveis estão organizados na matriz $\mathbf{X}$ e são expressos no preditor linear na forma

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{2.4}$$

e a relação funcional entre a média da variável resposta Y e o preditor linear é dada por

$$g(\mu_i) = \eta_i \tag{2.5}$$

em que $\boldsymbol{\beta}$ é o vetor dos parâmetros, $\boldsymbol{x}_i^\top$ é a i-ésima linha (observação) da matriz $\mathbf{X}$ e $g(\cdot)$ é uma função monotônica e diferenciável, denominada função de ligação. Tem-se na Tabela 2.2 as funções de ligação canônicas para as principais distribuições.

2.1.3 Estimação de parâmetros

Para a estimação do vetor de parâmetros β, seja o logaritmo da função de verossimilhança da família exponencial descrito como

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n}\left\{\frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\right\} \tag{2.6}$$

As estimativas de máxima verossimilhança de β serão obtidas pela solução do sistema de $p \times 1$ equações

Tabela 2.1: Termos da família exponencial para as distribuições Normal, Poisson, Binomial e Gama

Distribuição	$a(\phi)$	θ	$b(\theta)$	$c(y;\phi)$	$\mu(\theta)$	$V(\mu)$
Normal (μ, σ^2)	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$	θ	1
Poisson (μ)	1	$\ln(\mu)$	e^θ	$-\ln(y!)$	e^θ	μ
Binomial (n, π)	1	$\ln\left(\frac{\pi}{1-\pi}\right)$	$n\ln(1+e^\theta)$	$\ln\left[\frac{n!}{y!(n-y)!}\right]$	$n\frac{e^\theta}{1+e^\theta}$	$n\pi(1-\pi)$
Gama (μ, ν)	ν^{-1}	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\nu\ln(\nu y) - \ln(y) - \ln(\Gamma(\nu))$	$-\frac{1}{\theta}$	μ^2

Tabela 2.2: Ligação canônica para as distribuições Normal, Poisson, Binomial e Gama

Distribuição	Função de ligação canônica
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \ln(\mu)$
Binomial	Logística: $\eta = \ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{\mu}{n-\mu}\right)$
Gama	Recíproca: $\eta = 1/\mu$

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \boldsymbol{\theta}}\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\frac{\partial \boldsymbol{\eta}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)V(\mu_i)} \cdot \frac{x_{ij}}{g'(\mu_i)} = 0, \qquad j = 0, ..., p-1\,, \quad (2.7)$$

em que $g'(\mu_i) = d\eta_i/d\mu_i$ é a derivada da função de ligação.

Não há uma forma analítica fechada para a solução deste sistema de equações – exceto para a distribuição normal –, e portanto, faz-se necessária a utilização de um método numérico iterativo. As estimativas dos parâmetros são obtidas por meio do algoritmo IRLS (*Iteratively Reweighted Least Squares*):

1. $\boldsymbol{\eta}^{(k)}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}^{(k)}_{p\times 1}$;
2. $\boldsymbol{z}^{(k)}_{n\times 1} = \boldsymbol{\eta}^{(k)}_{n\times 1} + g'(\boldsymbol{\mu}^{(k)})_{n\times n}(\boldsymbol{y} - \boldsymbol{\mu}^{(k)})_{n\times 1}, \qquad \boldsymbol{\mu}^{(k)} = g^{-1}(\boldsymbol{\eta}^{(k)}), \quad g'(\boldsymbol{\mu}^{(k)}) =$

$diag(g'(\mu_1^{(k)}), ..., g'(\mu_n^{(k)}))$;

3. $\mathbf{W}_{n\times n}^{-1(k)} = [g'(\boldsymbol{\mu}^{(k)})]^2 V(\boldsymbol{\mu}^{(k)})\mathbf{W}_0^{-1}$; $\mathbf{W}_0 = diag(w_1, ..., w_n)$;

4. $\boldsymbol{\beta}_{p\times 1}^{(k+1)} = (\mathbf{X}^\top \mathbf{W}^{(k)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(k)} z^{(k)}$;

5. Repetir passos 1 a 4 até convergência;

Assintoticamente, o estimador de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ tem distribuição normal multivariada com média $\boldsymbol{\beta}$ e matriz de covariâncias $\phi(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \phi(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}).$$

Mais detalhes a respeito da estimação dos parâmetros dos MLGs, ver Demétrio (2002).

2.1.4 Resíduos em MLG

Considerando-se um modelo com p parâmetros, denominado modelo corrente (ou ajustado) e o modelo com n parâmetros, denominado modelo saturado, a *scaled deviance* é dada por

$$S_p = -2(\hat{l}_p - \hat{l}_n) \ , \tag{2.8}$$

em que $\hat{l}_p$ e $\hat{l}_n$ são os máximos da função de verossimilhança dos modelos corrente e saturado, respectivamente. Pode-se ainda escrever S_p da seguinte forma:

$$S_p = \frac{D_p}{\phi} = \frac{1}{\phi}\sum_{i=1}^{n} d_i^2 \tag{2.9}$$

em que D_p é chamada *deviance* e $d_i^2 = -2\int_{y_i}^{\hat{\mu}_i} \frac{y_i - t}{V(t)} dt$ é o componente de *deviance*. A Tabela ?? apresenta as expressões da função de *deviance* para algumas distribuições da família exponencial.

Definem-se, então, os resíduos para o modelo ajustado:

1. Resíduo componente de deviance:

$$d_i = sinal(y_i - \hat{\mu}_i)\sqrt{d_i^2}$$

2. Resíduo de Pearson:

$$r_{Pi} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)^{1/2}}.$$

Esses resíduos são utilizados para análise de diagnósticos.

2.1.5 Superdispersão

Para dados na forma de contagem ou de proporção, a superdispersão (ou sobredispersão) é caracterizada quando, na modelagem dos dados, a variação observada é maior que aquela assumida pelo modelo probabilístico (McCullagh; Nelder, 1989).

Para um modelo estatístico no qual se assume que a variável resposta tem distribuição binomial ou Poisson, considera-se que o parâmetro de dispersão ϕ é igual a 1. Assim, no caso binomial, se $Y_i \sim Bin(n_i, \pi_i)$, então, $E(Y_i) = n_i\pi_i$ e $Var(Y_i) = n_i\pi_i(1-\pi_i)$.

Porém, quando se verifica a presença de superdispersão dos dados, um procedimento comum é assumir valores maiores que 1 para o parâmetro de dispersão. Dessa maneira, a função de variância passa a ser escrita na forma $Var(Y_i) = \phi n_i\pi_i(1-\pi_i)$. Ocorre, no entanto, que a variável não mais possui distribuição binomial ou qualquer distribuição pertencente à família exponencial de distribuições.

2.1.6 Quase-Verossimilhança

Para os casos em que os dados não apresentam distribuição probabilística pertencente à família exponencial, como no caso de dados com superdispersão, mas em que se conhece a relação entre média e variância, Wedderburn (1974) propôs o método denominado Quase-Verossimilhança (QL), no qual a função QL é dada por

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{a(\phi)V(t)} dt. \tag{2.10}$$

Com relação aos parâmetros em μ_i, as estimativas são numericamente iguais às estimativas de máxima verossimilhança, considerando-se o parâmetro de dispersão ϕ fixo. Para ϕ, tipicamente, utiliza-se a estatística generalizada de Pearson dividida pelo respectivo número de graus de libredade como o estimador (McCullagh; Nelder, 1989):

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{2.11}$$

Para os modelos de quase-verossimilhança não se assume o conhecimento da distribuição da variável resposta, apenas especifica-se uma função de variância para a resposta assim como uma relação funcional entre a média e o vetor de parâmetros β. A distribuição da variável resposta é assumida conhecida quando a função de

variância especificada coincidir com a função de variância de alguma distribuição da família exponencial.

2.1.7 Distribuição Poisson Composta

A distribuição Poisson Composta, conforme descrita por Dunn e Smyth (2005), pode ser entendida como a distribuição de uma variável aleatória gerada por meio de um processo estocástico que, conforme Zhang (2012b), tem a forma

$$Y = \sum_{i=1}^{T} X_i, \quad T \sim Pois(\lambda), \quad X_i \overset{iid}{\sim} Gama(\alpha, \gamma), \quad T \perp X_i \tag{2.12}$$

Dessa forma, a distribuição conjunta de Y e T é dada por

$$f_{Y,T}(y,t) = \begin{cases} e^{-\lambda}, & (y,t) = (0,0) \\ \frac{y^{t\alpha-1}e^{-y/\gamma}}{\Gamma(t\alpha)\gamma^{t\alpha}} \frac{\lambda^t e^{-\lambda}}{t!}, & (y,t) \in \mathbb{R}_+ \times \mathbb{Z}_+ \end{cases} \tag{2.13}$$

A distribuição marginal de Y é, então, $f_Y(y) = \sum_{t=0}^{\infty} f_{Y,T}(y,t)$.

Tem-se, dessa forma, que dados gerados a partir da distribuição Poisson composta podem assumir valores discretos ($y = 0$) ou valores contínuos ($y > 0$). Esta distribuição pode, portanto, ser utilizada para analisar dados como, por exemplo, precipitação (em mm), valor pago por uma seguradora (em R$), ou massa (em mg).

2.1.8 Modelo exponencial de dispersão e distribuição Tweedie

O modelo exponencial de dispersão (MED), proposto por Jørgensen (1987), é uma família de distribuições de probabilidade que consiste de uma família exponencial com um parâmetro de dispersão adicional. Um MED pode ser caracterizado pela sua função de variância, que descreve a relação entre a média e a variância da distribuição quando o parâmetro de dispersão é constante.

Uma subclasse de interesse pertencente ao MED é aquela para a qual $V(\mu) = \mu^p$, para algum p, e que é denominada *família Tweedie de distribuições*. Esta classe de distribuições tem a forma

$$f(y|\theta,\phi) = a(y,\phi)\exp\left[\frac{y\theta - b(\theta)}{\phi}\right] \tag{2.14}$$

em que

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p}, & p \neq 1 \\ \ln(\mu), & p = 1 \end{cases} \tag{2.15}$$

e

$$b(\theta) = \begin{cases} \frac{\mu^{2-p}}{2-p}, & p \neq 2 \\ \ln(\mu), & p = 2 \end{cases} \tag{2.16}$$

uma vez que $b'(\theta) = \mu$ e $b''(\theta) = \mu^p$.

A classe de modelos Tweedie inclui, entre outras, algumas das distribuições associadas com modelos lineares generalizados tais como a Normal ($p = 0$), Poisson ($p = 1$), Gama ($p = 2$) e Gaussiana inversa ($p = 3$). A distribuição *Poisson composta* também pertence à classe Tweedie para $1 < p < 2$, com

$$a(y, \phi) = y^{-1} \sum_{t=1}^{\infty} \frac{y^t}{(p-1)^{t\alpha} \phi^{t(1+\alpha)} (2-p)^t t! \Gamma(t\alpha)} = y^{-1} \sum_{t=1}^{\infty} W_t \tag{2.17}$$

A expressão para $a(y, \phi)$ não possui forma analítica fechada, mas Dunn e Smyth (2005), apresentam métodos de aproximação numérica para achar um limite superior para a série infinita.

Além disso, os parâmetros da distribuição Poisson composta em (2.13), podem ser reescritos como

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)} \tag{2.18a}$$

$$\alpha = \frac{2-p}{p-1} \tag{2.18b}$$

$$\gamma = \phi(p-1)\mu^{p-1} \tag{2.18c}$$

Essa relação permite que observações de uma variável aleatória com distribuição Poisson composta sejam geradas conforme 2.12, fixando-se valores para μ, ϕ e p.

2.2 Modelos Lineares Generalizados Mistos

2.2.1 Introdução

A classe dos modelos lineares generalizados está inserida em um contexto de análise de dados em que os efeitos de fatores e covariáveis são denominados *efeitos fixos*

e em que se assume que as observações são independentes umas das outras. Entretanto, essa abordagem não é adequada quando existem estruturas de dependência entre as observações – por exemplo, quando as unidades observacionais/experimentais são aninhadas em uma unidade maior (bloco experimental, escola, hospital, etc.), denominados agrupamentos (*clusters*); ou em dados longitudinais, em que indivíduos são observados repetidamente ao longo do tempo.

Nestes casos, efeitos de *cluster* e de indivíduos podem ser introduzidos no modelo de análise para lidar com o impacto dessas estruturas no conjunto de dados. Esse tipo de efeito é chamado *aleatório*, uma vez que, ao contrário dos efeitos fixos, assume valores com uma distribuição de probabilidade. O modelo resultante da presença de efeitos fixos e aleatórios é um *modelo de efeitos mistos.*

Modelos de efeitos mistos para variáveis dependentes com distribuição normal passaram a ser intensamente desenvolvidos a partir do artigo de Laird e Ware (1982), no qual foi estruturado o modelo que serviu de base para posteriores extensões. Para outras distribuições, uma extensão do modelo linear generalizado de Nelder e Wedderburn (1972) foi popularizada por Breslow e Clayton (1993).

2.2.2 Definição dos MLG Mistos

Conforme Verbeke e Molenberghs (2005), seja uma amostra de N indivíduos (unidade observacional/experimental), $i = 1, ..., N$, cada um observado n_i vezes ($j = 1, ..., n_i$); tem-se que a resposta observada na j-ésima medição do indivíduo i, será denotada Y_{ij}. Dado um vetor de efeitos elemento-específicos q-dimensional $\mathbf{u}_i$, tem-se que os Y_{ij} são independentes e seguem uma distribuição da família exponencial:

$$Y_{ij}|\mathbf{u}_i \overset{i.i.d.}{\sim} f(y_{ij}|\mathbf{u}_i, \theta_{ij}, \phi) \tag{2.19}$$

em que a função de densidade é

$$f(y_{ij}|\mathbf{u}_i, \theta, \phi) = \exp\left[\frac{1}{a_{ij}(\phi)}[y_{ij}\theta_{ij} - b(\theta_{ij})] + c(y_{ij}, \phi)\right]. \tag{2.20}$$

Tem-se ainda, que

$$E(Y_{ij}|\mathbf{u}_i) = \mu_{ij} = b'(\theta_{ij}) \tag{2.21}$$

e

$$V(Y_{ij}|\mathbf{u}_i) = a_{ij}(\phi)b''(\theta_{ij}) = a_{ij}(\phi)V(\mu_{ij}). \tag{2.22}$$

Sejam, então, os vetores $\mathbf{x}_{ij}$, de dimensão $p \times 1$, e $\mathbf{z}_{ij}$, de dimensão $q \times 1$, de variáveis explicativas observadas referentes aos efeitos fixos e aleatórios, respectivamente. O preditor linear de um modelo de efeitos mistos para um indivíduo i na j-ésima observação, pode ser escrito como

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\top}\mathbf{u}_i \tag{2.23}$$

em que β é o vetor dos parâmetros de efeitos fixos. Supõe-se ainda, que $\mathbf{u}_i \sim N_q(\mathbf{0}, \mathbf{D})$, $\forall\, i = 1, ..., N$.

Assim, um modelo de efeitos mistos com um intercepto aleatório para cada indivíduo é descrito por

$$g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + u_i \tag{2.24}$$

com $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$.

2.2.3 Estimação

Ainda seguindo Verbeke e Molenberghs (2005), a estimação dos parâmetros $\boldsymbol{\beta}$ e $\mathbf{D}$ é feita via maximização da função de verossimilhança marginal, obtida da integração da distribuição condicional de $\mathbf{Y}_i|\mathbf{u}_i$, e é dada, para o i-ésimo indivíduo, por

$$f_i(y_i|\boldsymbol{\beta}, \mathbf{D}, \phi) = \int \cdots \int \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{u}_i, \boldsymbol{\beta}, \phi) f(\mathbf{u}_i|\mathbf{D}) d\mathbf{u}_i \tag{2.25}$$

Tendo-se que $\mathbf{u}_i \sim N_q(\mathbf{0}, \mathbf{D})$, a função de verossimilhança pode ser escrita como

$$L(\boldsymbol{\beta}, \mathbf{D}, \phi; \mathbf{Y}) \propto |\mathbf{D}|^{-N/2} \prod_{i=1}^{N} \int \cdots \int \exp\left\{ \sum_{j=1}^{n_i} l_i(\theta_{ij}; y_{ij}) - \frac{1}{2}\mathbf{u}_i^{\top}\mathbf{D}^{-1}\mathbf{u}_i \right\} d\mathbf{u}_i \tag{2.26}$$

em que $l_i(\theta_{ij}; y_{ij}) = \ln f_i(y_{ij}|\mathbf{u}_i, \boldsymbol{\beta}, \phi)$.

À exceção de alguns casos, como para $Y_{ij}|\mathbf{u}_i$ com distribuição normal ou para o modelo probit-normal-Bernoulli-beta como apresentado em Vieira (2008), as integrais acima são, de forma geral, intratáveis analiticamente, sendo necessário lançar mão de métodos numéricos de integração, como método de Laplace, quadratura de Gauss-Hermite ou, ainda, integração Monte Carlo (amostragem por importância, p. ex.). Entre outras abordagens de estimação dos parâmetros de um MLGM, podem ser citados: *Monte Carlo Expectation Maximization* (MCEM),

Monte Carlo Newton-Raphson e aproximação da verossimilhança (via linearização ou quase-verossimilhança penalizada), além de métodos bayesianos.

Abaixo, são descritas duas metodologias mais comumente encontradas nos *softwares* em que se tem implementados MLG's mistos, como *SAS* e *R*. São elas: Método da quase-verossimilhança penalizada e integração numérica (método de Laplace e quadratura de Gauss-Hermite).

2.2.4 Quase-Verossimilhança Penalizada

Breslow e Clayton (1993) utilizam uma abordagem de quase-verossimilhança penalizada (*Penalized Quasi-Likelihood*), que é definida pelos autores como se segue: Seja o logaritmo da quasi-verossimilhança correspondente ao i-ésimo indivíduo na j-ésima medição,

$$Q_{ij} = Q(\mu_{ij}; y_{ij}) = \phi^{-1} \int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij} - t}{V(t)} dt = -\frac{1}{2\phi} d_{ij} \tag{2.27}$$

em que d_{ij} é a medida de *deviance*, a função de quase-verossimilhança integrada é dada por

$$e^{ql(\beta,\mathbf{D})} \propto |\mathbf{D}|^{-1/2} \int \cdots \int \exp\left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} Q_{ij} - \frac{1}{2} \mathbf{u}_i^\top \mathbf{D}^{-1} \mathbf{u}_i \right\} d\mathbf{u}_i \tag{2.28}$$

Como a integral não possui forma analítica fechada, trabalha-se com uma aproximação para $PQL(\beta, \mathbf{u}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} Q_{ij} - \frac{1}{2} \mathbf{u}_i^\top \mathbf{D}^{-1} \mathbf{u}_i$, que é a quase-verossimilhança penalizada. Definindo-se o vetor de trabalho $\mathbf{Y}$ com elementos $Y_{ij} = \eta_{ij} + (y_{ij} - \mu_{ij}) g'(\mu_{ij})$, o sistema de equações do modelo misto, resultante da diferenciação do logaritmo da função de quase-verossimilhança penalizada com relação aos parâmetros e efeitos aleatórios, respectivamente, é dado por

$$\sum_i \sum_j \frac{y_{ij} - \mu_{ij}}{a_{ij}(\phi) V(\mu_{ij})} \cdot \frac{x_{ij}}{g'(\mu_{ij})} = 0 \tag{2.29}$$

$$\sum_i \sum_j \frac{y_{ij} - \mu_{ij}}{a_{ij}(\phi) V(\mu_{ij})} \cdot \frac{z_{ij}}{g'(\mu_{ij})} - \mathbf{D}^{-1} \mathbf{u} = 0 \tag{2.30}$$

Resolvendo-se o sistema, têm-se as estimativas dos parâmetros e os valores preditos dos efeitos aleatórios:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{Y} \tag{2.31}$$

$$\hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}^\top \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{2.32}$$

em que $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\mathbf{D}\mathbf{Z}^\top$ e $\mathbf{W}^{-1} = diag\left\{a_{ij}(\phi)V(\mu_{ij})[g'(\mu_{ij})]^2\right\}$

As estimativas das componentes de variância da matriz $\mathbf{D} = \mathbf{D}(\nu)$ são obtidas como solução para

$$-\frac{1}{2}\left[(\mathbf{Y} - \mathbf{X}\hat{\beta})^\top \mathbf{V}^{-1} \frac{\partial \mathbf{V}}{\partial \nu_k} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\beta}) - tr\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \nu_k}\right)\right] = 0 \tag{2.33}$$

em que $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}$.

2.2.5 Aproximação de Laplace e quadratura de Gauss-Hermite adaptativa

Aproximação de Laplace

O método de Laplace é comumente utilizado para aproximar integrais q-dimensionais da forma $\exp(Q(\mathbf{u}))$:

$$I = \int_{\mathbb{R}^q} e^{Q(\mathbf{u})} d\mathbf{u} \approx (2\pi)^{q/2} |-Q''(\hat{\mathbf{u}})|^{-1/2} e^{Q(\hat{\mathbf{u}})} \tag{2.34}$$

em que $\hat{\mathbf{u}}$ é a moda de $Q(\mathbf{u})$, isto é,

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u}} Q(\mathbf{u})$$

Bates (2010) considera um modelo com N observações e vetor de médias μ relacionado com o preditor linear pela forma

$$\eta(\mu) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} \tag{2.35}$$

em que $\boldsymbol{\beta}$ e b têm dimensão $p \times 1$ e $q \times 1$, respectivamente, e $\mathbf{X}$ e $\mathbf{Z}$ são as matrizes de delineamento associadas. Ainda, supõe-se $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D})$.

A matriz $\mathbf{D}$ dos componentes de variância pode ser reescrita como

$$\mathbf{D} = \phi \boldsymbol{\Lambda} \boldsymbol{\Lambda}^\top \tag{2.36}$$

em que Λ é uma matriz triangular inferior. Segue que (2.35) pode ser expressa por

$$\eta(\mu) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\Lambda}\mathbf{b} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u} \tag{2.37}$$

em que $u \sim N(\mathbf{0}, \phi\mathbf{I})$. Para calcular as modas condicionais dos efeitos aleatórios

$$\hat{\mathbf{u}} = \hat{\mathbf{u}}(\boldsymbol{\beta}, \phi, \boldsymbol{\Lambda}) = \arg\max_{\mathbf{u}} f(\mathbf{y}|\mathbf{u}, \boldsymbol{\beta}, \phi) f(\mathbf{u}|\phi)$$

necessários para a aproximação de Laplace, Bates (2010) propõe o algoritmo PIRLS (*Penalized Iteratively Reweighted Least Squares*). Este algoritmo, implementado na função *glmer* do pacote *lme4* da linguagem e ambiente de computação estatística **R** (R Development Core Team, 2012), é descrito, na r-ésima iteração:

1. $\boldsymbol{\eta}^{(r)} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^{(r)}$, $\boldsymbol{\mu}^{(r)} = g^{-1}(\boldsymbol{\eta}^{(r)})$;
2. $\mathbf{W}^{(r)} = \text{diag}\left\{[\phi V(\boldsymbol{\mu}^{(r)}) g'(\boldsymbol{\mu}^{(r)})^2]^{-1}\right\}$, $\mathbf{G}^{(r)} = \text{diag}\left\{g'(\boldsymbol{\mu}^{(r)})\right\}$;
3. $\mathbf{z}^{(r)} = \boldsymbol{\eta}^{(r)} + \mathbf{G}^{(r)}(\mathbf{y} - \boldsymbol{\mu}^{(r)})$;
4. $\mathbf{u}^{(r+1)} = \left(\mathbf{Z}^{*\top}\mathbf{W}^{(r)}\mathbf{Z}^* + \mathbf{I}\right)^{-1} Z^{*\top}\mathbf{W}^{(r)}\mathbf{z}^{(r)}$;
5. Repetir os passos 1 a 4 até convergência.

Tendo-se que

$$\text{Var}(\hat{\mathbf{u}}) = -E\left(\left.\frac{\partial^2 f(\mathbf{y}, \mathbf{u}|\boldsymbol{\beta}, \phi, \boldsymbol{\Lambda})}{\partial \mathbf{u} \partial \mathbf{u}^\top}\right|_{\mathbf{u}=\hat{\mathbf{u}}}\right)^{-1} = \phi\left(\mathbf{Z}^{*\top}\mathbf{W}\mathbf{Z}^* + \mathbf{I}\right)^{-1} = \phi\left(\mathbf{L}\mathbf{L}^\top\right)^{-1}$$

e fazendo as substituições na aproximação de Laplace, a expressão simplificada do logaritmo da funcção de verossimilhança é dada por

$$l(\mathbf{y}; \boldsymbol{\Theta}) \approx l(\mathbf{y}, \hat{\mathbf{u}}; \boldsymbol{\Theta}) + \frac{1}{2}\ln|\text{Var}(\hat{\mathbf{u}})| = \sum_{i=1}^{N} l(y_i; \boldsymbol{\Theta}, \hat{\mathbf{u}}) - \frac{\hat{\mathbf{u}}^\top\hat{\mathbf{u}}}{2\phi} - \frac{1}{2}\ln|\mathbf{L}|^2 \tag{2.38}$$

em que $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \phi, \boldsymbol{\Lambda})$.

Uma vez obtidas as estimativas pela maximização de (2.38), Bates (2010) apresenta a expressão da covariância das estimativas de $\boldsymbol{\beta}$ como resultado das expressões em (2.39)

$$\mathbf{L}\mathbf{L}^{\top} = \mathbf{Z}^{*\top}\mathbf{W}\mathbf{Z}^{*} + \mathbf{I}^{-1} \tag{2.39a}$$

$$\mathbf{L}\mathbf{R}_{\mathbf{XZ}}^{\top} = \mathbf{Z}^{*\top}\mathbf{W}^{1/2} \tag{2.39b}$$

$$\mathbf{R}_{\mathbf{X}}\mathbf{R}_{\mathbf{X}}{}^{\top} = \mathbf{X}^{\top}\mathbf{X} - \mathbf{R}_{\mathbf{XZ}}\mathbf{R}_{\mathbf{XZ}}^{\top} \tag{2.39c}$$

$$Cov(\hat{\boldsymbol{\beta}}) = \phi(\mathbf{R}_{\mathbf{X}}\mathbf{R}_{\mathbf{X}}{}^{\top})^{-1} \tag{2.39d}$$

Quadratura de Gauss-Hermite adaptativa

O método de integração denominado quadratura de Gauss-Hermite (Abramowitz; Stegun, 1972) é uma técnica numérica na qual a integral de interesse é aproximada por uma soma ponderada de valores obtidos da avaliação do integrando em certos pontos, mais especificamente nas raízes do polinômio de Hermite de ordem L. Seja então uma função $f(x)$ que deve ser integrada no conjunto dos números reais $\mathbb{R}$, tem-se

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} g(x)e^{-x^2}dx \approx \sum_{l=1}^{L} w_l g(\xi_l) \tag{2.40}$$

em que $g(x) = f(x)e^{x^2}$. Os nós ξ_l e os pesos w_l correspondentes podem ser encontrados em Abramowitz e Stegun (1972), mas também estão amplamente disponíveis em *softwares* estatísticos.

Para integrais multivariadas, o somatório pode ser aplicado da seguinte forma:

$$\begin{aligned}
\int f(\mathbf{x})d\mathbf{x} &= \int \cdots \int f(x_1, ..., x_q)dx_1...dx_q \\
&= \int g(x_1, ..., x_q)e^{-\mathbf{x}^{\top}\mathbf{x}}dx_1...dx_q \\
&\approx \sum_{l_1=1}^{L} w_{l_1}^{(1)} \cdots \sum_{l_q=1}^{L} w_{l_q}^{(q)} g(\xi_{l_1}^{(1)}, ..., \xi_{l_q}^{(q)})
\end{aligned}$$

em que $\xi_{l_j}^{(j)}$ e $w_{l_j}^{(j)}$ $(l = 1, ..., L$ e $j = 1, ..., q)$ são os nós e pesos de uma quadratura de Gauss-Hermite com L pontos na j-ésima coordenada de $\mathbf{x}$[1].

Uma versão aprimorada do método de quadratura descrito, é o denominado método de quadratura de Gauss-Hermite adaptativo (Liu; Pierce, 1994). Considerando

[1] P. ex., para $q = 2$ e $L = 2$,

$\hat{x}$, a moda de uma função $f(x)$, como aquela em (2.40), e $\hat{\tau}^2$ a curvatura estimada de $f(x)$ em $\hat{x}$, isto é,

$$\hat{\tau}^2 = \left[\left(-\frac{\partial^2 f(x)}{\partial x^2}\right)^{-1}\right]_{x=\hat{x}} \tag{2.41}$$

então

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{f(x)}{\phi(x;\hat{x},\hat{\tau}^2)}\phi(x;\hat{x},\hat{\tau}^2)dx = \int_{-\infty}^{\infty} h(x)\phi(x;\hat{x},\hat{\tau}^2)dx$$

em que $\phi(\cdot;\hat{x},\hat{\tau}^2)$ é a função de densidade da distribuição normal com média $\hat{x}$ e variância $\hat{\tau}^2$. Sob a reparametrização $z = (x-\hat{x})/\sqrt{2\hat{\tau}^2}$ a integral passa a ser

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} h(\hat{x}+\sqrt{2\hat{\tau}^2}z)\frac{1}{\sqrt{\pi}}e^{-z^2}dz \approx \sum_{l=1}^{L}\frac{w_l}{\sqrt{\pi}}h(\hat{x}+\sqrt{2\hat{\tau}^2}\xi_l).$$

Portanto, para um modelo

$$g(\mu_{ij}) = \eta_{hj} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + u_i, \quad u_i \overset{i.i.d.}{\sim} N(0,\sigma_u^2) \tag{2.42}$$

tem-se que a contribuição do i-ésimo indivíduo (ou *cluster*) pode ser aproximada por

$$\int_{-\infty}^{\infty} f(\mathbf{y}_i,u_i|\boldsymbol{\beta},\phi,\sigma^2)du_i \approx \sqrt{2\hat{\tau}^2}\sum_{l=1}^{L} w_l^* f(\mathbf{y}_i,\hat{u}_i+\sqrt{2\hat{\tau}^2}\xi_l|\boldsymbol{\beta},\phi,\sigma^2) \tag{2.43}$$

em que $w_l^* = w_l \exp(\xi_l^2)$ e

$$f(\mathbf{y}_i,u_i|\boldsymbol{\beta},\phi,\sigma^2) = \left[\prod_{j=1}^{n_i} f(y_{ij}|u_i,\boldsymbol{\beta},\phi)\right]\frac{e^{-u_i^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}. \tag{2.44}$$

O processo de maximização da função de verossimilhança condicional para calcular as modas $\hat{u}_i$ pode ser feito utlizando-se algum algoritmo numérico, como o

---

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{x})d\mathbf{x} \approx w_1^{(1)}\left[w_1^{(2)}g\left(\xi_1^{(1)},\xi_1^{(2)}\right) + w_2^{(2)}g\left(\xi_1^{(1)},\xi_2^{(2)}\right)\right] + w_2^{(1)}\left[w_1^{(2)}g\left(\xi_2^{(1)},\xi_1^{(2)}\right) + w_2^{(2)}g\left(\xi_2^{(1)},\xi_2^{(2)}\right)\right]$$

PIRLS, no qual a cada iteração tem-se valores fixados de $\boldsymbol{\beta}$, ϕ e σ^2, permitindo o cálculo aproximado da integral, como na Equação (2.43).

2.2.6 Diagnóstico de MLGs mistos

Como forma de avaliar a qualidade do ajuste de modelos lineares generalizados mistos, Vieira *et al.* (2000) e Hall e Wang (2005) sugerem a utlização de gráficos meio-normais com envelopes simulados para os resíduos do modelo estimado. O princípio geral é o de avaliar, utilizando resultados de simulações a partir do modelo estimado, se os resíduos observados são consistentes com aqueles produzidos pelo processo gerador descrito pelo modelo.

A construção de gráficos meio-normais com envelopes simulados para MLGs, conforme proposta de Aitkinson (1985) *apud* Yang e Sun (2006), pode também ser aplicada aos modelos lineares generalizados mistos e pode ser descrito por:

1. Ajuste do modelo e simulação de amostra com N observações da variável resposta a partir dos valores ajustados;
2. Reajuste do modelo para a amostra simulada e cálculo dos valores absolutos ordenados dos resíduos[2] obtidos;
3. Repetir (1) e (2) B vezes;
4. Para os N conjuntos de estatísticas de ordem com B elementos, calcular média, quantil α e $1-\alpha$;
5. Geração do gráfico: Plotar os resíduos do modelo original e os valores em (4) contra os escores meio-normais $\Phi^{-1}((t+n-1/8)/(2n+1/2))$.

Os quantis α e $1-\alpha$ das estatísticas de ordem resultam no envelope. A probabilidade de que um ponto se encontre fora desse envelope é de $1-2\alpha$. Além disso, se uma quantidade considerável de pontos está fora do envelope, tem-se a indicação de problemas no ajuste do modelo.

[2]Hall e Wang (2005) sugerem a utilização de resíduos de Pearson

Capítulo 3

MLGM Multivariado

3.1 Introdução

Para lidar com dados logitudinais com resposta multivariada, ainda é possível utilizar modelos mistos na modelagem conjunta das respostas (Fieuws *et al.*, 2006). Considera-se que para um conjunto de $k = 1, ..., m$ variáveis resposta medidas para N indivíduos ($i = 1, ..., N$) observados em n_i momentos diferentes ($j = 1, ..., n_i$), é possível especificar modelos da forma

$$g(\mu_{ijk}) = \eta_{ijk} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta}_k + \mathbf{z}_{ijk}^{\top}\mathbf{u}_{ik}. \tag{3.1}$$

Assume-se também que o caráter multivariado das observações pode ser tratado especificando-se uma distribuição conjunta para o vetor de efeitos aleatórios. Então, tipicamente, se para os m modelos há apenas um intercepto aleatório tal que $\mathbf{u}_i = (u_{i1}, ..., u_{im})^{\top}$, tem-se que

$$\begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{im} \end{pmatrix} \overset{iid}{\sim} N_m \left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \delta_1^2 & \delta_{12} & \cdots & \delta_{1m} \\ \delta_{21} & \delta_2^2 & \cdots & \delta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{m1} & \delta_{m2} & \cdots & \delta_m^2 \end{pmatrix} \right) \tag{3.2}$$

Denotando-se por $\boldsymbol{\Theta}^*$ o vetor contendo todos os parâmetros (efeitos fixos e parâmetros de covariância) – isto é, $\boldsymbol{\Theta}^* = (\boldsymbol{\beta}, \delta_1^2, \delta_{12}, ..., \delta_m^2)$ –, a contribuição do indivíduo i na função de verossimilhança para o modelo conjunto é escrito como

$$l_i(\boldsymbol{\Theta}^*|\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, ..., \mathbf{Y}_{im}) = \int_{\mathbb{R}^m} \left[\prod_{k=1}^{m}\prod_{j=1}^{n_i} f(y_{ijk}|\mathbf{x}, \mathbf{u}_i, \boldsymbol{\Theta})\right] \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(2\pi)^{m/2}} e^{-\frac{1}{2}\mathbf{u}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}_i} d\mathbf{u}_i \tag{3.3}$$

3.2 Aspectos computacionais

Embora o modelo (3.1) seja descrito como um modelo multivariado, sua forma bastante similar aos modelos univariados apresentados nas seções anteriores. De ato, se para cada indivíduo considera-se uma matriz de respostas $\mathbf{Y}_i$ com dimensão $_i \times m$ dada por

$$\mathbf{Y}_i = \begin{bmatrix} y_{i11} & \cdots & y_{i1m} \\ \vdots & \ddots & \vdots \\ y_{in_i1} & \cdots & y_{in_im} \end{bmatrix}_{n_i \times m}$$

uja correspondente matriz de médias $\boldsymbol{\mu}_i$ está relacionada com o preditor por meio le uma função de ligação $g(\cdot)$ na forma

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{X}_i\mathbf{B} + \mathbf{Z}_i\mathbf{U} \tag{3.4}$$

m que

$$\mathbf{X}_i = \begin{bmatrix} x_{i11} & \cdots & x_{i1p} \\ \vdots & \ddots & \vdots \\ x_{in_i1} & \cdots & x_{in_ip} \end{bmatrix}_{n_i \times p}$$

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{m1} \\ \vdots & \ddots & \vdots \\ \beta_{1p} & \cdots & \beta_{mp} \end{bmatrix}_{p \times m}$$

$$\mathbf{Z}_i = \mathbf{1}_{n_i} \otimes \mathbf{e}_i^\top = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}_{n_i \times N}$$

$$\mathbf{U} = \begin{bmatrix} u_{11} & \cdots & u_{m1} \\ \vdots & \ddots & \vdots \\ u_{1p} & \cdots & u_{mp} \end{bmatrix}_{N \times m},$$

então, se $\mathbf{Y}_i$ for reescrito de forma empilhada, tal que

$$\mathbf{Y}_i^* = vec(\mathbf{Y}_i) = \begin{bmatrix} y_{i11} & \cdots & y_{in_i1} & \cdots & y_{i1m} & \cdots & y_{in_im} \end{bmatrix}^\top_{n_im \times 1}$$

e reescrevendo-se também $\mathbf{X}_i$, $\mathbf{B}$, $\mathbf{Z}_i$ e $\mathbf{U}$ na forma

$$\mathbf{X}_i^* = \mathbf{I}_m \otimes \mathbf{X}_i = \begin{bmatrix} \mathbf{X}_i & & \\ & \ddots & \\ & & \mathbf{X}_i \end{bmatrix}_{n_im \times pm}$$

$$\boldsymbol{\beta} = vec(\mathbf{B}) = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1p} & \cdots & \beta_{m1} & \cdots & \beta_{mp} \end{bmatrix}^\top_{pm \times 1}$$

$$\mathbf{Z}_i^* = \mathbf{I}_m \otimes \mathbf{Z}_i = \begin{bmatrix} \mathbf{Z}_i & & \\ & \ddots & \\ & & \mathbf{Z}_i \end{bmatrix}_{n_im \times Nm}$$

$$\mathbf{u} = vec(\mathbf{U}) = \begin{bmatrix} u_{11} & \cdots & u_{1N} & \cdots & u_{m1} & \cdots & u_{mN} \end{bmatrix}^\top_{mN \times 1}$$

tem-se um modelo com um formato univariado

$$g(\boldsymbol{\mu}_i^*) = \boldsymbol{\eta}_i^* = \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{Z}_i^*\mathbf{u} \tag{3.5}$$

A vantagem do formato apresentado é a possibilidade de inserir um modelo multivariado em um contexto de modelos univariados, permitindo que os programas usuais para estimação de modelos mistos sejam utilizados, necessitando apenas de algumas adaptações na especificação. Entretanto, conforme a dimensionalidade dos vetores de efeitos aleatórios aumenta, podem surgir problemas computacionais que tornariam proibitiva a utilização desta abordagem.

3.3 Estimação por pares

A estimação de modelos mistos não é uma tarefa simples, já que a função de verossimilhança marginal não possui forma fechada e deve ser integrada. Para modelos multivariados, em que se especifica um vetor de efeitos aleatórios para cada

indivíduo, a estimação é ainda mais computacionalmente intensiva.

Levando-se em consideração casos em que se tem alta dimensionalidade, para os quais a estimação dos modelos seria computacionalmente muito difícil, Fieuws e Verbeke (2006) introduziram uma abordagem de modelagem conjunta por pares (ou par-a-par) das respostas de forma a reduzir a complexidade do problema. A idéia é que se realize o ajuste de todos os $m(m-1)/2$ pares de modelos bivariados separadamente via maximização do logaritmo das funções de verossimilhança

$$\sum_{i=1}^{N} l_{rsi}(\boldsymbol{\Theta}_{rs}|\mathbf{Y}_{ri}, \mathbf{Y}_{si}) \tag{3.6}$$

com $r = 1, ..., m-1$, $s = r+1, ..., m$, tendo-se que $\boldsymbol{\Theta}_{rs}$ é o vetor de todos os parâmetros do modelo misto bivariado para o par (r, s) de variáveis resposta.

Uma vez estimados os parâmetros de todos os pares, todos os vetores pares-específicos são empilhados em um único vetor $\hat{\boldsymbol{\Theta}}$. Tem-se que alguns parâmetros do vetor $\boldsymbol{\Theta}^*$ terão múltiplos correspondentes no vetor $\hat{\boldsymbol{\Theta}}$. Uma estimativa única destes parâmetros é obtida pela média das estimativas contidas em $\hat{\boldsymbol{\Theta}}$.

3.4 Inferência para $\boldsymbol{\Theta}$ e $\boldsymbol{\Theta}^*$

Empregando-se idéias de estimação por pseudo-verossimilhança (Besag, 1975), constrói-se uma matriz de covariâncias para os elementos de $\hat{\boldsymbol{\Theta}}$. A proposta da abordagem de pseudo-verossimilhança é substituir a verossimilhança conjunta por um produto de densidades marginais ou condicionais de tal forma que este produto seja computacionalmente mais tratável. De fato, a estimação por pares descrita acima é equivalente à maximização da função de pseudo-verossimilhança da forma

$$pl(\boldsymbol{\Theta}) = l(\boldsymbol{\Theta}_{1,2}|\mathbf{Y}_1, \mathbf{Y}_2) + l(\boldsymbol{\Theta}_{1,3}|\mathbf{Y}_1, \mathbf{Y}_3) + \cdots + l(\boldsymbol{\Theta}_{m-1,m}|\mathbf{Y}_{m-1}, \mathbf{Y}_m) = \sum_{p=1}^{P} l_p(\boldsymbol{\Theta}_p) \tag{3.7}$$

em que $p = 1, ..., P$, com $P = m(m-1)/2$. Das propriedades de estimadores de pseudo-verossimilhança, segue que $\hat{\boldsymbol{\Theta}}$ tem distribuição dada por

$$\sqrt{N}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \sim N(\mathbf{0}, \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}) \tag{3.8}$$

em que $\mathbf{J}$ é uma matriz bloco diagonal e $\mathbf{K}$ é uma matriz simétrica. Os blocos J_{pp} e K_{pq} são expressos como

$$J_{pp} = -\frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\partial^2 l_{pi}}{\partial\theta_p \partial\theta_p'}\right),$$

$$K_{pq} = -\frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\partial l_{pi}}{\partial\theta_p}\frac{\partial l_{qi}}{\partial\theta_q'}\right)$$

$p, q = 1, ..., P$. As estimativas de $\mathbf{J}$ e $\mathbf{K}$ são obtidas abstraindo-se as esperanças e substituindo os parâmetros pelos valores estimados em $\hat{\boldsymbol{\Theta}}$.

Finalmente, obtém-se $\hat{\boldsymbol{\Theta}}^*$ tomando a forma $\hat{\boldsymbol{\Theta}}^* = A\hat{\boldsymbol{\Theta}}$. Dessa forma, $\hat{\boldsymbol{\Theta}}^*$ tem distribuição normal multivariada com média $\boldsymbol{\Theta}^*$ e matriz de covariâncias $A\Sigma(\hat{\boldsymbol{\Theta}})A$, tal que A é uma matriz de pesos adequada e $\Sigma(\hat{\boldsymbol{\Theta}})$ é a matriz de covariâncias de $\hat{\boldsymbol{\Theta}}$.

3.5 Verossimilhança em modelos estimados par-a-par

Conforme exposto anteriormente, a estimação par-a-par do modelo multivariado equivale à maximização da função de pseudo-verossimilhança. Embora utilizada na estimação dos parâmetros, porém, Fieuws *et al.* (2006) não consideram adequada a utilização de estatísticas baseadas na pseudo-verossimilhança para seleção de modelos, por exemplo. Dessa forma, os autores propõem que o logartimo da função de verossimilhança do modelo multivariado estimado seja calculado a partir da soma dos logaritmos das funções de verossimilhanças marginais para cada indivíduo, isto é,

$$l^P = \sum_{i=1}^{N} \ln\left[\int_{-\infty}^{\infty} f(\mathbf{y}_i|\mathbf{u}_i, \hat{\boldsymbol{\beta}}, \hat{\phi}) f(\mathbf{u}_i|\hat{\boldsymbol{\Sigma}}) d\mathbf{u}_i\right] \tag{3.9}$$

Utilizando uma aproximação da integral via método Monte Carlo, tem-se, então,

$$l^P = \sum_{i=1}^{N} \ln\left[\frac{\sum_{r=1}^{R} f(\mathbf{y}_i|\mathbf{u}^{(r)}, \hat{\boldsymbol{\beta}}, \hat{\phi})}{R}\right] \tag{3.10}$$

em que $\mathbf{u}^{(r)}$, com $r = 1, ..., R$, é um vetor m-dimensional da distribuição $f(\mathbf{u}_i|\hat{\boldsymbol{\Sigma}})$.

Capítulo 4

Análise de dados de algodão – motivação

4.1 Introdução

Em resposta ao ataque de herbívoros, é comum que plantas aumentem a liberação de um certo conjunto de compostos orgânicos voláteis. Esses compostos podem servir como indicação da presença desses herbívoros para seus predadores, ou, por outro lado, pode ser utilizado por outros indivíduos para encontrar coespecíficos (Magalhães *et al.*, 2012; Hare, 2011).

No caso de plantas de algodão, Magalhães *et al.* (2012) verifica que insetos da espécie *Anthonomus grandis* são atraídos de forma mais intensa por plantas danificadas por indivíduos da mesma espécie que por plantas não danificadas, e que, ainda, não exibem preferência por plantas danificadas mecanicamente ou por insetos das espécies *S. frugiperda* e *E. heros* com relação a plantas não danificadas.

É importante, portanto, que se tente identificar compostos que contribuam na diferenciação dos perfis de compostos de plantas danificadas pelo *A. grandis* e plantas não danificadas ou danificadas por outros insetos. Uma aplicação de interesse seria a confecção de armadilhas ecologicamente amigáveis no controle do *A. grandis* em plantações de algodão (Magalhães *et al.*, 2012).

4.2 Descrição do experimento

A seguir, são analisados dados de perfil químico de algodão provenientes de um experimento aleatorizado longitudinal com plantas de algodão em estado reprodu-

tivo[1] no qual as plantas foram submetidas a um de cinco tratamentos de interesse (Controle, *A. grandis*, *E. heros*, *S. frugiperda* e dano mecânico), cada um com 8 repetições. Foram medidas, em 4 tempos distintos (24, 48, 72 e 96 horas após aplicação do tratamento), as massas (em μg) de 25 compostos químicos, que, para cada indivíduo em um determinado tempo, constituem um vetor de variáveis resposta.

Como motivação para a análise do conjunto completo dos dados, foram escolhidos 3 compostos, que são analisados a seguir sob três abordagens semelhantes, mas com graus de complexidade de especificação distintos: modelo linear generalizado univariado, MLG misto univariado e MLG misto multivariado.

4.3 Análise exploratória

Para análise inicial dos dados, são selecionados os compostos α-Pinene, β-Myrcene e β-Ocimene, que serão denominados C1, C5 e C8, respectivamente. Na Tabela 4.1 abaixo tem-se algumas medidas-resumo dos dados de massa (em $10^3\mu g$) por composto e tratamento. Observa-se que todos os valores são maiores ou iguais a zero e que em alguns casos há grande variabilidade nas respotas observadas, o que também pode ser observado na Figura 4.1.

Tabela 4.1: Medidas-resumo da variável massa, por tratamento e por composto

Composto	Trt	Média	Mediana	Mínimo	Q75%-Q25%	Máximo
C1	Controle	4.228	2.993	0.000	3.836	12.149
	frugiperda	13.252	10.233	1.242	11.677	50.823
	grandis	14.816	13.431	0.000	9.816	59.892
	heros	7.131	4.540	0.025	7.820	54.081
	mecânico	9.259	5.141	0.465	6.286	77.420
C5	Controle	1.104	0.649	0.098	0.814	4.837
	frugiperda	4.610	3.611	0.168	5.663	14.393
	grandis	7.464	3.332	0.000	9.256	55.769
	heros	2.519	1.464	0.048	2.647	12.976
	mecânico	3.623	1.484	0.137	4.535	22.491
C8	Controle	0.686	0.372	0.000	0.488	2.625
	frugiperda	5.133	3.833	0.102	6.743	18.377
	grandis	14.059	6.230	0.000	17.154	76.748
	heros	2.333	0.810	0.000	1.446	16.673
	mecânico	1.449	0.617	0.116	0.951	14.596

[1]Experimento realizado no Laboratório de Semioquímicos da unidade de Recursos Genéticos e Biotecnologia da EMBRAPA. Os dados foram cedidos pela pesquisadora Dra. Maria Carolina Blassioli Moraes para análise nesta dissertação.

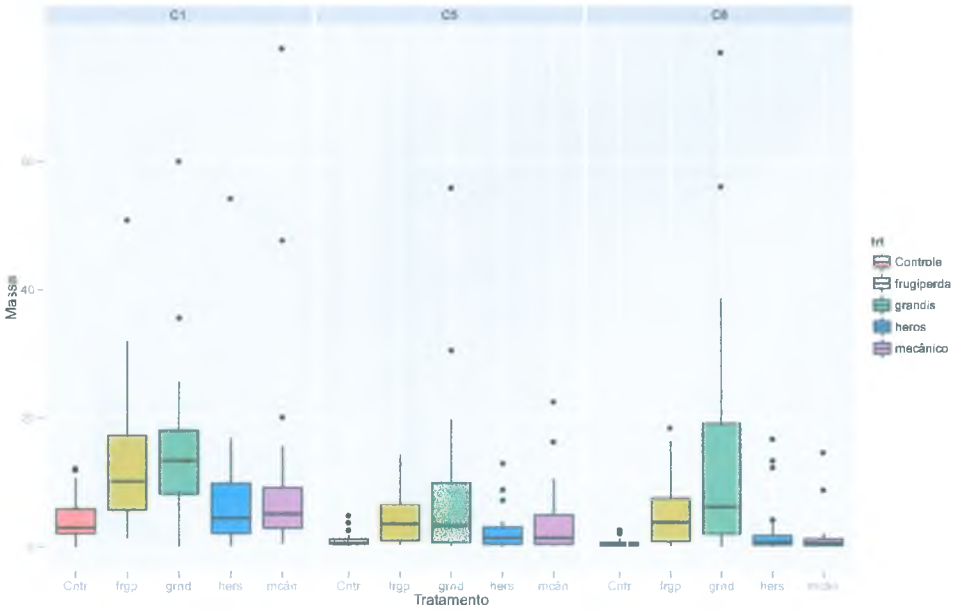

Figura 4.1: Boxplot da variável massa, por tratamento e por composto

A Figura 4.2 apresenta um histograma da distribuição dos dados por composto, no qual é constatado assimetria na distribuição dos dados nos três compostos. A Figura 4.3 mostra os gráficos de perfil das médias de cada composto por tratamento e tempo.

Uma vez que se tem um vetor de observações para cada indivíduo e, ainda, que são feitas medições ao longo do tempo, deseja-se verificar também se há correlação entre os compostos e entre as observações no tempo. Para isso, apresenta-se nas Figuras 4.4 e 4.5, gráficos de matriz de dispersão dos dados, em escala logarítmica.

Nota-se em ambos os gráficos que há certa correlação entre os grupos analisados. As matrizes de correlação calculadas abaixo confirmam estas relações:

$$\hat{\rho}_{Composto} = \begin{pmatrix} 1.000 & 0.426 & 0.479 \\ 0.426 & 1.000 & 0.724 \\ 0.479 & 0.724 & 1.000 \end{pmatrix}$$

e

$$\hat{\rho}_{Tempo} = \begin{pmatrix} 1.000 & 0.649 & 0.616 & 0.316 \\ 0.649 & 1.000 & 0.565 & 0.287 \\ 0.616 & 0.565 & 1.000 & 0.261 \\ 0.316 & 0.287 & 0.261 & 1.000 \end{pmatrix}$$

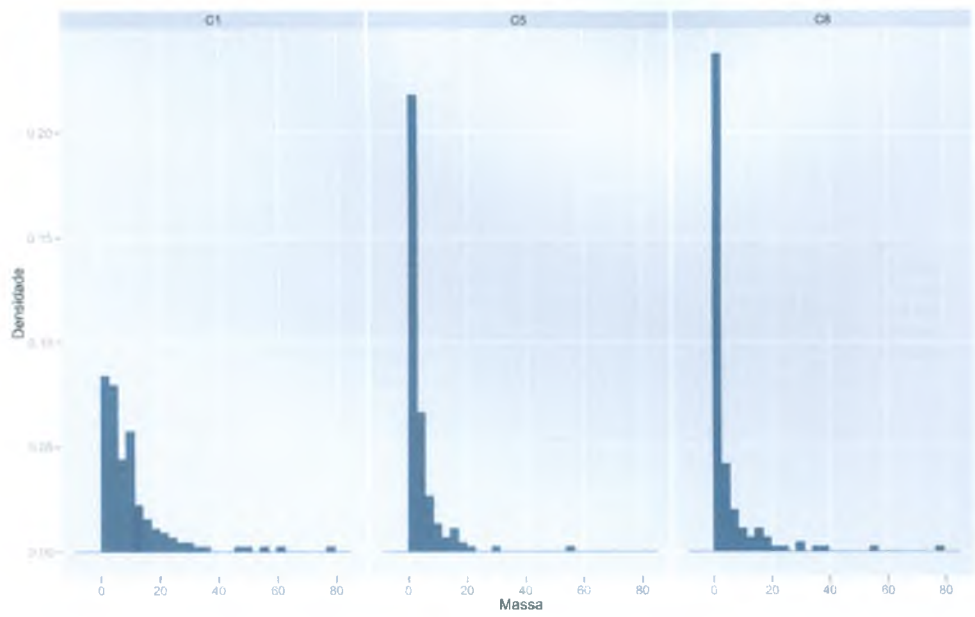

Figura 4.2: Histograma da distribuição da massa por composto

4.4 Análise dos dados

Para a modelagem da massa dos compostos (em $10^3 \mu g$) como função dos tratamentos e tempos de observação. são utilizadas 3 abordagens diferentes: Modelo linear generalizado univariado sem efeito aletório; MLG misto univariado e MLG misto multivariado. Além disso, em razão da natureza dos dados – valores contínuos positivos, mas com algumas observações iguais a zero – é assumida uma distribuição Poisson composta para a variável resposta.

As três abordagens de modelagem propostas levam em consideração diferentes pressupostos, e, com isto, busca-se verificar o impacto da adoção de modelos mais complexos na análise dos dados. Na primeira abordagem, MLGs univariados, tratam-se as observações ao longo tempo e dos diferentes compostos como independentes. Na segunda abordagem. MLGs mistos univariados, ainda é assumida a independência entre os compostos, mas são introduzidos efeitos aleatórios que modelam a dependência entre medições ao longo do tempo. Na terceira abordagem, efeitos aleatórios multivariados permitem que dependências tanto de medidas repetidas, como entre compostos sejam tratadas.

Abordagem 1: MLG univariado

Na primeira abordagem utilizada para modelagem dos dados, ajusta-se um MLG com distribuição Poisson composta e função de ligação logarítmica para cada com-

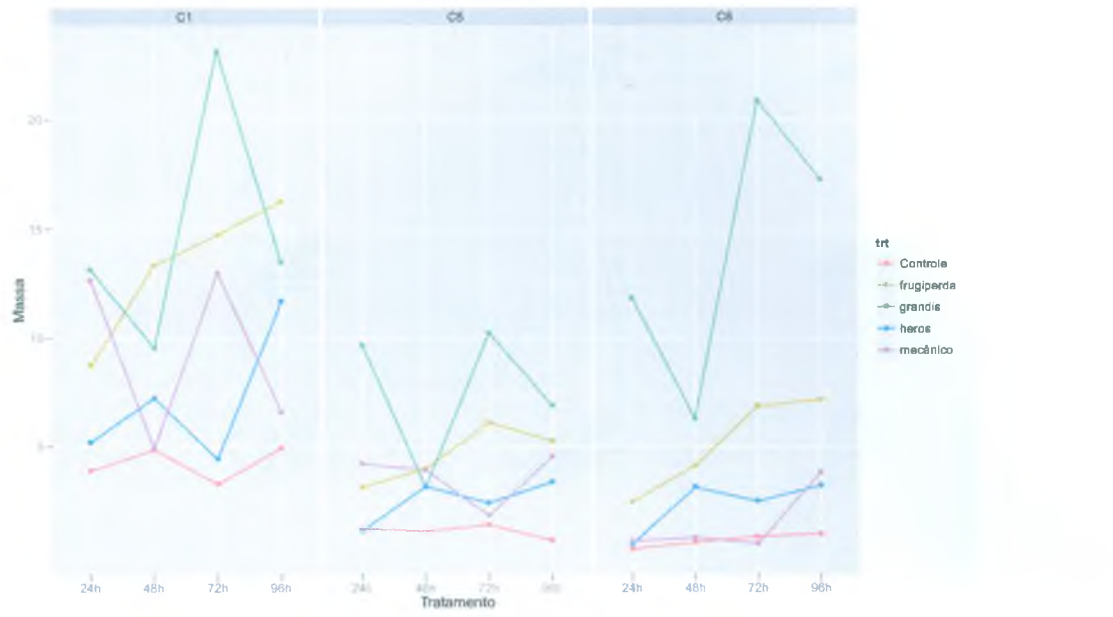

Figura 4.3: Gráfico de perfil das médias por composto, tratamento e tempo

posto j. Tem-se, portanto, a especificação:

$$Y_{jkt} \sim PC(\mu_{jkt}, \phi_j, p_j) \tag{4.1}$$

$$\ln(\mu_{jkt}) = \alpha_{jk} + \beta_j t + \gamma_{jk} t \tag{4.2}$$

em que

- α_{jk}: Efeito do k-ésimo tratamento para o composto j;
- β_j: Coeficiente angular para a variável tempo para o composto j;
- γ_{jk}: Efeito de interação entre tratamento e tempo para o composto j;

Sob esta abordagem, supõe-se que as observações ao longo tempo são independentes, bem como se supõe independência entre os compostos. Assim, pode-se escrever a função de verossimilhança para o composto j como

$$l_j(\Theta_j|\mathbf{x}_j, \mathbf{y}_j) = \prod_{i=1}^{40} \prod_{t=1}^{4} f(y_{ijt}|\mathbf{x}_j, \Theta_j) \tag{4.3}$$

em que $f(y_{ijt}|\mathbf{x}_j, \Theta_j)$ denota a função de densidade da distribuição Poisson composta, conforme (2.14)–(2.17), e $\Theta_j = (\alpha_{j1}, ..., \alpha_{j5}, \beta_j, \gamma_{j1}, ..., \gamma_{j5})$ é o vetor de parâmetros do modelo.

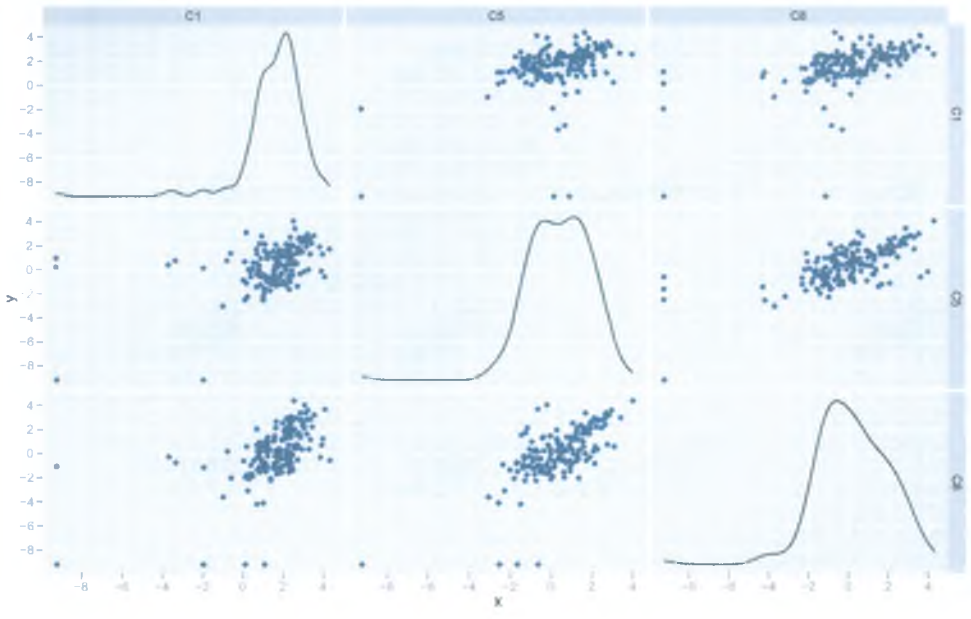

Figura 4.4: Matriz de dispersão da variável massa por composto

A seleção dos modelos univariados é feita por análise de deviance, cujos resultados são apresentados a seguir, na Tabela 4.2. Os P-valores são obtidos via *bootstrap* paramétrico.

Tabela 4.2: Análise de deviance para os modelos concorrentes – Modelos univariados

Composto	Modelo	GL	AIC	Deviance	GL_{X^2}	X^2_{obs}	P-Valor
C1	$\alpha_{1k} + \beta_1 t + \gamma_{1k} t$	150	1052.3	277.22			
	$\alpha_{1k} + \beta_1 t$	154	1048.1	280.93	4	3.71	0.472
	α_{1k}	155	1048.2	282.85	1	1.92	0.158
C5	$\alpha_{2k} + \beta_2 t + \gamma_{2k} t$	150	740.65	276.62			
	$\alpha_{2k} + \beta_2 t$	154	736.08	281.36	4	4.75	0.500
	α_{2k}	155	734.57	282.02	1	0.65	0.472
C8	$\alpha_{3k} + \beta_3 t + \gamma_{3k} t$	150	693.48	298.89			
	$\alpha_{3k} + \beta_3 t$	154	687.95	302.63	4	3.75	0.695
	α_{3k}	155	707.17	336.31	1	33.68	< 0.002

Com base nos resultados da análise deviance e considerando um nível de significância de 5%, selecionam-se, para os compostos 1 e 5, o modelo apenas com efeito de tratamento, enquanto para o composto 8, seleciona-se o modelo com efeito de tratamento e de tempo. As estimativas dos modelos finais são apresentadas na Tabela 4.3, juntamente com os erros-padrão, entre parênteses. Para o composto 8, nos períodos observados, tem-se que, em média, o efeito de tempo resulta em um au-

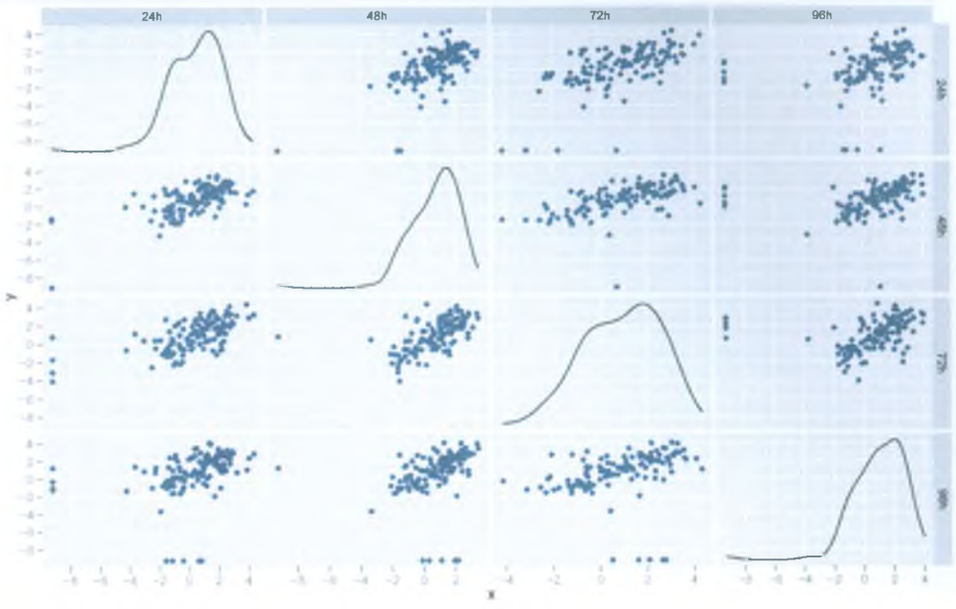

Figura 4.5: Matriz de dispersão da variável massa por tempo

mento de aproximadamente 1.6% ($e^{0.016}$) na massa desse composto a cada 24 horas. Os efeitos médio estimados são apresentados na Figura 4.6.

Como método para avaliar a qualidade do ajuste dos dados, utilizam-se gráficos meio-normais dos resíduos de Pearson com envelopes simulados, com $n = 100$ simulações e quantis de 98% para os envelopes (Figura 4.7). Considera-se que há um ajuste razoável dos dados, ainda que para os três compostos, haja pontos fora dos envelopes.

Tabela 4.3: Estimativas e erros padrão dos parâmetros dos MLG univariados

Parâmetro	Composto		
	C1	C5	C8
α_{j1} (Controle)	1.442 (0.223)	0.099 (0.237)	-1.406 (0.393)
α_{j2} (Frugiperda)	2.584 (0.188)	1.528 (0.212)	0.614 (0.355)
α_{j3} (Grandis)	2.696 (0.185)	2.010 (0.204)	1.671 (0.339)
α_{j4} (Heros)	1.964 (0.206)	0.924 (0.222)	-0.179 (0.368)
α_{j5} (Mecânico)	2.226 (0.198)	1.287 (0.216)	-0.757 (0.379)
β_j	–	–	0.016 (0.004)
ϕ_j	1.488 (0.171)	1.425 (0.131)	1.508 (0.137)
p_j	1.702 (0.049)	1.842 (0.033)	1.800 (0.029)

Abordagem 2: MLG misto univariado

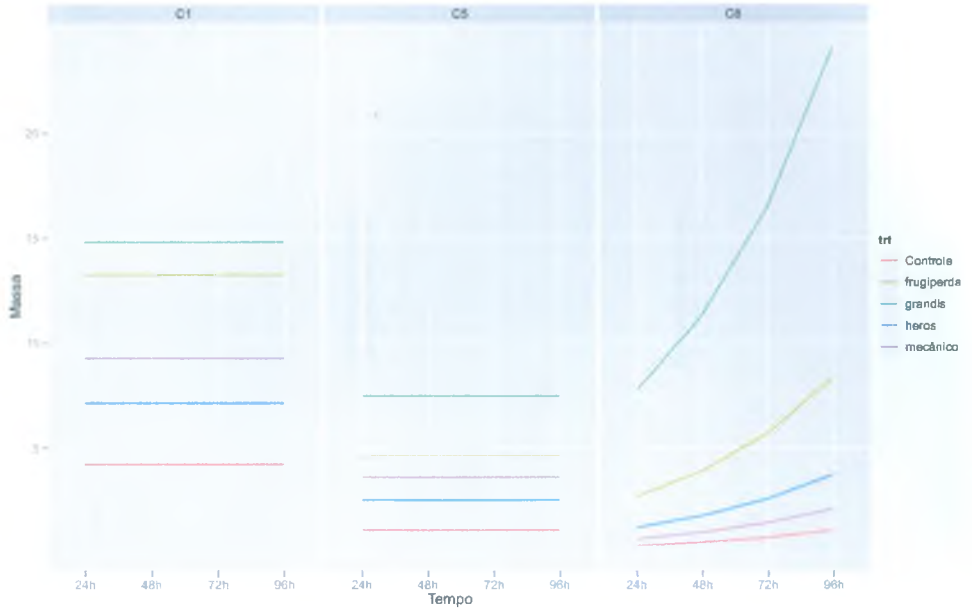

Figura 4.6: Valores preditos pelos MLGs univariados para as massas dos compostos C1, C5 e C8

Para acomodar o efeito referente à medição dos indivíduos em diferentes tempos, incorpora-se aos modelos univariados, um efeito aleatório para cada indivíduo (planta), isto é:

$$Y_{ijkt} \sim PC(\mu_{ijkt}, \phi_j, p_j) \tag{4.4}$$

$$\ln(\mu_{ijkt}) = \alpha_{jk} + \beta_j t + \gamma_{jk} t + u_{ji} \tag{4.5}$$

em que

- α_{jk}: Efeito do k-ésimo tratamento para o composto j;
- β_j: Coeficiente angular para a variável tempo para o composto j;
- γ_{kt}: Efeito de interação entre tratamento e tempo para o composto j;
- u_{ji}: Efeito aleatório no qual $u_{ji} \overset{iid}{\sim} N(0, \sigma_j^2)$, $j = 1, 2, 3$;

A função de verossimilhança marginal, para cada composto. do modelo descrito é dada por

$$l_j(\Theta_j|\mathbf{x}_j, \mathbf{y}_j) = \prod_{i=1}^{40} \int_{-\infty}^{+\infty} \prod_{t=1}^{4} f(y_{ijt}|\mathbf{x}_j, u_{ji}, \Theta_j) \frac{e^{-u_{ji}^2/2\sigma_j^2}}{\sqrt{2\pi\sigma_j^2}} du_{ji} \tag{4.6}$$

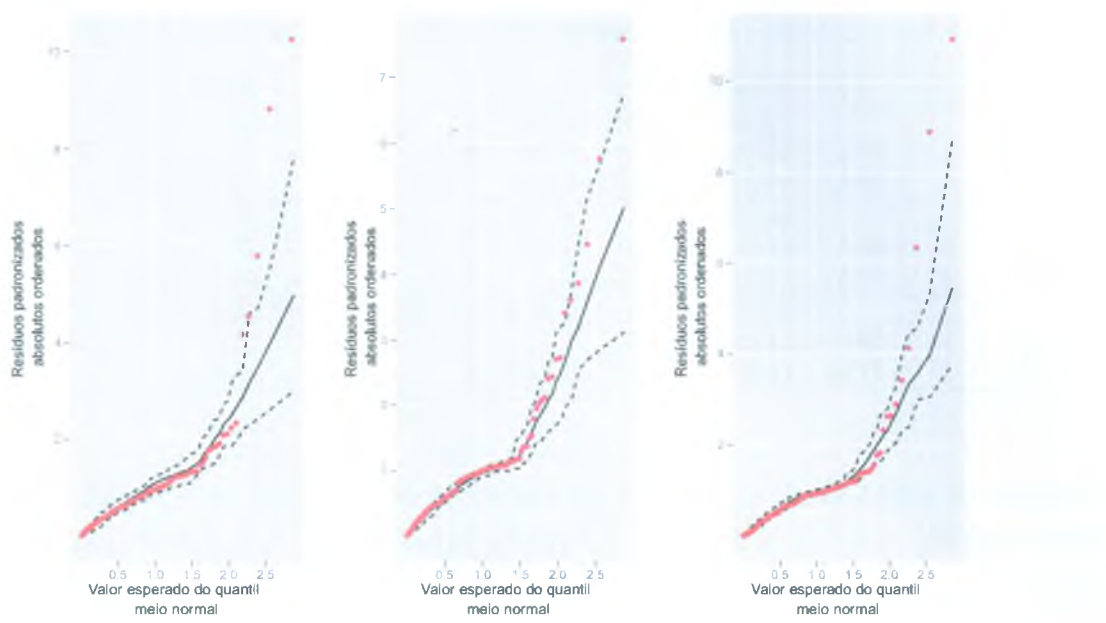

Figura 4.7: Gráficos meio-normais dos MLGs para C1 (esquerda), C5 (centro) e C8 (direita)

em que $f(y_{ijt}|\mathbf{x}_j, u_{ji}, \Theta_j)$ é a função de densidade da distribuição Poisson composta e $\Theta_j = (\alpha_{j1}, ..., \alpha_{j5}, \beta_j, \gamma_{j1}, ..., \gamma_{j5}, \sigma_j^2)$ é o vetor de parâmetros do modelo.

Para cada composto, foram ajustados também modelos aninhados sem interação e, posteriormente, sem efeito de tempo. A Tabela 4.4 apresenta os resultados dos testes de razão de verossimilhanças. Sob nível de significância de 5%, o modelos selecionados para os compostos C1 e C5 incluem nos efeitos fixos apenas o efeito de tratamento, indicando que para estes compostos não há variação significativa da massa medida ao longo do tempo, enquanto para C8 tem-se, além de efeito de tratamento, o coeficiente para tempo, mas sem interação, isto é, na escala do preditor linear, a evolução temporal da massa do composto não parece ser afetada pelo tratamento aplicado.

Para os modelos selecionados, as estimativas dos parâmetros são apresentadas na Tabela 4.5. Para o composto C8, no qual se tem efeito de tempo, verifica-se um coeficiente positivo, indicando que, para os tempos observados, há tendência crescente da massa do composto. Além disso, as estimativas pontuais indicam que, de maneira geral para os 3 compostos analisados, as plantas do grupo Controle são as que apresentam menor massa, enquanto as plantas submetidas ao tratamento *A. grandis* têm, em média, maior massa. A Figura 4.8 apresenta o gráfico com os valores preditos.

Tabela 4.4: Teste RV para os modelos concorrentes – Modelos univariados

Composto	Modelo	GL	AIC	BIC	log-Lik	X^2_{obs}	GL_{X^2}	P-valor
C1	$\alpha_{1k} + \beta_1 t + \gamma_{1k} t + u_{1i}$	12	1034.5	1071.4	-505.24			
	$\alpha_{1k} + \beta_1 t + u_{1i}$	8	1030.0	1054.6	-507.00	3.52	4	0.474
	$\alpha_{1k} + u_{1i}$	7	1030.1	1051.6	-508.04	2.07	1	0.149
C5	$\alpha_{2k} + \beta_2 t + \gamma_{2k} t + u_{2i}$	12	706.91	743.81	-341.45			
	$\alpha_{2k} + \beta_2 t + u_{2i}$	8	705.50	730.10	-344.75	6.58	4	0.159
	$\alpha_{2k} + u_{2i}$	7	704.34	725.87	-345.17	0.84	1	0.358
C8	$\alpha_{3k} + \beta_3 t + \gamma_{3k} t + u_{3i}$	12	674.83	711.73	-325.41			
	$\alpha_{3k} + \beta_3 t + u_{3i}$	8	667.22	691.82	-325.61	0.38	4	0.983
	$\alpha_{3k} + u_{3i}$	7	694.93	716.46	-340.47	29.71	1	5×10^{-8}

A avaliação da qualidade do ajuste dos modelos selecionados é feita por meio de gráficos meio-normais com envelopes simulados (Figura 4.9) com $n = 100$ simulações e limites de 98% para os envelopes. Tem-se um ajuste satisfatório dos dados, ainda que alguns pontos dos compostos C1 e C8 tenham ficado fora das regiões simuladas.

Tabela 4.5: Estimativa e erro padrão dos parâmetros dos MLG mistos univariados

Parâmetro	Composto		
	C1	C5	C8
α_{j1} (Controle)	1.314 (0.242)	-0.10 (0.334)	-1.736 (0.387)
α_{j2} (Frugiperda)	2.461 (0.225)	1.246 (0.322)	0.283 (0.367)
α_{j3} (Grandis)	2.600 (0.223)	1.634 (0.319)	1.252 (0.360)
α_{j4} (Heros)	1.728 (0.235)	0.542 (0.328)	-0.744 (0.376)
α_{j5} (Mecânico)	1.963 (0.231)	0.834 (0.326)	-1.044 (0.379)
β_j	–	–	0.017 (0.003)
ϕ_j	1.207 (0.097)	0.950 (0.064)	1.064 (0.071)
p_j	1.638 (0.037)	1.775 (0.031)	1.755 (0.023)
σ^2_j	0.281	0.652	0.645

Abordagem 3: MLGM multivariado

Para o modelo multivariado, segue-se a especificação de Fieuws *et al.* (2006), no qual além de acomodar o efeito relativo a repetidas medições de cada composto em cada planta, mas também a correlação entre os compostos, o que é feito por meio da especificação de um vetor de efeitos aleatórios para cada indivíduo. Dessa forma, tem-se:

$$Y_{ijkt} \sim PC(\mu_{ijkt}, \phi, p) \tag{4.7}$$

em que Y_{ijkt} é o valor da medida de massa para o i-ésimo indivíduo, j-ésimo composto, k-ésimo tratamento, no tempo t. O modelo para a média é ajustado pelo

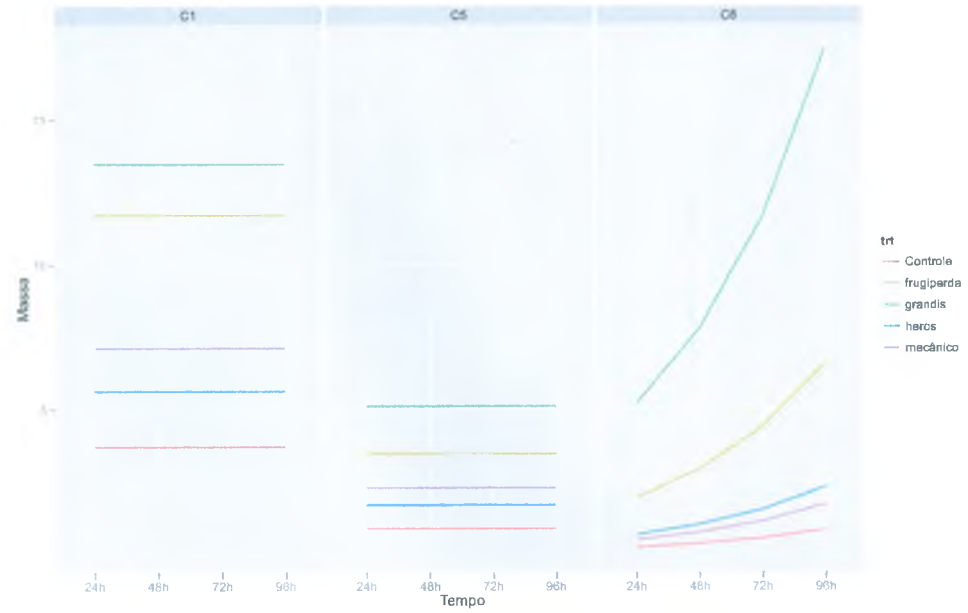

Figura 4.8: Valores preditos pelos modelos mistos univariados para as massas dos compostos C1, C5 e C8

modelo saturado, isto é, para cada composto tem-se os efeitos de tratamento e tempo, além da interação:

$$\ln(\mu_{ijkt}) = \alpha_{jk} + \beta_j t + \gamma_{jk} t + u_{ij} \tag{4.8}$$

em que

- α_{jk}: Efeito do k-ésimo tratamento para o composto j;
- β_j: Coeficiente angular para a variável tempo para o composto j;
- γ_{jk}: Efeito de interação entre tratamento e tempo para o composto j;
- u_{ij}: Efeito aleatório para o indivíduo i, composto j. O vetor de efeitos aleatórios do indivíduo i tem distribuição normal trivariada, com vetor de médias $(0, 0, 0)^\top$ e matriz de covariâncias não estruturada, como aquela expressa em (3.2), de dimensão 3×3;

Para o modelo multivariado, a função de verossimilhança do modelo é dada por

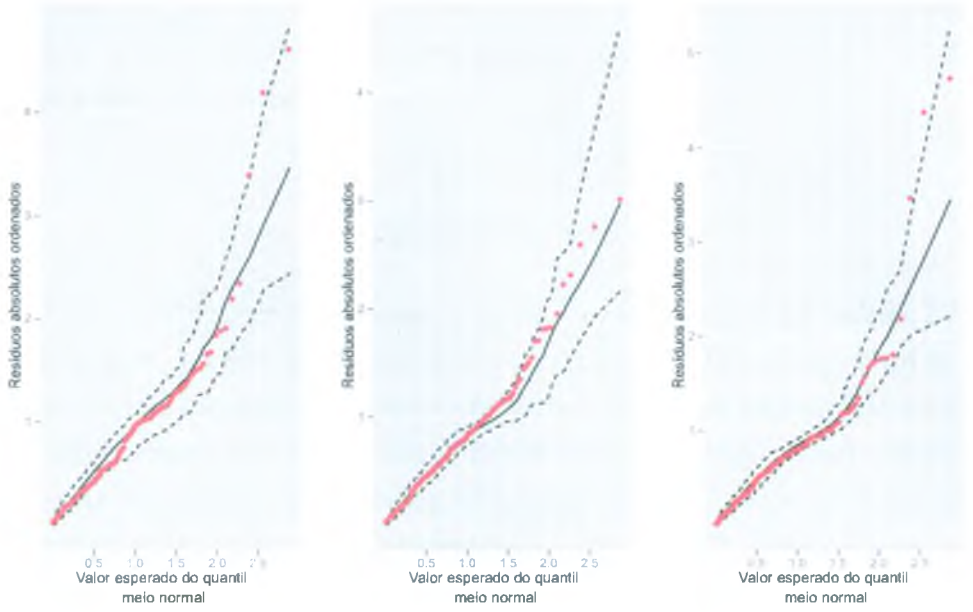

Figura 4.9: Gráficos meio-normais dos modelos mistos univariados para C1 (esquerda), C5 (centro) e C8 (direita)

$$l(\Theta|\mathbf{x},\mathbf{y}) = \prod_{i=1}^{40} \int_{\mathbb{R}^3} \left[\prod_{j=1}^{3} \prod_{t=1}^{4} f(y_{ijt}|\mathbf{x},\mathbf{u}_i,\Theta) \right] \frac{|\Sigma|^{-1/2}}{(2\pi)^{3/2}} e^{-\frac{1}{2}\mathbf{u}_i^T \Sigma^{-1} \mathbf{u}_i} d\mathbf{u}_i \tag{4.9}$$

em que, novamente, $f(y_{ijt}|\mathbf{x},\mathbf{u}_i,\Theta)$ é a função de densidade da distribuição Poisson composta, Θ é o vetor de parâmetros e Σ é a matriz de covariâncias dos efeitos aleatórios.

Ajustam-se, ainda, os modelos sem a interação γ_{jk} e sem efeito de tempo. Posteriormente, realiza-se o teste da razão de verossimilhança para verificar se estes termos devem ser mantidos no modelo. O resultado obtido é apresentado na Tabela 4.6.

Tabela 4.6: Teste RV para os modelos concorrentes – Modelo multivariado

Modelo	GL	AIC	BIC	log-Lik	X^2_{obs}	GL_{X^2}	P-valor
$\alpha_{jk} + \beta_j t + \gamma_{jk} t + u_{ij}$	37	2387.6	2542.0	-1156.8			
$\alpha_{jk} + \beta_j t + u_{ij}$	25	2373.3	2477.6	-1161.6	9.71	12	0.640
$\alpha_{jk} + u_{ij}$	22	2406.9	2498.7	-1181.4	39.56	3	1.31×10^{-8}

Sob nível de significância de 1%, não há evidências para rejeitar a hipótese de

nulidade de γ_{jk}. Entretanto, rejeita-se a nulidade do termo β_j. Dessa forma, o modelo a ser utilizado como base para as análises seguintes é o modelo com os termos de efeitos principais α_{jk} e β_j:

$$\ln(\mu_{ijkt}) = \alpha_{jk} + \beta_j t + u_{ij} \tag{4.10}$$

Uma vez selecionado o modelo descrito em 4.10, tem-se na Tabela 4.7 as estimativas e, entre parênteses, os respectivos erros-padrão do modelo. Para os compostos C1 e C5, nota-se pela magnitude dos erros-padrão das estimativas de β_j, $j = 1, 2$, que este termo não é significativo. Entretanto, a inspeção visual dos gráficos meio-normais com envelopes simulados, utilizados para verificação da qualidade do ajuste do modelo multivariado e apresentados na Figura 4.11, indica ajuste aceitável dos dados. A Figura 4.10 apresenta valores preditos pelo modelo misto multivariado.

Tabela 4.7: Estimativa e erro padrão dos parâmetros do modelo 4.10

Parâmetro	Composto		
	C1	C5	C8
α_{j1} (Controle)	1.127 (0.276)	-0.154 (0.372)	-1.714 (0.389)
α_{j2} (Frugiperda)	2.249 (0.265)	1.122 (0.359)	0.206 (0.369)
α_{j3} (Grandis)	2.386 (0.264)	1.604 (0.355)	1.252 (0.361)
α_{j4} (Heros)	1.499 (0.272)	0.440 (0.365)	-0.775 (0.378)
α_{j5} (Mecânico)	1.762 (0.269)	0.810 (0.362)	-1.055 (0.381)
β_j	0.003 (0.002)	0.002 (0.003)	0.017 (0.003)
ϕ	1.028 (0.040)		
p	1.735 (0.016)		
σ_j^2	0.278	0.648	0.679

Na Figura 4.12, tem-se o gráfico de matriz de dispersão dos efeitos aleatórios estimados no modelo. Nota-se que há uma relação linear quase perfeita entre os efeitos estimados para os compostos 5 e 8. A matriz de correlações estimada mostra forte correlação entre estes compostos e, ainda, correlação moderada entre o composto 1 e os demais.

$$\hat{\rho}_{u_{Composto}} = \begin{pmatrix} 1.000 & 0.457 & 0.492 \\ 0.457 & 1.000 & 0.999 \\ 0.492 & 0.999 & 1.000 \end{pmatrix}$$

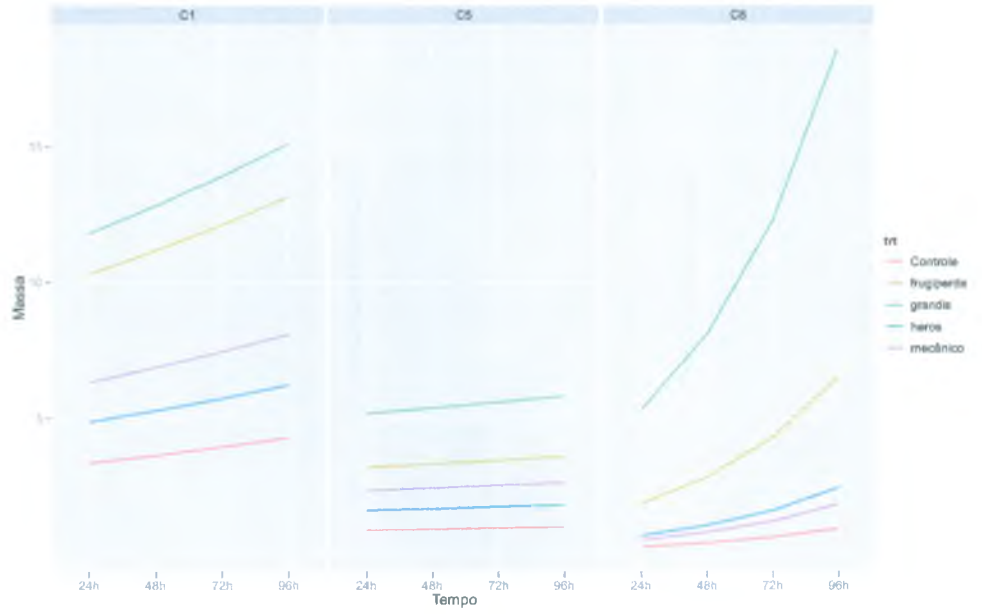

Figura 4.10: Valores preditos pelo modelo misto multivariado para as massas dos compostos C1, C5 e C8

4.5 Comparações entre tratamentos

Uma vez estimados os modelos para os compostos seguindo as três abordagens, prossegue-se com as comparações múltiplas dos tratamentos. Tendo-se que em nenhum dos casos há interação entre tratamento e tempo, e ainda, que quando há efeito de tempo, este é positivo – havendo, portanto, indícios de que para o período de observação do experimento o pico da massa média ocorre após 96 horas da aplicação dos tratamentos – faz-se apenas a comparação entre os tratamentos aplicados. Opta-se por realizar as comparações com aquele tratamento com estimativa pontual mais elevada nos modelos estimados – *A. grandis*, nas três abordagens utilizadas, para os três compostos. A Tabela 4.8 apresenta os contrastes estimados e, entre parênteses, os p-valores corrigidos pelo método de Holm-Bonferroni obtidos para os modelos MLG, MLG misto univariado e MLG misto multivariado, respectivamente.

Para o composto 1, α-Pinene, nota-se que, sob nível de significância de 5% há concordância entre os três modelos – tratamento Controle e *E. heros* são significativamente diferentes do tratamento *A. grandis*, mas o mesmo não pode ser afirmado para *S. frugiperda* e dano mecânico.

Já para o composto 5, β-Myrcene, os modelos que incorporam efeitos aleatórios tendem a ser mais conservadores, gerando resultados conflitantes nas comparações dos tratamentos *E. heros* e dano mecânico com o tratamento *A. grandis*. De forma

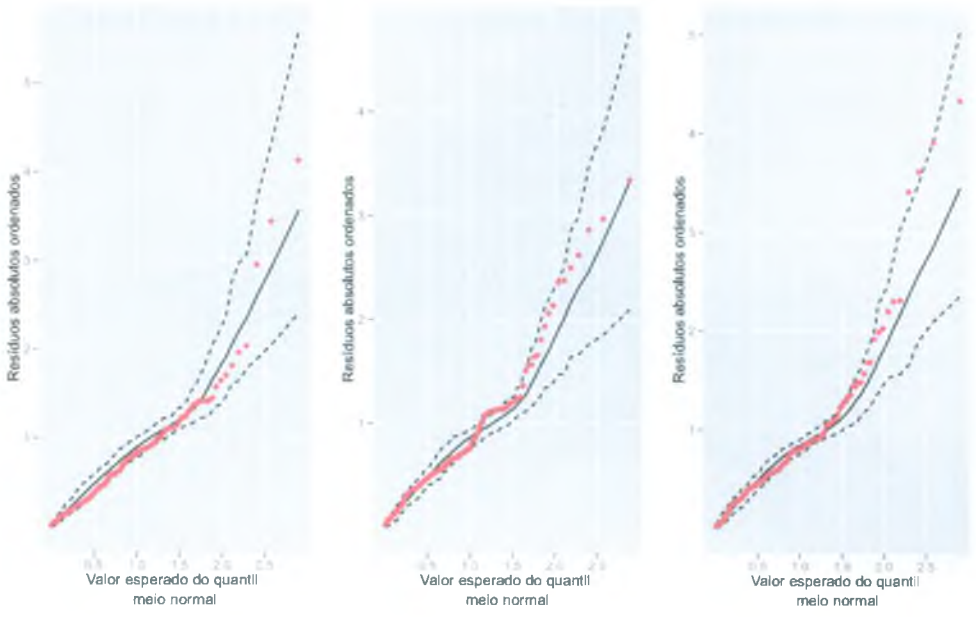

Figura 4.11: Gráficos meio-normais do modelo multivariado para C1 (esquerda), C5 (centro) e C8 (direita)

Tabela 4.8: Estimativas e p-valores corrigidos dos contrastes para C1, C5 e C8

Abordagem	Composto	trtG - trtC	trtG - trtF	trtG - trtH	trtG - trtM
MLG	C1	1.254 (<1e-4)	0.112 (0.6725)	0.731 (0.025)	0.47 (0.1664)
	C5	1.911 (<1e-4)	0.482 (0.1008)	1.086 (9e-04)	0.723 (0.0295)
	C8	3.078 (<1e-4)	1.057 (0.001)	1.851 (<1e-4)	2.428 (<1e-4)
MLGM (uni)	C1	1.287 (4e-04)	0.139 (0.6642)	0.873 (0.0235)	0.637 (0.1008)
	C5	1.734 (9e-04)	0.388 (0.3998)	1.091 (0.0571)	0.8 (0.1688)
	C8	2.988 (<1e-4)	0.969 (0.0347)	1.996 (<1e-4)	2.295 (<1e-4)
MLGM (multi)	C1	1.259 (5e-04)	0.137 (0.6681)	0.887 (0.0196)	0.624 (0.1081)
	C5	1.758 (6e-04)	0.482 (0.2884)	1.164 (0.0337)	0.794 (0.164)
	C8	2.966 (<1e-4)	1.046 (0.0228)	2.027 (<1e-4)	2.307 (<1e-4)

geral. porém, parece haver diferença significativa entre os tratamentos controle e *A. grandis*. o que não ocorre na comparação entre *S. frugiperda* e *A. grandis*.

Para o composto 8, β-Ocimene, os três modelos apontam significativa diferença entre *A. grandis* e os tratamentos controle, *E. heros* e dano mecânico. Com relação ao contraste entre *A. grandis* e *S. frugiperda*. o modelo sem efeito aleatório indica diferença significativa, enquanto os outros modelos são, novamente, mais conservadores e indicam maior incerteza, embora ainda apresentem p-valores abaixo do limiar de 5%.

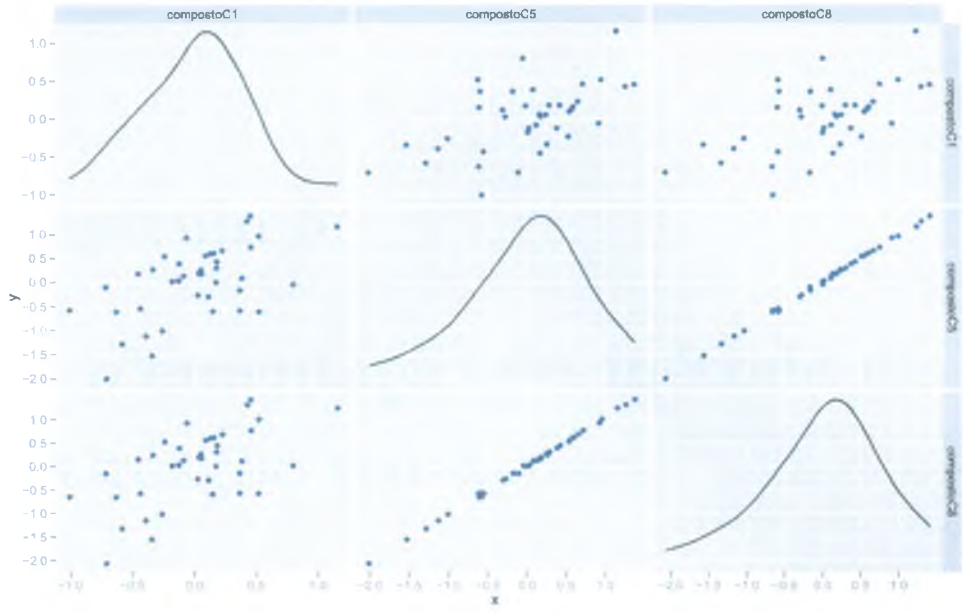

Figura 4.12: Matriz de dispersão dos efeitos aleatórios preditos em 4.10

4.6 Considerações

No ajuste dos modelos para os compostos C1, C5 e C8, verifica-se que as três aborgadens utilizadas apresentam ajuste aceitável, mas é possível fazer uma distinção qualitativa entre os modelos que incluem efeitos aleatórios com relação aos modelos sem efeitos aleatórios e concluir que os modelos de efeitos mistos apresentam ligeira melhora no ajuste dos dados.

Ainda assim, as três abordagens apresentam indícios de que os compostos C1 e C5, não apresentam variação significativa com relação ao tempo, enquanto o composto C8 tem, em média, um comportamento crescente. O modelo multivariado permite ainda constatar que há forte correlação entre os compostos C5 e C8. β-Myrcene e β-Ocimene, uma vez isolados os efeitos de tratamento e de tempo.

Na comparação entre os tratamentos, avaliando-se os resultados das abordagens de efeitos mistos uni e multivarados, o tratamento *A. grandis*, utilizado como tratamento de referência, se mostra superior ao tratamento controle para os três compostos. Comparado aos outros tratamentos, para o composto C8 é significativamente distinto de *S. frugiperda*, *E. heros* e dano mecânico, sob nível de significância de 5%.

Capítulo 5

Implementação do método de estimação par-a-par

5.1 Introdução

Como parte do trabalho computacional, fez-se necessária a implementação da abordagem par-a-par proposta para modelos multivariados. Tendo em vista que o pacote *cplm* (Zhang, 2012a) do software *R* (R Development Core Team 2012) permite o ajuste de modelos mistos para a distribuição Poisson composta, com especificação flexível dos efeitos fixos e aleatórios, por meio da função *cpglmm*, este pacote foi tomado como base para a implementação do ajuste par-a-par dos modelos multivariados[1]. Para o cálculo dos erros padrões das estimativas obtidas, adota-se o método de Bates (2010) expresso em (2.39), uma vez que seu cálculo não depende de derivações numéricas de primeira e segunda ordem da verossimilhança marginal, reduzindo o esforço computacional.

5.2 Comparação da performance entre ajuste multivariado e par-a-par

Conforme discutido anteriormente, a abordagem de ajuste par-a-par de modelos multivariados proposta por Fieuws e Verbeke (2006) tem como potencial atrativo a redução do custo computacional. Entretanto, é preciso avaliar de forma mais sistemática se efetivamente há redução do tempo de computação ao estimar

[1] De fato, até o momento da elaboração deste trabalho, apenas o pacote *cplm* permite o ajuste de modelos mistos para a distribuição Poisson composta, incluindo modelos com efeitos aleatórios aninhados e cruzados.

$m(m-1)/2$ modelos bivariados em vez de estimar apenas um modelo que depende do cálculo de integrais com três ou mais dimensões.

A fim de comparar a performance das duas abordagens de estimação para modelos multivariados, são simulados dados com 30 indivíduos (efeitos aleatórios), um fator com 4 níveis (efeitos fixos) e número de dimensões m (variáveis que compõem a resposta multivariada) variando entre 5 e 20. A variável resposta condicionada ao vetor de efeitos aleatórios tem distribuição Poisson composta com parâmetros $\mu_{ijk} = \exp(\alpha_k + \beta_{jk} + u_{ik})$ $(i = 1, ..., 30;\ j = 1, .., 4,\ k = 1, ..., m)$, $\phi = 1$ e $p = 1.6$.

Os resultados obtidos[2] são apresentados na Tabela 5.1 e Figura 5.1 abaixo. Como pode ser percebido, há uma redução substancial no tempo de processamento conforme o número de dimensões aumenta – para o caso $m = 20$, por exemplo, foram necessários aproximadamente 10 minutos (600.26 segundos) para finalizar o ajuste par-a-par do modelo, ao passo que o ajuste multivariado foi interrompido após mais de 4.5 horas sem chegar ao fim do processo de estimação. Levando-se em consideração que o ajuste dos pares de modelos é feito de forma paralelizada, o tempo computacional poderia ser reduzido ainda mais com um número maior de núcleos de processamento.

Tabela 5.1: Comparação dos tempos (em segundos) para ajuste multivariado e par-a-par

m	Multivariado	Par-a-par
3	18.13	18.66
5	104.55	43.08
7	294.98	79.15
10	1740.00	150.72
13	4684.29	260.01
15	8509.26	331.05
20	> 16800	600.26

Analisando-se os tempos obtidos para o ajuste par-a-par, verifica-se ainda que o tempo de estimação é aproximadamente proporcional ao número de pares ajustados, o que pode ser observado na Figura 5.2.

[2]Para este estudo, foi utilizado um computador com CPU Intel Core^{TM}i5 750, com 4 núcleos de processamento.

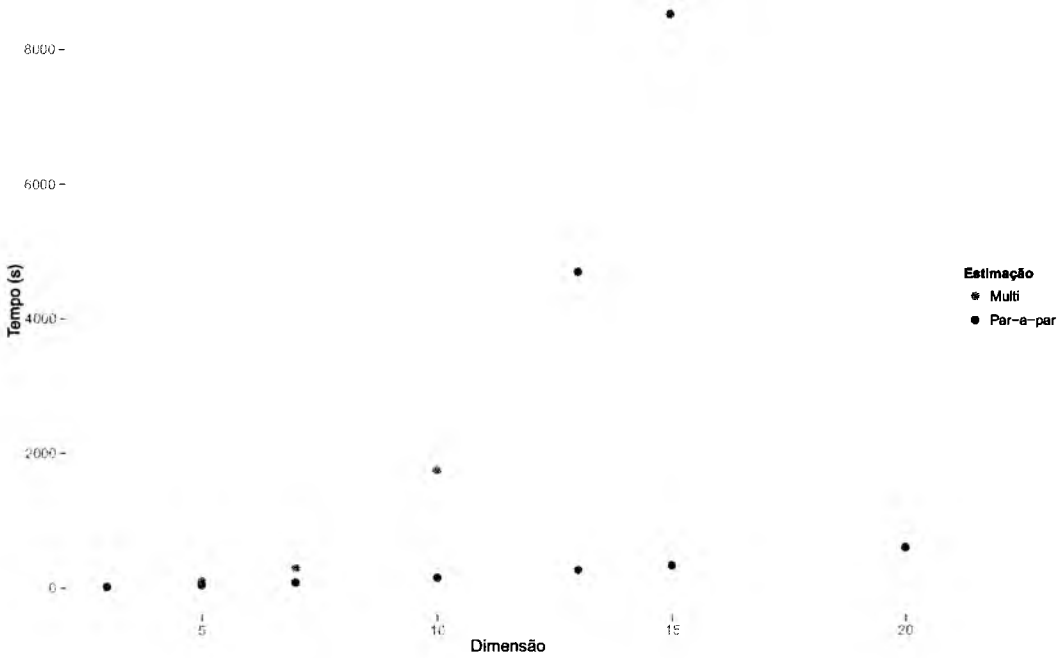

Figura 5.1: Tempos para os ajustes multivariado e par-a-par

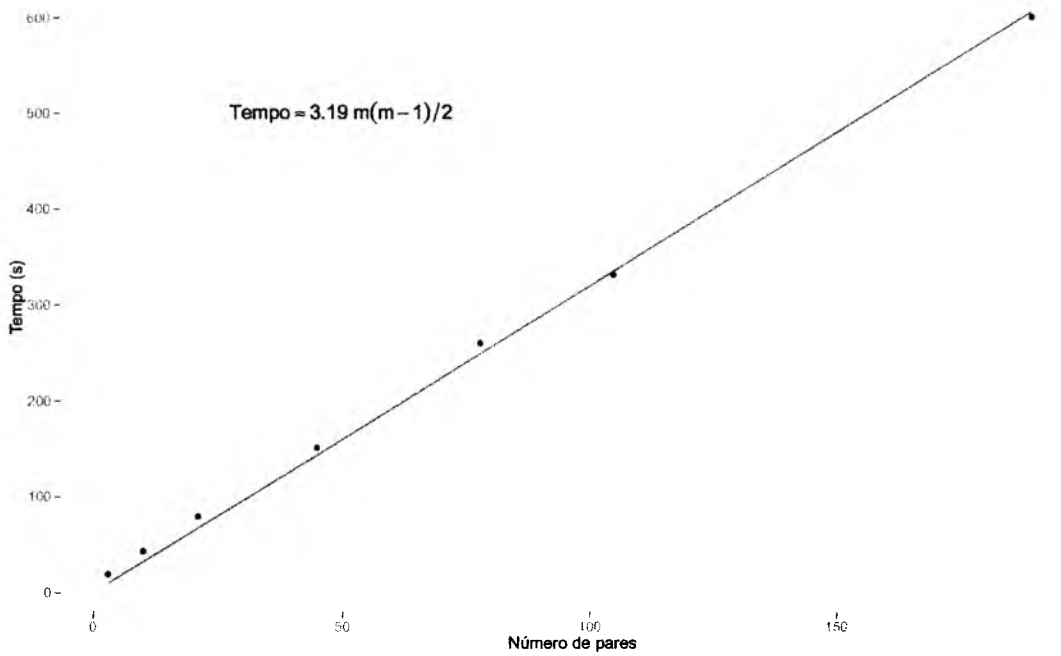

Figura 5.2: Tempos do ajuste par-a-par com relação ao número de pares

5.3 Modelo multivariado para dados de algodão – compostos 1, 5 e 8

Para verificar as diferenças produzidas entre o ajuste de modelos estimados de forma multivariada com relação à implementação da estimação par-a-par, os dados de algodão para os compostos 1, 5 e 8 são novamente analisados, desta vez utilizando-se a função implementada. A Figura 5.3 mostra que há pouca diferença nas estimativas produzidas, indicando a validade do método e que não há problemas na implementação.

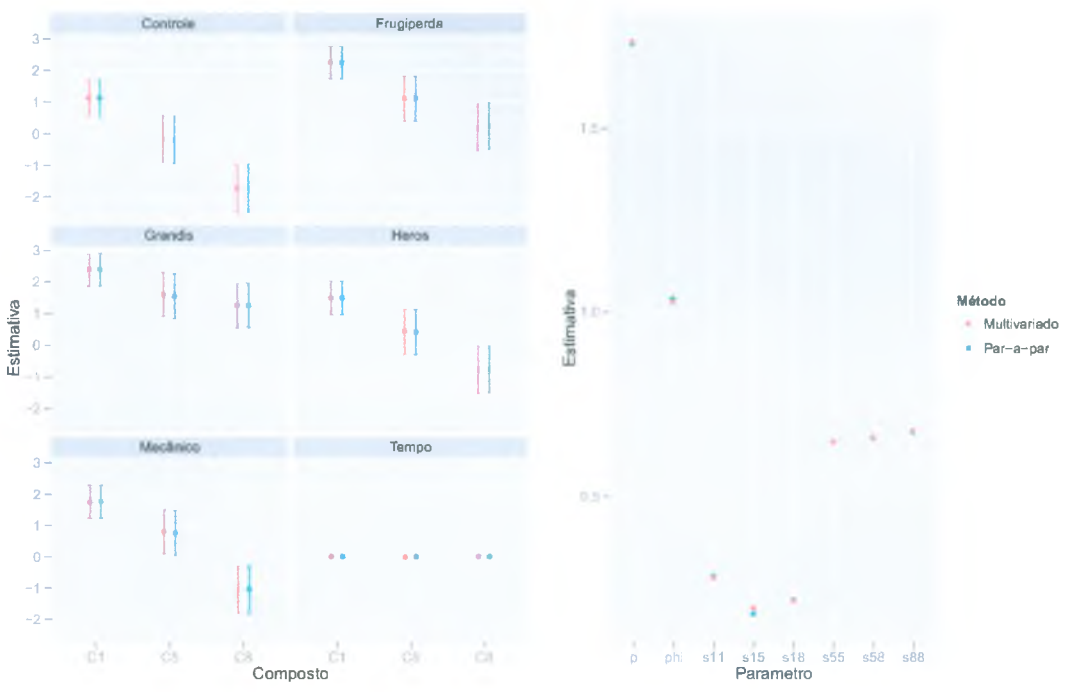

Figura 5.3: Comparação das estimativas do ajuste multivariado e par-a-par

5.4 Considerações

Conforme visto, tem-se que o ajuste par-a-par dos modelos multivariados não somente produzem resultados semelhantes, mas apresentam performance melhor em termos computacionais, uma vez que o tempo necessário para estimação de modelos multivariados pela resolução de integrais multidimensionais explode rapidamente conforme o número de variáveis aumenta, enquanto para a abordagem par-a-par, o tempo varia de forma aproximadamente linear com relação ao número de pares a

serem ajustados. Uma outra vantagem da abordagem proposta, é a possibilidade de que os $m(m-1)/2$ pares de modelos sejam ajustados em paralelo, já que não há qualquer tipo de dependência na estimação de dois pares distintos de modelos. De fato, esta potencialidade foi explorada na implementação por meio da utilização do pacote *doParallel* (Revolution Analytics, 2012), possibilitando uma redução substancial no tempo de computação.

As funções para simulação de dados multivariados, bem como para estimação de modelos via abordagem de pseudo-verossimilhança (estimação par-a-par) foram reunidas em um pacote para a linguagem *R*. Os *links* para obtenção do pacote e descrição de utilização do mesmo podem ser encontrados no Apêndice B.

Capítulo 6

Análise dos dados de experimento de Algodão

6.1 Introdução

No Capítulo 4, introduziu-se o experimento em que plantas de algodão em estado reprodutivo são submetidas a um de cinco tratamentos (Controle, *A. grandis*, *E. heros*, *S. frugiperda* e dano mecânico) e para as quais são medidas as massas (em μ g) de 25 compostos químicos, em 4 momentos (24, 48, 72 e 96 horas após aplicação do tratamento).

Anteriormente, um subconjunto de três compostos foi selecionado a fim de comparar diferentes abordagens na modelagem dos dados de interesse. Aqui, porém, será dado enfoque apenas à modelagem multivariada dos 25 compostos observados, como aplicação da metodologia par-a-par de estimação dos parâmetros do modelo linear generalizado misto multivariado.

Dessa forma, tem-se que a modelagem dos dados de algodão é feita por meio de um modelo misto multivariado com distribuição Poisson composta para a variável resposta, análogo àquele especificado na *abordagem 3* do capítulo 4, equações (4.7)–(4.9). Em virtude da dimensão do modelo, a abordagem par-a-par, apresentada no Capítulo 3 e implementada no Capítulo 5, é adotada.

6.2 Modelagem multivariada

O modelo para o processo gerador dos dados pode ser especificado supondo-se distribuição Poisson composta para a variável resposta, tal que:

$$Y_{ijkt} \sim PC(\mu_{ijkt}, \phi, p) \tag{6.1}$$

em que Y_{ijkt} é o valor da medida de massa (em $10^3 \mu g$) para o i-ésimo indivíduo, j-ésimo composto, k-ésimo tratamento, no tempo t. O modelo saturado para a média de cada composto, com os efeitos principais de tratamento e tempo, além do termo de interação entre tratamento e tempo é escrito na forma

$$\ln(\mu_{ijkt}) = \alpha_{jk} + \beta_j t + \gamma_{jk} t + u_{ij} \tag{6.2}$$

em que

- α_{jk}: Efeito do k-ésimo tratamento para o composto j;
- β_j: Coeficiente angular para a variável tempo para o composto j;
- γ_{jk}: Efeito de interação entre tratamento e tempo para o composto j;
- u_{ij}: Efeito aleatório para o indivíduo i, composto j. O vetor de efeitos aleatórios do indivíduo i tem distribuição normal multivariada, tal que

$$\begin{pmatrix} u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,25} \end{pmatrix} \sim N_{25} \left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \delta_1^2 & \delta_{1,2} & \cdots & \delta_{1,25} \\ \delta_{2,1} & \delta_2^2 & \cdots & \delta_{2,25} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{25,1} & \delta_{25,2} & \cdots & \delta_{25}^2 \end{pmatrix} \right) \tag{6.3}$$

Como modelos concorrentes para a média, são ajustados, ainda, um modelo somente com os efeitos principais (sem o termo de interação) e um modelo apenas com efeito de tratamento (sem efeito de tempo). No ajuste par-a-par dos 25 compostos, verificaram-se problemas de convergência para os compostos Benzothiazol e 17 β-cariofileno, o que levou à exclusão destes dois compostos do ajuste dos modelos.

Os testes de razão de verossimilhanças para seleção do modelo são apresentados na Tabela 6.1. Tomando-se um nível de significância de 5%, decide-se pelo modelo com interação e efeitos principais de tratamento e tempo, cujas estimativas são apresentadas na Tabela A.1.

Os gráficos das Figuras 6.1 e 6.2, nos quais têm-se, respectivamente, as médias e valores ajustados pelo modelo estimado, permitem uma comparação visual dos dados observados e os efeitos capturados pelo modelo multivariado. É possível observar,

Tabela 6.1: Teste RV para os modelos concorrentes

Modelo	GL	logLik	X^2_{obs}	GL_{X^2}	P-valor
$\alpha_{jk} + \beta_j t + \gamma_{jk} t + u_{ij}$	507	-4524.22			
$\alpha_{jk} + \beta_j t + u_{ij}$	415	-4585.65	122.85	92	0.017
$\alpha_{jk} + u_{ij}$	392	-4825.79	480.28	23	0.000

ainda, em ambos os gráficos, que alguns compostos são liberados de forma mais intensa que outros.

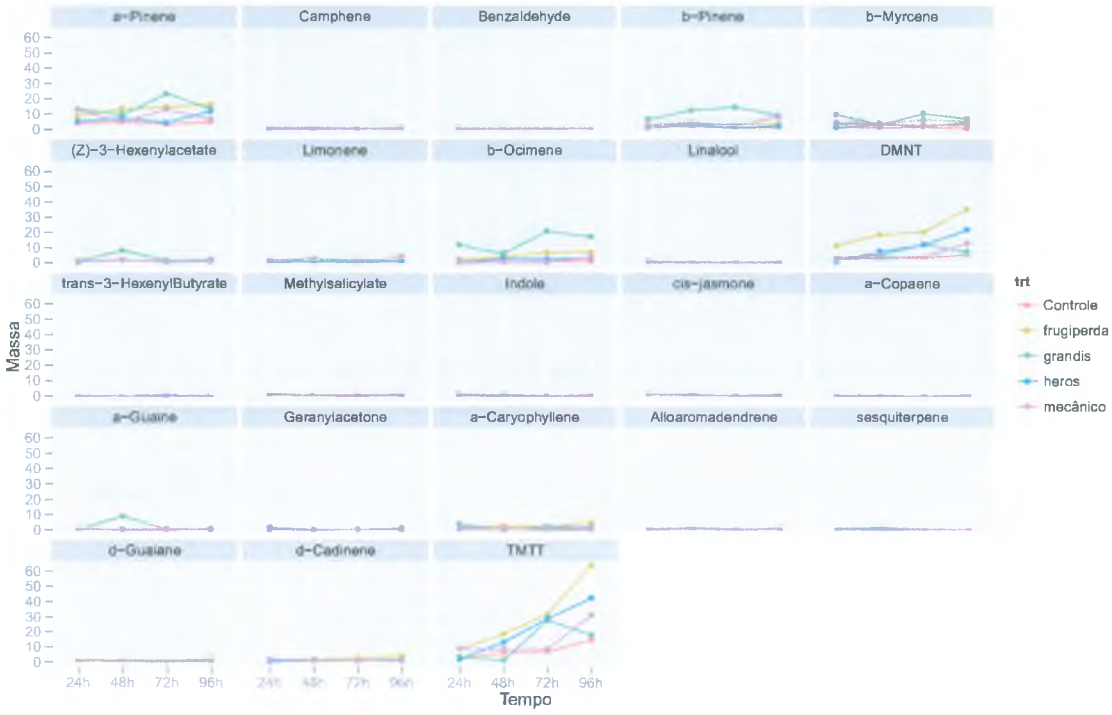

Figura 6.1: Gráfico das médias de cada tratamento ao longo do tempo para cada composto

A partir da matriz de componentes de variância estimada, a matriz de correlação é obtida e pode ser visualizada na Figura 6.3. Observa-se um predomínio de relações positivas entre os compostos, o que poderia ser esperado, uma vez que certos conjuntos de compostos compartilham rotas metabólicas.

6.3 Comparação dos tratamentos

Uma vez estimados os parâmetros do modelo, faz-se a comparação entre os tratamentos, utilizando-se o método de Holm-Bonferroni para correção dos p-valores

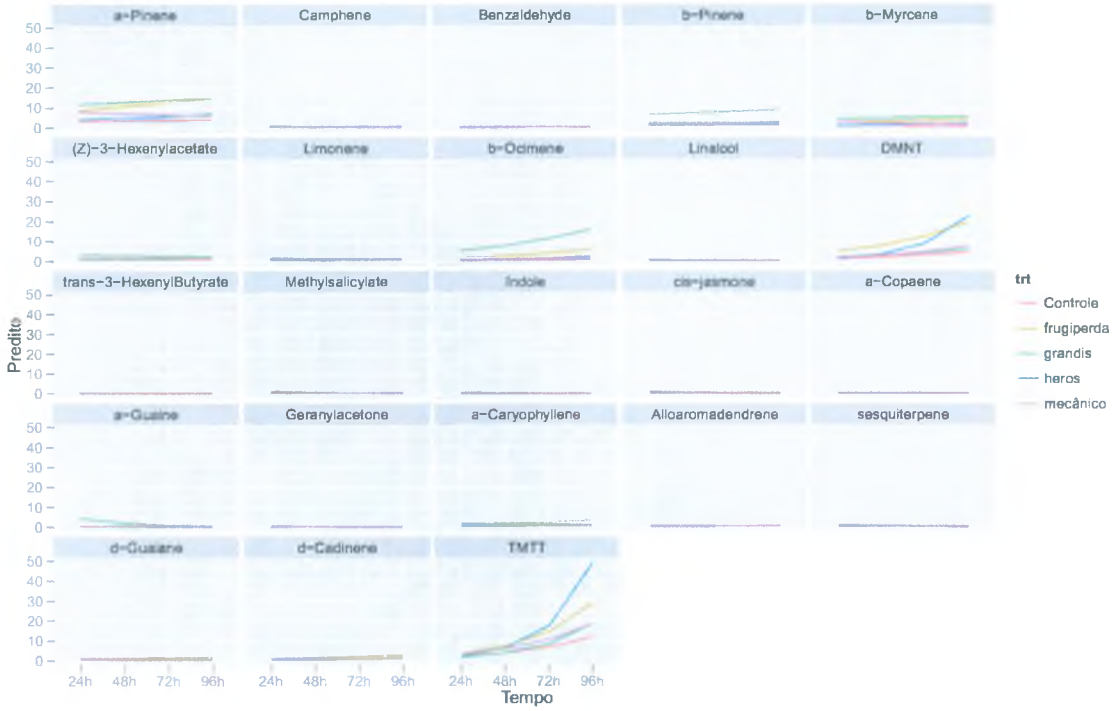

Figura 6.2: Gráfico de valores preditos pelo modelo com efeitos de tratamento, tempo e interação para cada composto

Tabela 6.2: Compostos em que *A. grandis* difere do tratamento controle

Composto	24h	48h	72h	96h
α-Pinene		✓	✓	✓
β-Pinene	✓	✓	✓	✓
β-Myrcene		✓	✓	✓
(Z)-3-Hexenylacetate		✓		
β-Ocimene	✓	✓	✓	✓
α-Guaine	✓	✓		
α-Caryophillene		✓	✓	✓
δ-Guaiene		✓	✓	✓

das comparações múltiplas. As comparações são feitas por linhas e colunas. isto é, fixando-se um determinado tratamento aplicado e comparando-se os tempos, e, posteriormente, fixando-se um tempo e comparando-se os tratamentos aplicados. As estimativas pontuais por tratamento, tempo e composto são apresentados na Tabela A.2, acompanhadas de letras para codificar os diferentes tratamentos – letras iguais indicam contrastes não significativos a 5%, caso contrário, são atribuídas letras diferentes.

Os resultados (Tabela A.2) podem ser resumidos conforme as Tabelas 6.2, 6.3 e 6.4. A Tabela 6.2 apresenta os compostos e tempos para os quais o tratamento *A. grandis* e Controle diferem significativamente. Analogamente, na Tabela 6.3 tem-se os casos em que há diferenças significativas entre *S. frugiperda* e Controle. Na Tabela

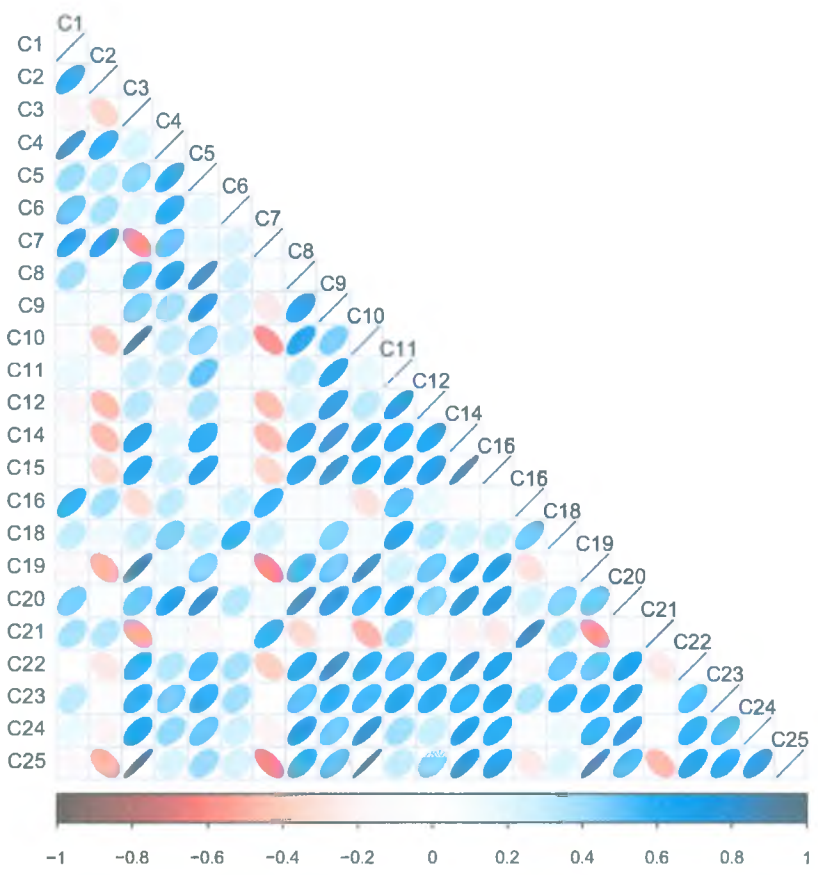

Figura 6.3: Gráfico de visualização das correlações entre compostos a partir da matriz de componentes de variância estimada

Tabela 6.3: Compostos em que *S. frugiperda* difere do tratamento controle

Composto	24h	48h	72h	96h
α-Pinene			✓	
β-Ocimene	✓	✓	✓	✓

6.4 tem-se os compostos e tempos para os quais o tratamento *A. grandis* difere de outros tratamentos, com indicação abreviada dos tratamentos dos quais difere.

Verifica-se que tratamento *A. grandis* difere do tratamento controle nos compostos α-Pinene, β-Pinene, β-Myrcene, (Z)-3-Hexenylacetate, β-Ocimene, α-Caryophillene e δ-Guaiene. com estimativas pontuais superiores. Diferenças significativas ocorrem também na comparação do tratamento *A. grandis* com *E. heros* nos compostos β-Pinene, β-Ocimene e α-Guaine, e com o tratamento dano mecânico para β-Ocimene e α-Guaine.

Para *A. grandis*, *E. heros*, *S. frugiperda* e dano mecânico, há um aumento significativo nas massas dos compostos β-Ocimene, DMNT e TMTT ao longo do tempo

Tabela 6.4: Compostos em que *A. grandis* difere de outro tratamentos

Composto	24h	48h	72h	96h
β-Pinene	Con, Her	Con, Her	Con, Her	Con, Her
β-Ocimene	Con, Her, Mec	Con, Her, Mec	Con, Her, Mec	Con, Her, Mec
α-Guaine	Con, Her, Mec, Fru	Con, Her, Mec, Fru	Mec	

– 24, 48 ,72 e 96 horas após aplicação do tratamento. Ainda, para *A. grandis* há uma redução significativa ao longo tempo na produção do composto α-Guaine.

6.4 Considerações

Como aplicação do modelo multivariado com estimação par-a-par de parâmetros, segundo a metodologia de pseudo-verossimilhança de Fieuws *et al.* (2006), objetivou-se trabalhar com os dados de 25 compostos de plantas de algodão. Entretanto, em certos pares de modelos envolvendo os compostos Benzothiazol e β-cariofileno (separadamente), diagnosticaram-se problemas de convergência e estes compostos foram, então, retirados da análise.

Utilizando-se 3 núcleos de processamento, o modelo com termo de interação ente tratamento e tempo foi estimado em 4855.66 segundos ($\sim$ 1 hora 21 minutos), o modelo apenas com os efeitos principais foi estimado em 1554.31 segundos ($\sim$ 26 minutos) e o modelo somente com efeito de tratamento, em 1215.37 segundos ($\sim$ 20 minutos). Considerando-se a dimensionalidade dos modelos – consequência da presença de 23 compostos por planta de algodão – a abordagem par-a-par tem performance notável. Levando-se em consideração, ainda, os resultados obtidos no Capítulo 5, em que foram comparados os tempos das abordagens multivariada e par-a-par, é razoável admitir que o ajuste dos modelos candidatos pela abordagem multivariada demandasse um tempo de computação extremamente longo.

A partir do conjunto de dados com os 23 compostos restantes, os modelos multivariados candidatos foram ajustados, e o modelo completo para tratamento e tempo – contendo efeitos pricipais e interação nos efeitos fixos – foi selecionado. A posterior comparação entre os tratamentos mostra que os tratamentos se diferenciam para apenas alguns compostos. Em geral, o tratamento aplicado *A. grandis* se mostra superior ao controle em 8 compostos, e difere de outros tratamentos para os compostos β-Pinene, β-Ocimene e α-guaine. Há pouca diferenciação entre os demais tratamentos entre si, isto é, não são verificadas diferenças significativas, sob nível de 5%, entre os tratamentos *E. heros*, *S. frugiperda* e dano mecânico para nenhum composto, bem como não são verificadas diferenças para *E. heros* e dano mecânico com relação ao controle. Ao longo do tempo, para *A. grandis*, *E. heros*, *S. frugiperda* e

dano mecânico, há um aumento significativo nas massas dos compostos β-Ocimene, DMNT e TMTT.

Capítulo 7

Considerações finais

No trabalho apresentado, utiliza-se a abordagem de modelos lineares generalizados de efeitos mistos para lidar com conjuntos de dados quando a suposição de normalidade não é atendida e, ainda, para os quais há algum tipo de dependência entre as observações. Como caso geral, quando se têm dados multivariados longitudinais, isto é, vetores de variáveis resposta observados ao longo do tempo, apresenta-se o que pode ser considerada uma extensão natural do caso univariado de modelos lineares generalizados mistos. No caso mais simples, quando se especifica um vetor de efeitos aleatórios de indivíduo (ou *cluster*) referentes a cada variável resposta, passa-se a ter a possibilidade de analisar as correlações entre as variáveis de interesse de forma mais direta, uma vez que os efeitos de tratamentos e outras covariáveis são, ao menos em parte, capturados pelos efeitos fixos.

Além da interpretabilidade, a formulação abordada permite também que as ferramentas computacionais existentes para casos univariados – que são mais freqüentes e para os quais há maior disponibilidade de implementações – possam ser utilizadas na estimação dos modelos, bastando a especificação adequada dos mesmos. Entretanto, existem duas limitações a serem notadas neste caso: A primeira refere-se à especificação do preditor linear, uma vez que este será comum para todas as variáveis resposta; a segunda, diz respeito à distribuição do vetor de variáveis resposta, já que apenas uma única distribuição é especificada. Para os casos em que se deseja maior flexibilidade nestes aspectos, o pacote *sabreR*, por exemplo, oferece a possibilidade de especificar, para respostas trivariadas, diferentes preditores lineares e diferentes distribuições (normal, binomial e/ou Poisson).

Em particular, tem-se interesse em tratar MLGs com efeitos mistos quando se supõe que as variáveis resposta de interesse têm distribuição Poisson composta, com suporte em $\mathbb{R}_+$ e que assume como processo gerador uma mistura de distribuições

gama com número de componentes distribuídos de acordo com uma distribuição Poisson. Esta distribuição permite, portanto, que dados como massa observada, por exemplo, sejam modelados sem a necessidade de procedimentos *ad hoc* para lidar com a presença de valores nulos.

Por não apresentar forma analítica fechada, porém, o cálculo da função de densidade desta distribuição depende de métodos numéricos de aproximação, fazendo com que sua implementação seja não trivial. Aliando-se a isso o fato de que modelos lineares generalizados mistos também são de difícil implementação, o ajuste destes modelos para dados com distribuição Poisson composta atualmente é bastante restrito. De fato, até o momento da elaboração deste trabalho, apenas o pacote *cplm* do *R* tem este tipo de modelo disponível.

Como motivação do uso de MLGs mistos com efeitos aleatórios multivariados, foram analisados dados em que as variáveis de interesse são as massas (em $10^3\mu g$) de três compostos bioquímicos, retirados de um conjunto de dados de 25 compostos. Na análise destes três compostos, faz-se uma avaliação da utilização do modelo multivariado frente a modelos mais simples, com suposições mais restritivas. Dessa maneira, foram ajustadas três variações de modelos para resposta com distribuição Poisson composta: MLG univariado sem efeito aleatório, MLG misto univariado e MLG misto multivariado. Para os compostos selecionados, verifica-se que os modelos de efeitos mistos – uni e multivariado – apresentam melhora no ajuste dos dados com relação ao MLG sem efeito aleatório. Além disso, na comparação entre os tratamentos, as duas abordagens de efeitos mistos apresentam resultados similares e são, de maneira geral, mais conservadoras que o MLG, mostrando que, ao não levar em consideração as dependências entre observações, modelos mais simples podem levar a resultados enganosos.

Para aqueles casos em que se tem alta dimensionalidade do vetor de variáveis resposta, apresenta-se a metodologia de pseudo-verossimilhança de estimação dos parâmetros do modelo multivariado. A metodologia, introduzida por Fieuws *et al.* (2006) e que se caracteriza pela estimação dos parâmetros via ajuste de modelos bivariados, tem como principal atrativo a redução do custo computacional associado à estimação de parâmetros de modelos lineares generalizados mistos via integração multivariada em grandes dimensões.

De fato, conforme observado no Capítulo 5, a abordagem de pseudo-verossimilhança apresenta redução no tempo de computação e se mostra vantajosa mesmo em modelos com vetores de resposta de dimensão 3. A metodologia par-a-par foi implementada com o pacote estatístico *R* (R Development Core Team, 2012), tomando-se as funções do pacote *cplm* (Zhang, 2012a) como base para os ajustes dos modelos

bivariados, e explorando as possibilidades de paralelização para redução do tempo de estimação.

Como aplicação da estimação de modelos multivariados via pseudo-verossimilhança, foram analisados dados de experimento agronômico com plantas de algodão, submetidas a um de cinco tratamentos (Controle, *A. grandis*, *E. heros*, *S. frugiperda* e dano mecânico) e as massas de 25 compostos observadas em 4 tempos distintos (24, 48, 72 e 96 horas após aplicação do tratamento).

No ajuste dos dados, verificaram-se problemas numéricos de convergência envolvendo dois compostos. Uma vez retirados estes dois compostos, o modelo multivariado pôde ser ajustado, e permitiu verificar que, dos 23 compostos analisados, em 10 são verificadas diferenças significativas entre efeitos de tratamentos ou de tempo. Há indicação de que para 8 compostos o tratamento *A. grandis* é superior ao tratamento controle.

De forma geral, modelos lineares generalizados de efeitos mistos, com efeitos aleatórios multivariados se mostram uma ferramenta importante na análise de dados longitudinais com respostas multivariadas, uma vez que a partir desta classe de modelos, é possível capturar o efeito de dependências entre observações, característico de estudos longitudinais, bem como o efeito das possíveis correlações entre as variáveis de interesse. Com o advento do método de pseudo-verossimilhança na estimação dos parâmetros, há um ganho computacional que permite que esses modelos possam ser utilizados para dados com dimensionalidade elevada.

Bibliografia

Abramowitz, M; Stegun, I. **Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables**. Now York: Dover Publications, 1972. 470p.

Aitkinson, A.C. **Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis**. New York: Oxford University Press, 1985.

Bates, D. M. **Linear Mixed Model Implementation in lme4**. 2011. *http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf*

Besag, J. **Statistical analysis of non-lattice data**. *Statistician* **24** (1975), 179-195.

Box, G.E.P.; Cox, D.R. **An analysis of transformations**. *Journal of the Royal Statistical Society, Series B* **26** (1964), 211-252.

Breslow, N.E.; Clayton, D.G. **Approximate Inference in Generalized Linear Mixed Models**. *Journal of the American Statistical Association* **88** (1993), 9-25.

Demétrio, C.G.B. **Modelos Lineares Generalizados em Experimentação Agronômica**. Apostila, ESALQ/USP, 121p.

Dunn, P.K.; Smyth, G.K. **Series evaluation of Tweedie exponential dispersion model densities**. *Statistics and Computing* **15**(4) (2005). 267-280.

Fieuws, S.; Verbeke, G. **Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles**. *Biometrics* **62** (2006). 424-431.

Fieuws, S.; Verbeke, G.; Boen, P.; Delecluse, C.; **High dimensional multivariate mixed models for binary questionnaire data**. *Applied Statistics* **55**(4) (2006). 449-460.

Hall, B.; Wang, L. **Two-component mixtures of generalized linear mixed effects models for cluster correlated data.** *Statistical Modelling* **5** (2005), 21-37.

Hare, J. D. **Ecological Role of Volatiles Produced by Plants in Response to Damage by Herbivorous Insects.** *Annu. Rev. Entomol.* **56** (2011), 161-180.

Jørgensen, B. **Exponential dispersion models (with discussion).** *J. R. Stat. Soc. Ser. B Stat. Methodol.* **49** (1987). 127-162.

Laird, N. M.; Ware, J. H. **Random-effects models for longitudinal data.** *Biometrics* **38**(4) (1982). 963-974.

Liu, Q.; Pierce, D. A. **A note on Gauss-Hermite quadrature.** *Biometrika* **81**(3) (1994). 624-629.

Magalhães, D. M.; Borges, M; Laumann, R. A.; Sujii, E. R.; Mayon, P.; Caulfield, J. C.; Midega, C. A. O.; Khan, Z. R.; Pickett, J. A.; Birkett, M. A.; Blassioli-Moraes, M. C. **Semiochemicals from Herbivory Induced Cotton Plants Enhance the Foraging Behavior of the Cotton Boll Weevil, *Anthonomus grandis*.** *J. Chem. Ecol.* **38** (2012). 1528-1538.

McCullagh, P.; Nelder, J. A. **Generalized Linear Models.** 2. ed. London: Chapman & Hall, 1989.

Nelder, J. A.; Wedderburn, R. W. M. **Generalized Linear Models**, *Journal of the Royal Statistical Society* **135** (1972), 370-384.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing: Vienna, Austria. 2012. **ISBN** 3-900051-07-0. *http://www.R-project.org/.*

Revolution Analytics. **doParallel: Foreach parallel adaptor for the parallel package.** 2012. http://CRAN.R-project.org/package=doParallel

Verbeke, G.; Molenberghs, G. **Models for Discrete Longitudinal Data.** 1. ed. New York: Springer, 2005.

Vieira, A.M.C. **Modelagem Simultânea de média e dispersão e aplicações na pesquisa agronômica.** 2008. 117 f.. Tese (Doutorado em Estatística) - Escola

Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, São Paulo, 2008.

Vieira, A.M.C.; Hinde, J.P.; Demétrio, C.G.B. **Zero-inflated proportion data models applied to a biological control assay**. *Journal of Applied Statistics* **27**(3) (2000), 373-389.

Wedderburn, R.W.M. **Quasi-likelihood functions, generalized linear models, and Gauss-Newton method**. *Biometrika* **61** (1974), 439-447.

Yang, Z; Sun, X. **Generating Half-normal Plot for Zero-inflated Binomial Regression**. 2006. *http://www.lexjansen.com/pharmasug/2006/statisticspharmacokinetics/sp05.pd*

Zhang, W. **cplm: Compound Poisson linear models. R package version 0.6-4**. 2012a. http://CRAN.R-project.org/package=cplm

Zhang, W. **Likelihood-based and Bayesian Methods for Tweedie Compound Poisson Linear Mixed Models**. 2012b. http://www.actuaryzhang.com/publication/publi

Apêndice A

Tabelas - Análise de dados de algodão (Todos os compostos)

Tabela A.2: Tabela de valores preditos por composto, tratamento e tempo – letras diferentes indicam tratamentos que diferem significativamente, com nível de significância de 5%

Composto	Tratamento	24h	48h	72h	96h
α-Pinene	Controle	$3.42^{a,a}$	$3.61^{a,a}$	$3.8^{a,a}$	$4.01^{a,a}$
	Frugiperda	$8.96^{a,a}$	$10.59^{a,ab}$	$12.52^{a,b}$	$14.8^{a,ab}$
	Grandis	$11.88^{a,a}$	$12.86^{a,b}$	$13.92^{a,b}$	$15.06^{a,b}$
	Heros	$4.16^{a,a}$	$4.97^{a,ab}$	$5.94^{a,ab}$	$7.09^{a,ab}$
	Mecânico	$8.03^{a,a}$	$7.3^{a,ab}$	$6.64^{a,ab}$	$6.04^{a,ab}$
Camphene	Controle	$0.61^{a,a}$	$0.47^{a,a}$	$0.37^{a,a}$	$0.28^{a,a}$
	Frugiperda	$0.47^{a,a}$	$0.43^{a,a}$	$0.39^{a,a}$	$0.36^{a,a}$
	Grandis	$0.77^{a,a}$	$0.61^{a,a}$	$0.49^{a,a}$	$0.39^{a,a}$
	Heros	$0.27^{a,a}$	$0.34^{a,a}$	$0.43^{a,a}$	$0.53^{a,a}$
	Mecânico	$0.36^{a,a}$	$0.36^{a,a}$	$0.36^{a,a}$	$0.37^{a,a}$
Belzaldehyde	Controle	$0.41^{a,a}$	$0.44^{a,a}$	$0.48^{a,a}$	$0.52^{a,a}$
	Frugiperda	$0.47^{a,a}$	$0.44^{a,a}$	$0.41^{a,a}$	$0.38^{a,a}$
	Grandis	$0.51^{a,a}$	$0.49^{a,a}$	$0.47^{a,a}$	$0.45^{a,a}$
	Heros	$0.34^{a,a}$	$0.38^{a,a}$	$0.43^{a,a}$	$0.48^{a,a}$
	Mecânico	$0.49^{a,a}$	$0.55^{a,a}$	$0.61^{a,a}$	$0.67^{a,a}$
	Controle	$1.28^{a,a}$	$1.3^{a,a}$	$1.32^{a,a}$	$1.33^{a,a}$
	Frugiperda	$2.55^{a,ab}$	$2.68^{a,ab}$	$2.82^{a,ab}$	$2.96^{a,ab}$

β-Pinene	Grandis	$6.72^{a,b}$	$7.42^{a,b}$	$8.2^{a,b}$	$9.05^{a,b}$
	Heros	$1.23^{a,a}$	$1.37^{a,a}$	$1.52^{a,a}$	$1.69^{a,a}$
	Mecânico	$1.76^{a,ab}$	$1.99^{a,ab}$	$2.26^{a,ab}$	$2.57^{a,ab}$
β-Myrcene	Controle	$1.11^{a,a}$	$0.96^{a,a}$	$0.83^{a,a}$	$0.72^{a,a}$
	Frugiperda	$2.65^{a,a}$	$3.13^{a,ab}$	$3.69^{a,ab}$	$4.35^{a,ab}$
	Grandis	$4.55^{a,a}$	$4.92^{a,b}$	$5.31^{a,b}$	$5.74^{a,b}$
	Heros	$1.06^{a,a}$	$1.41^{a,ab}$	$1.88^{a,ab}$	$2.51^{a,ab}$
	Mecânico	$2.73^{a,a}$	$2.43^{a,ab}$	$2.16^{a,ab}$	$1.92^{a,ab}$
(Z)-3-Hexenylacetate	Controle	$1.07^{a,a}$	$1.01^{a,a}$	$0.95^{a,a}$	$0.9^{a,a}$
	Frugiperda	$1.5^{a,a}$	$1.59^{a,ab}$	$1.69^{a,a}$	$1.79^{a,a}$
	Grandis	$3.24^{a,a}$	$2.86^{a,b}$	$2.53^{a,a}$	$2.23^{a,a}$
	Heros	$0.98^{a,a}$	$1.04^{a,ab}$	$1.09^{a,a}$	$1.15^{a,a}$
	Mecânico	$1.22^{a,a}$	$1.23^{a,ab}$	$1.24^{a,a}$	$1.26^{a,a}$
Limonene	Controle	$0.67^{a,a}$	$0.68^{a,a}$	$0.69^{a,a}$	$0.71^{a,a}$
	Frugiperda	$0.65^{a,a}$	$0.74^{a,a}$	$0.84^{a,a}$	$0.96^{a,a}$
	Grandis	$1.45^{a,a}$	$1.39^{a,a}$	$1.33^{a,a}$	$1.28^{a,a}$
	Heros	$0.49^{a,a}$	$0.56^{a,a}$	$0.65^{a,a}$	$0.76^{a,a}$
	Mecânico	$0.95^{a,a}$	$1.02^{a,a}$	$1.1^{a,a}$	$1.19^{a,a}$
β-Ocimene	Controle	$0.28^{a,a}$	$0.4^{a,a}$	$0.58^{a,a}$	$0.84^{a,a}$
	Frugiperda	$2.07^{a,bc}$	$2.98^{b,bc}$	$4.28^{c,bc}$	$6.15^{d,bc}$
	Grandis	$5.6^{a,b}$	$8^{b,b}$	$11.43^{c,b}$	$16.35^{d,b}$
	Heros	$0.67^{a,ac}$	$1.04^{b,ac}$	$1.62^{c,ac}$	$2.52^{d,ac}$
	Mecânico	$0.47^{a,ac}$	$0.75^{b,ac}$	$1.2^{c,ac}$	$1.92^{d,ac}$
Linalool	Controle	$0.27^{a,a}$	$0.3^{a,a}$	$0.34^{a,a}$	$0.39^{a,a}$
	Frugiperda	$0.24^{a,a}$	$0.32^{a,a}$	$0.42^{a,a}$	$0.56^{a,a}$
	Grandis	$0.75^{a,a}$	$0.56^{a,a}$	$0.41^{a,a}$	$0.31^{a,a}$
	Heros	$0.25^{a,a}$	$0.27^{a,a}$	$0.28^{a,a}$	$0.3^{a,a}$
	Mecânico	$0.18^{a,a}$	$0.2^{a,a}$	$0.21^{a,a}$	$0.23^{a,a}$
DMNT	Controle	$1.32^{a,a}$	$2^{b,a}$	$3.02^{c,a}$	$4.55^{d,a}$
	Frugiperda	$5.1^{a,a}$	$8.01^{b,a}$	$12.57^{c,a}$	$19.73^{d,a}$
	Grandis	$2.2^{a,a}$	$3.07^{b,a}$	$4.28^{c,a}$	$5.98^{d,a}$
	Heros	$1.4^{a,a}$	$3.54^{b,a}$	$8.98^{c,a}$	$22.77^{d,a}$
	Mecânico	$1.71^{a,a}$	$2.81^{b,a}$	$4.63^{c,a}$	$7.63^{d,a}$
trans-3-HexenylButyrate	Controle	$0.23^{a,a}$	$0.19^{a,a}$	$0.15^{a,a}$	$0.12^{a,a}$
	Frugiperda	$0.32^{a,a}$	$0.32^{a,a}$	$0.32^{a,a}$	$0.32^{a,a}$
	Grandis	$0.4^{a,a}$	$0.24^{a,a}$	$0.14^{a,a}$	$0.09^{a,a}$

	Heros	$0.22^{a,a}$	$0.23^{a,a}$	$0.24^{a,a}$	$0.25^{a,a}$
	Mecânico	$0.19^{a,a}$	$0.13^{a,a}$	$0.09^{a,a}$	$0.06^{a,a}$
Methylsalicylate	Controle	$0.89^{a,a}$	$0.67^{a,a}$	$0.51^{a,a}$	$0.38^{a,a}$
	Frugiperda	$1.13^{a,a}$	$0.76^{a,a}$	$0.51^{a,a}$	$0.35^{a,a}$
	Grandis	$0.76^{a,a}$	$0.55^{a,a}$	$0.39^{a,a}$	$0.28^{a,a}$
	Heros	$0.56^{a,a}$	$0.52^{a,a}$	$0.48^{a,a}$	$0.45^{a,a}$
	Mecânico	$0.46^{a,a}$	$0.49^{a,a}$	$0.52^{a,a}$	$0.55^{a,a}$
Indole	Controle	$0.2^{a,a}$	$0.17^{a,a}$	$0.14^{a,a}$	$0.12^{a,a}$
	Frugiperda	$0.61^{a,a}$	$0.47^{a,a}$	$0.36^{a,a}$	$0.27^{a,a}$
	Grandis	$0.48^{a,a}$	$0.35^{a,a}$	$0.25^{a,a}$	$0.18^{a,a}$
	Heros	$0.34^{a,a}$	$0.22^{a,a}$	$0.14^{a,a}$	$0.09^{a,a}$
	Mecânico	$0.11^{a,a}$	$0.09^{a,a}$	$0.07^{a,a}$	$0.06^{a,a}$
cis-Jasmone	Controle	$0.51^{a,a}$	$0.34^{a,a}$	$0.23^{a,a}$	$0.15^{a,a}$
	Frugiperda	$0.81^{a,a}$	$0.74^{a,a}$	$0.68^{a,a}$	$0.62^{a,a}$
	Grandis	$1.11^{a,a}$	$0.66^{a,a}$	$0.39^{a,a}$	$0.24^{a,a}$
	Heros	$0.55^{a,a}$	$0.4^{a,a}$	$0.3^{a,a}$	$0.22^{a,a}$
	Mecânico	$0.48^{a,a}$	$0.39^{a,a}$	$0.31^{a,a}$	$0.25^{a,a}$
α-Copaene	Controle	$0.19^{a,a}$	$0.22^{a,a}$	$0.24^{a,a}$	$0.27^{a,a}$
	Frugiperda	$0.25^{a,a}$	$0.32^{a,a}$	$0.4^{a,a}$	$0.51^{a,a}$
	Grandis	$0.38^{a,a}$	$0.33^{a,a}$	$0.29^{a,a}$	$0.25^{a,a}$
	Heros	$0.13^{a,a}$	$0.16^{a,a}$	$0.19^{a,a}$	$0.23^{a,a}$
	Mecânico	$0.2^{a,a}$	$0.19^{a,a}$	$0.17^{a,a}$	$0.15^{a,a}$
α-Guaine	Controle	$0.28^{a,a}$	$0.36^{a,a}$	$0.47^{a,ab}$	$0.62^{a,a}$
	Frugiperda	$0.53^{a,a}$	$0.48^{a,a}$	$0.44^{a,ab}$	$0.4^{a,a}$
	Grandis	$4.36^{a,b}$	$2.07^{b,b}$	$0.98^{c,b}$	$0.47^{d,a}$
	Heros	$0.25^{a,a}$	$0.24^{a,a}$	$0.24^{a,ab}$	$0.23^{a,a}$
	Mecânico	$0.39^{a,a}$	$0.27^{a,a}$	$0.18^{a,a}$	$0.12^{a,a}$
Geranylacetone	Controle	$0.29^{a,a}$	$0.23^{a,a}$	$0.18^{a,a}$	$0.14^{a,a}$
	Frugiperda	$0.23^{a,a}$	$0.28^{a,a}$	$0.34^{a,a}$	$0.41^{a,a}$
	Grandis	$0.14^{a,a}$	$0.2^{a,a}$	$0.27^{a,a}$	$0.36^{a,a}$
	Heros	$0.52^{a,a}$	$0.36^{a,a}$	$0.25^{a,a}$	$0.17^{a,a}$
	Mecânico	$0.26^{a,a}$	$0.2^{a,a}$	$0.16^{a,a}$	$0.12^{a,a}$
α-Caryophyllene	Controle	$0.49^{a,a}$	$0.48^{a,a}$	$0.47^{a,a}$	$0.45^{a,a}$
	Frugiperda	$1.66^{a,a}$	$2.12^{a,b}$	$2.71^{a,b}$	$3.47^{a,b}$
	Grandis	$1.93^{a,a}$	$1.72^{a,ab}$	$1.54^{a,ab}$	$1.37^{a,ab}$
	Heros	$0.62^{a,a}$	$0.67^{a,ab}$	$0.72^{a,ab}$	$0.78^{a,ab}$

	Mecânico	$1.06^{a,a}$	$1^{a,ab}$	$0.96^{a,ab}$	$0.91^{a,ab}$
Alloaromadendrene	Controle	$0.34^{a,a}$	$0.46^{a,a}$	$0.61^{a,a}$	$0.82^{a,a}$
	Frugiperda	$0.78^{a,a}$	$0.66^{a,a}$	$0.56^{a,a}$	$0.48^{a,a}$
	Grandis	$0.82^{a,a}$	$0.74^{a,a}$	$0.67^{a,a}$	$0.61^{a,a}$
	Heros	$0.33^{a,a}$	$0.4^{a,a}$	$0.48^{a,a}$	$0.58^{a,a}$
	Mecânico	$0.51^{a,a}$	$0.54^{a,a}$	$0.57^{a,a}$	$0.6^{a,a}$
Sesquiterpene	Controle	$0.35^{a,a}$	$0.34^{a,a}$	$0.32^{a,a}$	$0.3^{a,a}$
	Frugiperda	$0.41^{a,a}$	$0.43^{a,a}$	$0.45^{a,a}$	$0.47^{a,a}$
	Grandis	$1.17^{a,a}$	$0.77^{a,a}$	$0.5^{a,a}$	$0.33^{a,a}$
	Heros	$0.33^{a,a}$	$0.29^{a,a}$	$0.25^{a,a}$	$0.22^{a,a}$
	Mecânico	$0.29^{a,a}$	$0.25^{a,a}$	$0.21^{a,a}$	$0.18^{a,a}$
δ-Guaiane	Controle	$0.4^{a,a}$	$0.35^{a,a}$	$0.31^{a,a}$	$0.27^{a,a}$
	Frugiperda	$1.29^{a,a}$	$1.43^{a,b}$	$1.59^{a,b}$	$1.76^{a,b}$
	Grandis	$0.67^{a,a}$	$0.7^{a,ab}$	$0.73^{a,ab}$	$0.77^{a,ab}$
	Heros	$0.53^{a,a}$	$0.53^{a,ab}$	$0.53^{a,ab}$	$0.52^{a,ab}$
	Mecânico	$0.65^{a,a}$	$0.65^{a,ab}$	$0.64^{a,ab}$	$0.64^{a,ab}$
δ-Cadinene	Controle	$0.78^{a,a}$	$0.94^{a,a}$	$1.12^{a,a}$	$1.34^{a,a}$
	Frugiperda	$1.08^{a,a}$	$1.51^{a,a}$	$2.13^{a,a}$	$2.99^{a,a}$
	Grandis	$0.54^{a,a}$	$0.64^{a,a}$	$0.76^{a,a}$	$0.91^{a,a}$
	Heros	$0.39^{a,a}$	$0.62^{a,a}$	$1^{a,a}$	$1.61^{a,a}$
	Mecânico	$0.99^{a,a}$	$1.06^{a,a}$	$1.14^{a,a}$	$1.22^{a,a}$
TMTT	Controle	$2.23^{a,a}$	$3.91^{b,a}$	$6.85^{c,a}$	$11.99^{d,a}$
	Frugiperda	$3.71^{a,a}$	$7.37^{b,a}$	$14.66^{c,a}$	$29.13^{d,a}$
	Grandis	$1.8^{a,a}$	$3.94^{b,a}$	$8.6^{c,a}$	$18.81^{d,a}$
	Heros	$2.31^{a,a}$	$6.39^{b,a}$	$17.67^{c,a}$	$48.84^{d,a}$
	Mecânico	$3.55^{a,a}$	$6.2^{b,a}$	$10.8^{c,a}$	$18.84^{d,a}$

Tabela A.1: Estimativas dos parâmetros do modelo (6.2) e seus respectivos erros padrões

Composto	Tmp	Controle	Frugiperda	Frugiperda:Tmp	Grandis	Grandis:Tmp	Heros	Heros:Tmp	Mecânico	Mecânico:Tmp
C1	0.002 (0.004)	1.176 (0.362)	2.025 (0.321)	0.005 (0.006)	2.396 (0.312)	0.001 (0.006)	1.246 (0.348)	0.005 (0.006)	2.177 (0.329)	-0.006 (0.006)
C2	-0.011 (0.007)	-0.23 (0.501)	-0.68 (0.515)	0.007 (0.01)	-0.034 (0.484)	0.001 (0.009)	-1.514 (0.536)	0.02 (0.01)	-1.034 (0.515)	0.011 (0.009)
C3	0.003 (0.007)	-0.965 (0.512)	-0.684 (0.505)	-0.006 (0.009)	-0.636 (0.499)	-0.005 (0.009)	-1.188 (0.523)	0.001 (0.009)	-0.809 (0.498)	0.001 (0.009)
C4	0.001 (0.005)	0.235 (0.447)	0.887 (0.411)	0.002 (0.007)	1.806 (0.37)	0.004 (0.007)	0.105 (0.445)	0.004 (0.008)	0.437 (0.422)	0.005 (0.007)
C5	-0.006 (0.006)	0.251 (0.48)	0.811 (0.429)	0.013 (0.007)	1.438 (0.408)	0.009 (0.007)	-0.234 (0.464)	0.018 (0.008)	1.12 (0.434)	0.001 (0.007)
C6	-0.003 (0.006)	0.131 (0.404)	0.347 (0.376)	0.005 (0.008)	1.298 (0.34)	-0.003 (0.007)	-0.067 (0.406)	0.005 (0.008)	0.183 (0.394)	0.003 (0.008)
C7	0.001 (0.006)	-0.423 (0.482)	-0.564 (0.477)	0.005 (0.009)	0.417 (0.436)	-0.003 (0.008)	-0.87 (0.498)	0.005 (0.009)	-0.134 (0.447)	0.002 (0.008)
C8	0.015 (0.007)	-1.637 (0.552)	0.364 (0.44)	0 (0.008)	1.365 (0.404)	0 (0.007)	-0.848 (0.49)	0.003 (0.008)	-1.218 (0.512)	0.004 (0.009)
C9	0.005 (0.007)	-1.444 (0.52)	-1.726 (0.519)	0.007 (0.01)	0.014 (0.457)	-0.018 (0.01)	-1.433 (0.528)	-0.003 (0.01)	-1.757 (0.555)	-0.002 (0.011)
C10	0.017 (0.005)	-0.13 (0.521)	1.178 (0.472)	0.002 (0.006)	0.453 (0.501)	-0.003 (0.006)	-0.596 (0.506)	0.022 (0.006)	0.035 (0.508)	0.004 (0.006)
C11	-0.009 (0.008)	-1.254 (0.581)	-1.146 (0.547)	0.009 (0.011)	-0.402 (0.555)	-0.012 (0.011)	-1.557 (0.57)	0.011 (0.011)	-1.304 (0.609)	-0.006 (0.012)
C12	-0.012 (0.006)	0.161 (0.46)	0.517 (0.45)	-0.005 (0.009)	0.057 (0.475)	-0.002 (0.009)	-0.509 (0.483)	0.009 (0.009)	-0.844 (0.488)	0.014 (0.009)
C14	-0.007 (0.008)	-1.435 (0.707)	-0.22 (0.637)	-0.004 (0.01)	-0.395 (0.657)	-0.007 (0.011)	-0.637 (0.676)	-0.011 (0.011)	-1.978 (0.737)	-0.002 (0.012)
C15	-0.017 (0.007)	-0.264 (0.56)	-0.121 (0.52)	0.013 (0.009)	0.622 (0.52)	-0.005 (0.01)	-0.292 (0.547)	0.004 (0.01)	-0.522 (0.554)	0.008 (0.01)
C16	0.005 (0.008)	-1.764 (0.55)	-1.62 (0.521)	0.005 (0.01)	-0.844 (0.51)	-0.01 (0.01)	-2.178 (0.582)	0.002 (0.011)	-1.488 (0.562)	-0.009 (0.011)
C18	0.011 (0.007)	-1.533 (0.523)	-0.55 (0.489)	-0.015 (0.009)	2.217 (0.391)	-0.042 (0.009)	-1.373 (0.548)	-0.012 (0.01)	-0.57 (0.531)	-0.027 (0.01)
C19	-0.01 (0.008)	-0.984 (0.655)	-1.64 (0.639)	0.018 (0.01)	-2.241 (0.674)	0.023 (0.011)	-0.282 (0.621)	-0.005 (0.01)	-1.082 (0.661)	0 (0.011)
C20	-0.001 (0.007)	-0.677 (0.491)	0.262 (0.403)	0.011 (0.008)	0.772 (0.406)	-0.004 (0.008)	-0.547 (0.468)	0.004 (0.009)	0.102 (0.439)	-0.001 (0.009)
C21	0.012 (0.006)	-1.359 (0.499)	-0.08 (0.459)	-0.019 (0.009)	-0.096 (0.458)	-0.016 (0.009)	-1.289 (0.509)	-0.004 (0.009)	-0.735 (0.484)	-0.01 (0.009)
C22	-0.002 (0.007)	-0.987 (0.521)	-0.947 (0.502)	0.004 (0.01)	0.581 (0.449)	-0.016 (0.009)	-0.967 (0.53)	-0.004 (0.01)	-1.081 (0.539)	-0.004 (0.01)
C23	-0.005 (0.007)	-0.778 (0.49)	0.152 (0.396)	0.01 (0.009)	-0.447 (0.443)	0.007 (0.009)	-0.62 (0.461)	0.005 (0.01)	-0.423 (0.448)	0.005 (0.009)
C24	0.007 (0.006)	-0.424 (0.433)	-0.267 (0.403)	0.007 (0.008)	-0.804 (0.458)	0 (0.008)	-1.426 (0.468)	0.012 (0.008)	-0.077 (0.421)	-0.005 (0.008)
C25	0.023 (0.004)	0.243 (0.491)	0.624 (0.467)	0.005 (0.005)	-0.193 (0.489)	0.009 (0.006)	-0.179 (0.478)	0.019 (0.005)	0.712 (0.474)	0 (0.006)
$\phi = 0.761$										
$p = 1.6$										

Apêndice B

Códigos em *R*

Os códigos em *R* utilizados na confecção deste trabalho podem ser encontrados na página https://github.com/rceratti, divididos em três repositórios:

- *pair.mglmm*: Repositório contendo código fonte, binário windows e documentação do pacote *pair.mglmm* para linguagem *R*, usado nesta dissertação para ajustar modelos mistos multivariados para a distribuição Poisson composta. A documentação fornecida provê exemplos da função *mglmmCP* desenvolvida para o ajuste dos modelos mencionados. O pacote ainda tem limitações tais como restrição ao ajuste de modelos apenas para a distribuição Poisson composta, e forma de especificação do modelo ("formula") pouco intuitiva. Estas limitações serão abordadas em versões futuras do pacote, bem como será adotada a utilização de classes formais e métodos que permitam a utilização das funções de forma mais rotineira no *R*;

- *Multivariado_vs_Par-a-par*: Script *R* contendo o estudo de simulação do Capítulo 5, comparando o ajuste de modelos multivariados mistos Poisson composto via abordagem par-a-par e multivariado. Depende do pacote *pair.mglmm*;

- *Dados_de_algodao*: Scripts *R* para análise dos dados de algodão dos Capítulos 4 (compostos 1, 5 e 8) e 6 (compostos 1 a 25). Apenas para a segunda análise existe dependência do pacote *pair.mglmm*;

B.1 Exemplo de uso do pacote *pair.mglmm*

No pacote desenvolvido, a principal função de interesse é a função `mglmmCP`. Como forma de ilustrar seu uso, toma-se um conjunto de dados simulados pela

função `data.sim`, com vetor de variável resposta de dimensão `m`= 3, isto é, resposta trivariada, e distribuição Poisson composta com $p = 1.6$ e $\phi = 1$:

```
phi <- 1; p <- 1.6
mydat <- data.sim(m = 3, distr = 'CP', link.inv = exp, xi = p, phi = phi)
dat <- mydat$Data
```

Os dados são gerados supondo-se um fator com 4 níveis denominado 'period', para o qual uma matriz de parâmetros de efeitos fixos `beta` de dimensão $4 \times m$ é gerada aleatoriamente – que também pode ser especificada com o argumento `beta`. O objeto `mydat` gerado é do tipo lista contendo os dados, a matriz de efeitos fixos e a matriz de componente de variâncias usada na geração dos efeitos aleatórios. O conjunto de dados, alocado no objeto `dat` no código acima, tem formato longo contendo as variáveis 'period', 'ID' (p. ex., identificação da planta observada ao longo dos períodos), 'variable' (fator indicando o nível da variável resposta) e 'value' (resposta observada), tal como abaixo:

```
> head(dat)
  period ID variable       value
1      1  1       C1 2.995789451
2      2  1       C1 0.387148781
3      3  1       C1 0.807502808
4      4  1       C1 0.000000000
5      1  2       C1 2.276448618
6      2  2       C1 0.007743925
```

Gerados os dados, o modelo é especificado na função `mglmmCP` com os argumentos:

```
mglmmCP(formula, id, data, cl)
```

em que `formula` é a especificação do modelo propriamente dita, seguindo a formulação da função `glmer` do pacote `lme4`. Para os dados gerados, tem-se:

```
Form0 <- value ~ 0 + variable + variable:period + (0 + variable|ID)
```

Os efeitos fixos são especificados em `0 + variable + variable:period`, em que o termo `0` é usado para remover o intercepto geral e `variable:period` cria o termo de interação entre 'variable' e 'period' sem termos de efeito principal para 'period'. Já os efeitos aleatórios são especificados em `(0 + variable|ID)`, gerando um intercepto aleatório para 'ID' em cada nível de 'variable'.

O argumento `id` da função é um vetor que indica os níveis da variável resposta. Para os dados acima, tem-se:

```
id <- dat$variable
```

Além disso, especificam-se o objeto que contém os dados e o *cluster* a ser utilizado para paralelizar o processamento. Assim tem-se que o ajuste do modelo é finalmente realizado fazendo-se:

```
cl <- makeCluster(4)  # 4 núcleos de processamento
registerDoParallel(cl)
clusterEvalQ(cl, library(pair.mglmm))

m0 <- mglmmCP(formula = Form0, id = id, data = dat, cl = cl)
```

O objeto `m0` criado é uma lista contendo efeitos fixos, matriz de componentes de variância, p e ϕ estimados, log-verossimilhança, graus de liberdade, matriz de efeitos aleatórios preditos, valores ajustados e resíduos.